Habitat suitability and community modelling of marine benthos

Marine Biology Research Group

Campus Sterre – S8

Krijgslaan 281

B-9000 Ghent

Belgium

Academic Year 2010-2011

Publically defended on July 8th, 2011

Cover picture of nematode by Sofie Derycke
Printed by *Druk in de weer*, Ghent on 100% recycled paper and with vegetal ink.

Co-authored one or more chapters:

Tim Ferrero, Tom Gheskiere, Peter Goethals, John Lambshead, Andrea McEvoy, Michaela Schratzberger, Maaike Steyaert, Jan Vanaverbeke, Marc Van Meirvenne, Ann Vanreusel, Magda Vincx

For citation to published work reprinted in this thesis, please refer to the original publications as mentioned at the beginning of each chapter.

To refer to this thesis, please cite as:

Merckx, B., 2011. Habitat suitability and community modelling of marine benthos. Ghent University, Ghent, Belgium, pp. 309 + xvi.

ISBN: 978-90-77713-87-7

This PhD was financially supported by the Research Foundation - Flanders (FWO-Vlaanderen) and Ghent University.

Habitat suitability and community modelling of marine benthos

Modeleren van habitatgeschiktheid en gemeenschapsstructuren van marien benthos

Bea Merckx

Promotor: Prof. Dr. Magda Vincx
Co-promotor: Dr. Jan Vanaverbeke
Academic year 2010-2011

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor in Science (Marine Sciences)

Members of the reading committee

Prof. Dr. Ann Vanreusel
Ghent University, Ghent, Belgium

Prof. Dr. Peter Goethals
Ghent University, Ghent, Belgium

Dr. Paul Somerfield
Plymouth Marine Laboratory, Plymouth, U.K.

Dr. Tom Ysebaert
NIOO-CEME, Yerseke, The Netherlands

Members of the examination committee

Prof. Dr. Tom Moens, Chairman
Ghent University, Ghent, Belgium

Prof. Dr. Magda Vincx, Promotor
Ghent University, Ghent, Belgium

Dr. Jan Vanaverbeke, Co-promotor
Ghent University, Ghent, Belgium

Prof. Dr. Ann Vanreusel
Ghent University, Ghent, Belgium

Prof. Dr. Peter Goethals
Ghent University, Ghent, Belgium

Prof. Dr. Steven Degraer
KBIN-MUMM, Brussels, Belgium

Prof. Dr. Karline Soetaert
NIOO-CEME, Yerseke, The Netherlands

Prof. Dr. Jan Mees
VLIZ, Ostend, Belgium

Dr. Paul Somerfield
Plymouth Marine Laboratory, Plymouth, U.K.

Dr. Tom Ysebaert
NIOO-CEME, Yerseke, The Netherlands

TABLE OF CONTENTS

DANKWOORD

‘Het boekske’ is er. In exact 89 985 woorden vind je hier het werk van enkele jaren samengevat. Zo’n boekje komt niet tot stand door in een ivoren toren te leven, een misverstand dat soms de ronde doet. Integendeel, onderzoek vraagt een intense samenwerking tussen wetenschappers van verschillende disciplines. Niet alleen (mariene) biologen hebben bijgedragen tot het tot stand komen van dit boekje, ook geologen, geografen, statistici, programmeurs, ja zelfs ruimtewetenschappers, vormden een bouwsteentje in dit werk.

Een onmisbare bouwsteen is mijn promotor Magda Vincx. Zij bood mij de kans om hier als doctoraatsstudent aan de slag te gaan. Niettegenstaande mijn beperkte kansen om een beurs vast te krijgen wegens ‘aanvaardbare leeftijdsgrens overschreden’, gaf ze me toch de kans om een doctoraatsonderzoek te starten. Ze gaf me steeds de vrijheid om mijn eigen weg te zoeken in het doolhof van modeleer- en programmeertechnieken. Af en toe informeerde ze naar de voortgang van ‘het boekske’, maar vooral ervoer ik veel steun bij het verloop van het onderzoek.

Verder kon ik altijd terecht bij mijn co-promotor Jan Vanaverbeke. Hij bekeek mijn schrijfsels altijd met een kritische blik, gaf terechte commentaren en steunde me in mijn keuzes. Ik kon altijd bij hem terecht wanneer ik nood had aan nematologische ‘expert knowledge’. Hij is iemand die zijn doctoraatsstudenten door dik en dun steunt, waarvoor dank.

Daarnaast wens ik ook Peter Goethals te bedanken om mij te helpen bij mijn eerste stappen bij het gebruik van neurale netwerken, de artificiële, wel te verstaan. Telkens kon ik bij hem terecht met een reeks vragen over wat, hoe, waarom en waarmee die neurale netwerken getemd konden worden.

Ook Marc Van Meirvenne mag niet op het lijstje ontbreken. Nooit is een cursus statistiek mij zo duidelijk uitgelegd als door hem. Stap voor stap leidt hij je binnen in de wondere wereld van de geostatistiek, het kriggen, de variogrammen en de nuggets. Ook zijn kritische stem bij het artikel dat voortkwam uit deze hersenspinsels was meer dan welkom.

Although far away in the US, I would like to thank both Richard Pearson and Steven Phillips for the wonderful workshop on Maxent modelling in the mountains of Arizona. Although, we did not see any bears or mountain lions, this was a workshop that I wil never forget: superb location, nice people, fantastic teachers and great software.

Furthermore, I would like to thank all the members of the examination committee for their critical but fair comments to this PhD: Paul Somerfield, Tom Ysebaert, Karline Soetaert, Ann Vanreusel, Peter Goethals, Steven Degraer, Jan Mees en Tom Moens.

Bij het modelleren ben je steeds afhankelijk van data en geen data zonder dat er mensen achter die data zitten. Dit onderzoek is dan ook maar tot stand kunnen komen omdat ik op de schouders van reuzen kon staan. Voor de nematodendata wil ik graag alle dataleveranciers bedanken in het bijzonder Magda Vincx en Ann Vanreusel die hun zorgvuldig getrieerde, gedetermineerde en gepubliceerde data graag ter beschikking stelden. In dit lijstje mogen ook Jan Vanaverbeke en Maaike Steyaert niet ontbreken.Maaike, bedankt voor het geduld dat je opbracht om mijn eindeloze vragenreeksen te blijven beantwoorden. Daarnaast wens ik ook Carlo Heip, Maarten Raes, Sandra Vanhove, Saskia Van Gaever, Tom Gheskiere, Matthew Lammertyn en Jyotsna Sharma te bedanken. Hun data is nu vereeuwigd in de UGent-databank. Ook alle andere mensen van wie ik data verzamelde, maar die ik nooit persoonlijk ontmoette: Chen Guotong, Gonda Bisschop, Jian Li, Preben Jensen, Zhang Derong en Regine Vandenberghe. Bedankt, zonder jullie was dit doctoraat nooit tot stand kunnen komen. Ook wens ik Tom Moens te bedanken voor het nakijken en corrigeren van de lijst met voedingstypes van nematoden.

Hoewel de nematoden de hoofdmoot uitmaken in dit werk, was ik steeds blij ook eens een groter 'beest' te mogen modelleren. Daarbij wil ik graag Marijn Rabaut bedanken omdat ik mijn modellen ook eens mocht loslaten op zijn geliefde schelpkokerworm, *Lanice conchilega*. In het kader van het Ensis project mocht ik, in samenwerking met Jean-Sébastien Houziaux en Steven Degraer, ook de Amerikaanse zwaardschede, een exoot en daarom minder geliefd in biologische kringen, eens door de PC halen. Er werd er duchtig op los gemodelleerd: aan de Belgische kust, aan de Nederlandse kust, kleine ensissen, grote ensissen, veel ensissen, weinig ensissen, wat maar kon gemodelleerd worden, moest eraan geloven.

Naast biologische data had ik ook omgevingsvariabelen nodig. Voor de onafhankelijke variabelen was ik vooral afhankelijk van onafhankelijke instituten. Instituten met betrokken mensen die steeds bereid waren hun data te delen. Ik denk daarbij in de eerste plaats aan Els Verfaillie die steeds behulpzaam was bij het vinden van nog een extra kaart van nog een andere omgevingsvariabele. Ook Dries Vanden Eynde (stromingskaartjes), Barbara Van Mol (Belcolour kaartjes) en Koen Degrendele (blackbox data zandextractie) bedankt voor jullie hulp bij het bezorgen van de geschikte data.

Daarnaast wil ik ook al mijn collega's op het labo bedanken voor de tussendoorgesprekjes, de staalnames die vaak een leuke afwisseling vormden, de koffiepauzes, de tractaties en de fijne recepties. Het feit dat ik 'into' databanken ben, heb ik volledig te danken aan Tim; bedankt Annick en Isolde om mij wegwijs te maken in het papieren doolhof; bedankt Guy voor de hulp bij soft- & hardware problemen; bedankt Pieter voor de pogingen om last minute mijn PC te reanimeren; bedankt Julie en Ulrike om mij bij de laatste loodjes nog bij te staan met jullie 'last minute' goeie raad; bedankt ook aan alle andere collega's en ex-collega's voor bij- en tussen koffie gesprekjes: Anneke, Bart, Hendrik, Tania, Carl, Joke, Sofie, Marijn, Katja, Delphine, Ellen, Clio, Annelies, Saskia, Jan, Dirk, Karen, Annick, Jeroen, Marleen, Nele, Freija, Thibaud, Francesca, Jelle, Sarah, Annelien, Niels.

Naast de werkuren, heb ik de laatste jaren veel tijd doorgebracht als ‘activist’ bij het Gents MilieuFront. Ook de GMF’ers wil ik bedanken: we hebben de laatste jaren al vele uurtjes brainstormend met elkaar doorgebracht. Dat was niet alleen leuk, maar ook zinvol. Merci, aan alle wisselstroom- en groene loper-fanaten!

Ook mijn vrienden wil ik zeker bedanken. Bij hen kan ik altijd terecht wanneer ik nood heb aan een luisterend oor. Marijke, Barbara en Elke de vele etentjes, kaart- en praatavonden waren altijd iets waar ik enorm naar uitkeek. Annelies toen ik nog biologie studeerde, heb ik ongelooflijk veel gehad aan jouw praktische en morele steun om door die drukke periode heen te geraken. Iets wat ik nooit zal vergeten. Steven, merci om mij tijdens de laatste maanden af en toe het huis uit te sleuren voor een filmpje of een etentje. Lieven, het was altijd een leuke afwisseling wanneer je eens binnensprong na een examen voor een wijntje of een mojito. Noémie, merci voor je aanmoedigingsmailtjes vanuit Portugal. Ook Alice, Antoon, Janne, Luc, Luk, Hendrik, en alle anderen, merci voor de bezoekjes en de steun...

And last but not least, wil ik mijn ouders bedanken. Zij hebben me ondanks mijn ietwat vreemde loopbaan altijd moreel gesteund. Als er iets nodig was, kon ik altijd bij hen terecht voor zowel praktische als morele steun. Niets is hen te veel: maandenlang renoveren, liters soep koken, biologische groentjes kweken in de tuin, vlierbloesems verzamelen, een wesp uit de houtstapel toveren, … Ook mijn zus en broers mogen niet ontbreken in het lijstje: we vormen op onze manier een hechte familie en achter de grappen weet ik dat er wederzijdse appreciatie schuilgaat. Ook Rein hoort hier thuis: de winterse maanden dat we beide achter onze laptop aan het werk waren, jij voor je examens, ik voor dit doctoraat. Ik denk dat er zelden in de familiegeschiedenis zoveel uurtjes na elkaar gewerkt is als toen ;-)

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under the Curve
BCS	Belgian Continental Shelf
Belcoulor	Project coordinated by MUMM focussing on optical remote sensing of coastal waters.
BPI	Bathymetric Position Index
BPNS	Belgian Part of the North Sea
CCC	Concordance Correlation Coefficient
Chl *a*	Chlorophyll *a*
CI	Confidence Interval
ES(n)	Expected Species richness (if the sample was of the smaller size n)
GLS	Generalised Least Squares
HSM	Habitat Suitability Model
IDH	Intermediate Disturbance Hypothesis
MAEE	Mean Absolute Estimation Error
MANUELA	Meiobenthic And Nematode biodiversity Unravelling Ecological and Latitudinal Aspects
Maxent	Maximum Entropy Modelling. More specifically, here it refers to the software developed by Phillips *et al.* (2004) for developing HSMs.
MEE	Mean Estimation Error
MPA	Marine Protected Area
MUMM	Management Unit of the North Sea Mathematical Models and the Scheldt estuary
OK	Ordinary Kriging
OLS	Ordinary Least Squares
p/a	presence/absence
PCA	Principle Components Analysis
PO	Presence-Only
RA	Relative Abundances
RCMG	Renard Centre of Marine Geology at Ghent University
RK	Regression Kriging
RMSEE	Root Mean-Square Estimation Error
ROC	Receiver Operating Characteristic
SA	Spatial Autocorrelation
SBNS	Southern Bight of the North Sea
TSM	Total suspended matter

SAMENVATTING

Het marien milieu staat onder grote druk. Vele processen zoals visserij, zandextractie, klimaatsopwarming, verzuring en de introductie van invasieve soorten hebben een grote invloed op het mariene ecosysteem. De bescherming van het mariene milieu staat wereldwijd op de politieke agenda. Op Europees niveau vertaalt zich dat in de bescherming van gebieden in het kader van de eerder uitgevaardigde Vogelrichtlijn en Habitatrichtlijn. Deze richtlijnen laten toe om waardevolle gebieden af te bakenen, wat resulteert in een Europees netwerk van natuurgebieden, nl. Natura 2000. Het afgelopen decennium werden bovendien twee richtlijnen uitgevaardigd die zich specifiek op het aquatisch en het marien milieu richten: de Kaderrichtlijn Water en de Kaderrichtlijn Mariene Strategie. De implementatie van de Kaderrichtlijn Water en de Kaderrichtlijn Mariene Strategie moet er voor zorgen dat het mariene milieu een goede ecologische status bereikt en/of behoudt. Terwijl alleen de kustwateren (gelegen binnen één zeemijl van de kust) vallen onder de Kaderrichtlijn Water, gelden de doelstellingen van de Kaderrichtlijn Mariene Strategie voor het volledige gebied waarover een Europese staat jurisdictie heeft. De implementatie van het mariene luik van de Vogelrichtlijn en de Habitatrichtlijn vereist de ontwikkeling van een netwerk van mariene beschermde gebieden die rekening houdt met soorten, habitats en ecologische processen. De afbakening van deze waardevolle gebieden wordt bij voorkeur gebaseerd op wetenschappelijke gegevens van onder andere de biotische componenten van het ecosysteem. Op basis van deze gegevens kunnen modellen en gebiedsdekkende kaarten van habitats en de verspreiding van soorten ontwikkeld worden. Het is belangrijk dat deze modellen representatief en betrouwbaar zijn. Daarom moeten potentiële valkuilen bij het modelleren zoals ruimtelijke autocorrelatie, preferentiële staalname en overfitting omzeild worden. Deze kunnen immers aanleiding geven tot het aanvaarden van foute modellen of modellen die slechts representatief zijn voor een klein gedeelte van het gebied. In het kader van voorliggend onderzoek werd op uiteenlopende manieren rekening gehouden met deze problemen. Daartoe werden bestaande technieken gecombineerd of nieuwe technieken ontwikkeld. Voor dit proefschrift werd gefocust op het mariene benthos en meer specifiek op de nematodengemeenschap en twee macrobenthische soorten *Lanice conchilega* en *Ensis directus*. De nematodengemeenschap wordt gekenmerkt door een zeer grote diversiteit op zowel de schaal van enkele vierkante centimeter als op een grote regionale schaal. De factoren die bijdragen aan deze diversiteit zijn reeds onderzocht op het niveau van individueel onderzoek. In het kader van een internationaal samenwerkingsverband (MarBEF, MANUELA) tussen verschillende wetenschappelijke instellingen werd deze informatie recent samengebracht in een overkoepelende databank. Deze databank maakt het mogelijk om de op beperkte schaal waargenomen patronen te toetsen op grote schaal, zowel ruimtelijk als naar de hoeveelheid informatie over de gemeenschappen.

In de inleiding, **Hoofdstuk 1**, wordt een overzicht gegeven van de onderzoeksdoelstellingen van dit proefschrift. Deze doelstellingen situeren zich op twee niveaus: enerzijds wordt onderzoek gedaan naar verbeterde modelleertechnieken die de invloed van modelleeruitdagingen zoals preferentiële staalname, ruimtelijke autocorrelatie en overfitten behandelen, anderzijds tracht dit onderzoek ook inzicht te brengen in de factoren die de nematodengemeenschap en beide macrobenthische soorten beïnvloeden.

Nematodengemeenschappen vertonen een grote diversiteit op kleine schaal, zo zijn 50 verschillende nematodensoorten op een oppervlakte van 10 cm² niet uitzonderlijk. Deze enorme diversiteit is ondermeer toegeschreven aan de mogelijke invloed van lokale soorteninteracties, die er enerzijds toe geleid zouden hebben dat soorten evolueren zodat competitie voor specifieke voedselbronnen minder intens wordt, of waardoor anderzijds soorten minder samen voorkomen dan verwacht wordt op basis van kansberekening. De eerste theorie kan niet gevalideerd worden op basis van mathematische modellen, maar de tweede theorie heeft aanleiding gegeven tot het ontwikkelen van nulmodellen. Deze nulmodellen gaan na of soorten eerder een geaggregeerd of een gesegregeerd patroon vormen dan wat verwacht wordt op basis van toeval. In **Hoofdstuk 2** werden deze nulmodellen toegepast op de nematodengemeenschappen. Voor dit onderzoek werd het wisselalgoritme dat ontworpen werd om nulmodellen te maken, enkel toegepast tussen replica's van één staalname. Dit heeft als voordeel dat de omgevingsvariabelen geen of slechts een beperkte invloed zullen hebben op het eindresultaat. Bovendien werd er nagegaan of de neiging van soorten om elkaar te vermijden eerder optreedt voor soorten binnen hetzelfde voedingstype. Op basis van deze drie nulmodellen kon besloten worden dat nematodensoorten meer samen voorkomen dan wat verwacht wordt op basis van een toevalsverdeling, en dat ze dus eerder een geaggregeerd patroon vormen. Deze patronen kunnen mogelijk gelinkt worden aan het fragmentarisch verspreidingspatroon dat gekend is voor nematodengemeenschappen. Op een kleine schaal van enkele centimeters kunnen immers zeer grote diversiteits- en densiteitsverschillen optreden. Het is dus onwaarschijnlijk dat competitieve exclusie plaatsvindt binnen een replica.

Diversiteit kan uitgedrukt worden op verschillende manieren nl. soortenrijkdom, een gelijkmatige verdeling van de soorten in de gemeenschap (evenness) en de taxonomische diversiteit. Deze aspecten worden niet even sterk beïnvloed door de omgevingsvariabelen. In **Hoofdstuk 3** werd nagegaan in hoeverre deze diversiteitsaspecten kunnen verklaard worden door de omgevingsvariabelen op regionale schaal. Aan de hand van artificiële neurale netwerken (ANN) en op basis van de omgevingsvariabelen en de gekende lokale diversiteit werden verklarende modellen ontwikkeld voor het Belgisch deel van de Noordzee. ANN worden vaak beschouwd als een 'zwarte doos': het verband tussen de verklarende variabele en de afhankelijke variabele is niet duidelijk. Om dit verband toch zichtbaar te maken werden drie methodieken toegepast: de Perturb, de Profile en een aangepaste Profile methode. Ook werd aan de hand van Moran's I nagegaan of de residuelen nog ruimtelijke autocorrelatie vertonen. Uit de modelanalyse blijkt dat

voornamelijk evenness het best verklaard kan worden door de omgevingsvariablen, gevolgd door soortenrijkdom. De taxonomische diversiteit bleek moeilijker te verklaren op basis van de omgevingsvariabelen. Voornamelijk de sedimentkarakteristieken en meer bepaald de zand en slib-kleifractie spelen een grote rol bij het voorspellen van de soortenrijkdom en de evenness van een gemeenschap. De zandfractie vertoont een positieve correlatie met deze diversiteitsaspecten, terwijl de slib-kleifractie een negatieve relatie vertoonde. Ook het minimumgehalte aan gesuspendeerd materiaal (TSM) en het gehalte aan chlorofyl *a* in de waterkolom bleken een negatieve relatie te vertonen met deze aspecten van biodiversiteit. De soortenrijkdom bleek positief gecorreleerd te zijn met het aandeel grind in de zeebodem en de intensiteit van zandextractie. Zandextractie gebeurt voornamelijk in sedimenten met een hoge zandfractie. Dit laatste effect kan dus onrechtstreeks te wijten zijn aan een hoge zandfractie.

Aangezien evenness en soortenrijkdom vrij goed te voorspellen zijn, werd verder onderzocht wat de beste manier is om gebiedsdekkende kaarten te maken van deze diversiteitsaspecten. In **Hoofdstuk 4** werden twee krigingtechnieken met elkaar vergeleken nl. ordinary kriging en regression kriging. Bovendien werd bij regression kriging twee lineare modelleertechnieken met elkaar vergeleken: ordinary least squares (OLS) en generalised least squares (GLS), de laatste is een regressietechniek die rekening houdt met de aanwezigheid van ruimtelijke autocorrelatie. Het best voorspellende model voor de diversiteitsindices ES(25) en S werden bekomen met Regressie Kriging op basis van GLS. Opnieuw bleek dat het aandeel slib-klei en TSM een significante relatie vertonen met de diversiteit. De resulterende kaarten vertonen een zeer lage diversiteit ten Zuiden van de monding van de Schelde en een hogere diversiteit verder verwijderd van de kust.

Naast algemene gemeenschapskenmerken zoals diversiteit, kan ook de potentiële verspreiding van individuele soorten, de zogenaamde habitatgeschiktheidskaarten, voorspeld worden aan de hand van gebiedsdekkende kaarten van de omgevingsvariabelen. In dit kader werd de Maxent software toegepast, die enkel gebruik maakt van aanwezigheidsdata. Dit is te verantwoorden aangezien afwezigheid moeilijk vast te stellen is bij onopvallende organismen zoals nematoden. Bovendien vertonen zij een fragmentarische distributie, dus de afwezigheid van een soort in een staal betekent niet vanzelfsprekend dat de omgeving niet geschikt is voor de soort. Nu blijkt wel dat deze techniek gevoelig is voor preferentiële staalname. Staalnames gebeuren immers niet altijd at random over het volledige gebied en voor alle omgevingen. In **Hoofdstuk 5** wordt dieper ingegaan op het effect van preferentiële staalname op de kwaliteitsparameters van het model, in dit geval de 'area under the curve' (AUC). Deze invloed werd nagegaan met behulp van nulmodellen nl. fictieve soortsmodellen. Deze fictieve soortsmodellen werden gebaseerd op enerzijds willekeurige vindplaatsen verspreid over het volledige gebied en anderzijds vindplaatsen beperkt tot de effectieve staalnameplaatsen. Het vergelijken van duizenden van dergelijke nulmodellen duidt duidelijk de aanwezigheid van preferentiële staalname aan. Daarnaast werd ook de invloed van ruimtelijke autocorrelatie en overfitting op de AUC nagegaan. Uit

de analyse bleek dat de drie effecten aanwezig zijn en een grote invloed uitoefenen op de resulterende AUC. Dat betekent dat in sommige gevallen de drempelwaardes die meestal gehanteerd worden om onderscheid te maken tussen een goed en een slecht model niet toepasbaar zijn. Verder bleek dat soorten die een breed verspreidingspatroon vertonen vaak moeilijk te voorspellen zijn. Op basis van de modellen die beter presteerden dan de fictieve modellen, werden habitatgeschiktheidskaarten afgeleid die terug te vinden zijn in **Appendix 3**. In deze appendix wordt voor elke soort ook de relatieve bijdrage van elke omgevingsvariabele weergegeven. De manier waarop de omgevingsvariabelen bijdragen tot het model is terug te vinden op de bijgevoegde DVD.

Algemeen voorkomende soorten zijn door hun breed verspreidingspatroon en weinig specifieke habitatvoorkeur vaak moeilijker te voorspellen dan soorten die in een welbepaald habitat voorkomen en een beperkt verspreidingspatroon kennen. Toch kunnen deze soorten ook van belang zijn in kader van natuurbehoud omdat zij dankzij hun hoge densiteiten een belangrijke bijdrage kunnen leveren aan de structuur en het functioneren van ecosystemen. In plaats van enkel rekening te houden met de aanwezigheid van een soort, kan het dus interessant zijn habitatgeschiktheidskaarten te maken die weergeven waar het organisme in hogere densiteiten voorkomt. Deze techniek werd in **Hoofdstuk 6** uitgewerkt voor zes nematodensoorten die in sterk fluctuerende relatieve densiteiten voorkomen in de Zuidelijke Bocht van de Noordzee. Uit de analyse blijkt dat het mogelijk is relevante habitatgeschiktheidskaarten op te stellen op basis van drempelwaarden voor de relatieve densiteit of door de relatieve densiteit voor te stellen als individuele observaties.

In **Hoofdstuk 7** werden de voorgenoemde technieken toegepast op twee macrobenthische soorten nl. *Lanice conchilega* en *Ensis directus*. Beide soorten komen in hoge densiteiten voor aan de Belgische kust, maar hun belang is verschillend: *L. conchilega* is een kokerbouwende polychaet die, wanneer hij voorkomt in hoge densiteiten, aanleiding geeft tot de vorming van riffen. Op deze manier oefent deze soort een positieve invloed uit op de macrobenthische diversiteit. *Ensis directus* is een invasieve soort die afkomstig van het Atlantisch kustgebied van Noord-Amerika. Dit schelpdier kan in hoge densiteiten interessant zijn voor de schelpdiervisserij en voor zeevogels zoals de Zwarte Zeeëend.

In **Hoofdstuk 8** wordt het onderzoek ruimer gekaderd. Het potentiële gevaar van modeleeruitdagingen zoals ruimtelijke autocorrelatie, preferentiële staalname en overfitten worden besproken. Ook worden verschillende methodes aangereikt om deze problemen te omzeilen. Daarnaast worden de verschillende theorieën die de ruimtelijke verschillen in biodiversiteit kunnen verklaren, zoals interacties tussen soorten, habitat heterogeniteit, productiviteit en verstoring overlopen en gekoppeld aan de resultaten van dit onderzoek. Tenslotte worden nog enkele potentiële doelstellingen voor toekomstig onderzoek gesuggereerd.

Algemeen kan besloten worden dat de methodieken gebruikt en verder verfijnd tijdens dit onderzoek eenvoudig toe te passen zijn op de andere biotische componenten van het

marien milieu. De bundeling van deze gegevens kan de basis vormen om op een gefundeerd wetenschappelijk wijze te komen tot de afbakening van te beschermen gebieden.

SUMMARY

The marine environment is under strong pressure: many processes such as fisheries, sand extraction, climate change, ocean acidification and the introduction of invasive species are rapidly changing the marine environment. Policy makers worldwide have put the protection of the marine environment on the agenda. On the European level this is translated into the protection of habitats by means of the Habitat Directive and the Birds Directive. These legal instruments allow the designation of nature conservation areas, which will result in a European network of protected areas, known as Natura 2000. Moreover, the past decade two directives focussing on the aquatic and marine environment have been issued: the Water Framework Directive and the Marine Strategy Framework Directive. The implementation of these directives aims at preserving or improving a good ecological status of the marine environment. While the Water Framework Directive aims at the coastal area only (up to one nautical mile off coast), the Marine Strategy Framework Directive aims at the total marine area under the jurisdiction of a European state. The implementation of the marine part of the Habitat Directive and the Birds Directive requires the development of a network of Marine Protected Areas (MPAs) taking into account species, habitats and ecological processes. The delineation of these valuable areas should be based, among other things, on scientific data about the biotic components of the ecosystem. Based on these data, models and area covering maps of habitats and the distribution of species can be developed. It is of major importance that these models are reliable. Therefore, tackling potential modelling pitfalls such as spatial autocorrelation, preferential sampling and overfitting is essential since these issues could result in accepting faulty models or models which are only representative for a small part of the area. In this research, these issues have been addressed in different ways: different techniques have been combined and new techniques have been developed. Here, the focus was on marine benthic species, more specifically the nematode community and two macrobenthic species: *Lanice conchilega* and *Ensis directus*. The nematode community is characterised by large species diversity both on a small scale, within square centimetres, and on a large regional scale. In the past, the factors contributing to this diversity were investigated at the level of individual research. Recently, this information has been gathered in a database in the framework of an international collaboration between different research institutes (MarBEF, MANUELA). These data enable us to test the individual observed patterns on a large spatial scale and with a large amount of data.

In the general introduction, **Chapter 1**, a brief overview of the aims of this research is given. These are situated on two levels: on the one hand, the research focuses on improving the known modelling techniques with an emphasis on preferential sampling, spatial autocorrelation and overfitting; on the other hand, this research attempts to reveal the factors influencing the nematode community and both macrobenthic species under study.

Nematode communities are characterised by a large diversity on a small scale, 50 different species per 10 cm² are not unusual. This high diversity has been ascribed, among other things, to the influence of local species interactions. These species interactions could have led on one hand to the evolution of species in such a way that less competition is to be expected. On the other hand, competing species are expected to co-occur less frequently than expected from a random distribution. The first theory cannot be validated by mathematical models, but the second theory has given rise to the development of specific null models. These models check if species tend to form more segregated or aggregated patterns compared to a random distribution. In **Chapter 2** these null models were applied to the nematode communities. Here, the swapping algorithm used to produce null models was restricted to the replicate samples taken during a sampling event. This removes or strongly reduces the influence of the environmental gradients on the result of the analysis. Segregated community patterns are rather expected between species of the same feeding types. Therefore, the analyses were repeated for the different feeding types. The results of these swapping algorithms indicate the presence of aggregated species, thus nematode species tend to co-occur more often than expected by chance. These patterns could be explained by the patchy distribution known for nematode communities. At the scale of some square centimetres large differences in density and diversity have been observed. Thus, it is unlikely that competitive exclusion is achieved within a sample.

Biodiversity can be expressed in different ways; namely species richness, evenness and taxonomic diversity. These aspects are not equally influenced by the environmental variables. In **Chapter 3** the relation between these diversity indices and the environmental variables is investigated on a regional scale for the Belgian Part of the North Sea. By the use of artificial neural networks (ANNs) and maps of the environmental variables, predictive models of the diversity indices were developed. ANNs are often considered to be 'black boxes': the link between the explaining variables and the dependent variable is not clear. Three techniques were applied for revealing the importance of each variable: the Perturb, the Profile and the Modified Profile method. Moran's I was used to investigate if spatial autocorrelation remains in the model residuals. The model analysis shows that evenness is best explained by the environmental variables, followed by species richness. Pure taxonomic diversity shows high spatial variability and is difficult to model. Especially the sediment characteristics such as the sand and the silt-clay fraction strongly contribute to the species richness and evenness models. The sand fraction reveals a positive effect on the diversity, while the opposite is true for the silt-clay fraction. Also the minimum total suspended matter (TSM) and chlorophyll *a* content show a negative relation with these aspects of biodiversity. The gravel content and the intensity of sand extraction seem to have a positive effect on species richness. Sand extraction typically occurs in sediments with a high sand fraction. Thus the latter effect may be indirectly related to the sand fraction.

Since evenness and species richness are well predicted, we investigated further which technique results in the best area covering maps of these diversity aspects. In **Chapter 4** two

kriging techniques were compared: ordinary kriging and regression kriging. Moreover, for regression kriging two linear modelling techniques were applied: ordinary least squares (OLS) and generalised least squares (GLS). The latter is a regression technique which compensates for spatial autocorrelation. The best models for the diversity indices ES(25) and S are realised by regression kriging combined with GLS. Again, the silt-clay fraction and TSM show a significant relationship with biodiversity. The resulting maps reveal very low diversity to the south of the mouth of the Scheldt estuary. Off coast diversity and evenness are generally higher.

Besides general community characteristics such as diversity, also the habitat suitability of individual species can be predicted based on area covering environmental variables. Here, we used Maxent, a presence-only modelling technique. This is reasonable, since absence is difficult to assess in case of inconspicuous organisms such as nematodes. Moreover, these animals show a patchy distribution, thus absence does not necessary imply an unsuitable environment. However, this technique seems to be more prone to preferential sampling. Preferential sampling occurs when sampling stations are not randomly selected across the area, or when certain habitat types are undersampled. In **Chapter 5** the effect of preferential sampling on the quality parameter of the model the 'area under the curve' (AUC) was verified by means of null models, in this case 'random species' models. These random models were built with randomly chosen stations. These stations were selected in two ways; stations chosen across the entire area, and stations randomly selected from the actual sampling stations. The presence of preferential sampling is revealed by comparing thousands of such null models. Moreover, the influence of spatial autocorrelation and overfitting on the AUC was verified. The analyses clearly show that the three issues strongly affect the AUC of the models. This means that the predefined thresholds between a good and a bad model cannot be applied. Besides, it seems that common species are generally harder to model. The actual species models which perform better than random were further optimised. The resulting habitat suitability maps can be found in **Addendum 3** together with the table of the relative contribution of each environmental variable to each species model. The response curves showing the influence of these variables on the model output can be found on the DVD annexed to this thesis.

Common species have a broad distribution pattern and have less stringent habitat requirements which make them often harder to predict than specialist species with specific habitat requirements. However, common species can be of importance in nature conservation since they can strongly attribute to the structure and functioning of ecosystems when they appear in high densities. Thus, habitat suitability models taking into account the density of the species can be valuable in such cases. In **Chapter 6** different models incorporating information on the frequencies of six nematode species were developed. The analyses show that it is possible to develop reliable maps based on frequency thresholds or by considering the relative abundances as being separate observations of the species.

In **Chapter 7** the above mentioned techniques were applied to two macrobenthic species: *Lanice conchilega* and *Ensis directus*. Both species appear in high densities in the Belgian Part of the North Sea, but their importance is of a different order: *L. conchilega* is a tube-building polychaete, which forms reefs when appearing in high densities. In this way it induces a positive effect on the macrobenthic diversity. *Ensis directus* is an invasive species originating from the Atlantic coast of North-America. High densities of this clam can be of interest to shellfish fisheries or to sea birds, such as the Common Scoter.

In **Chapter 8** the results of this research are put in a broader perspective. The potential pitfalls such as spatial autocorrelation, preferential sampling and overfitting are discussed and different ways to counteract these issues are suggested. Furthermore, different theories explaining the spatial differences in diversity, such as species interactions, habitat heterogeneity, productivity and disturbance are considered and linked to the results of this research. Finally, some recommendations for future research are suggested.

Generally, it can be concluded that the techniques used and refined during this research can easily be applied to other biotic components of the marine environment. Combining these data can be the foundation of delineating conservation areas based on sound scientific knowledge.

CHAPTER 1

GENERAL INTRODUCTION

CHAPTER 1

GENERAL INTRODUCTION

THE OCEANS

A world under pressure

Strong anthropogenic pressure rapidly induces changes in the oceans. Oceans are directly exploited by fisheries, gravel and mineral extraction, construction of oil platforms and wind mill farms, the use of wave and tidal energy, installation of pipe lines and distribution cables (Halpern *et al.*, 2008). Moreover, other non-intentional effects such as organic and inorganic pollution, the emergence of invasive species, ocean acidification and climate change operate. As a result, over 40% of the world's oceans are heavily affected by human activities and few, if any areas remain untouched (Halpern *et al.*, 2008) (Fig. 1.1).

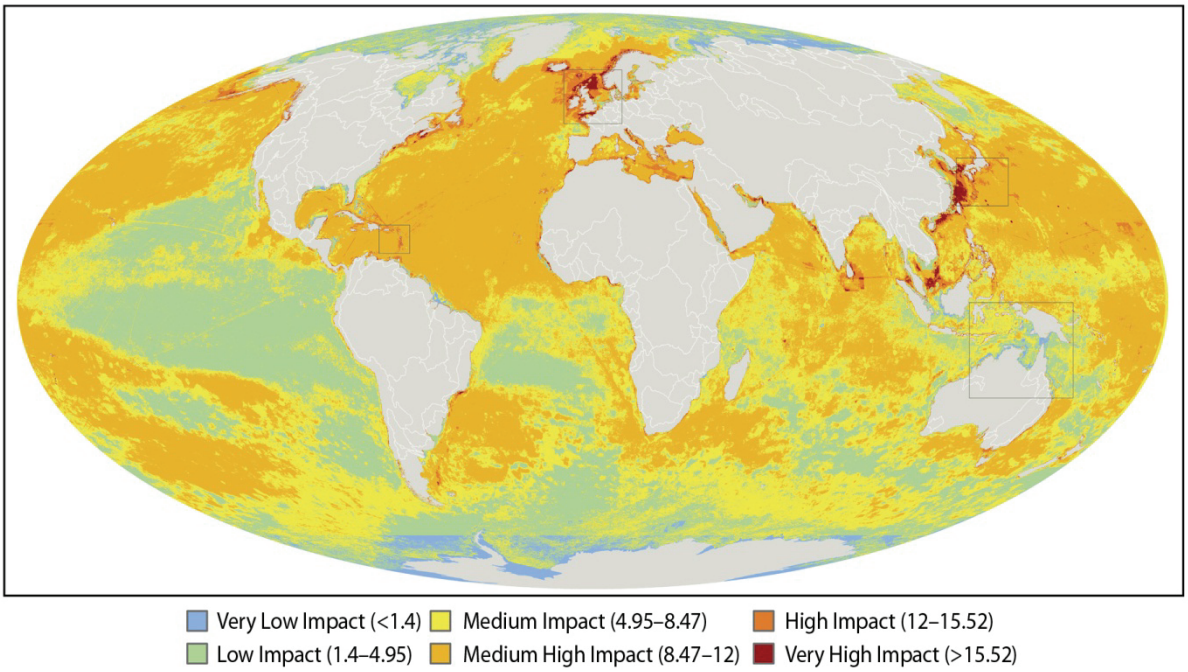


*Fig. 1.1. Global Map of Human Impacts to Marine Ecosystems (*from *Halpern* et al.*, 2008)*

Some of these impacts strongly affect local biodiversity: numerous publications report on the impacts of fisheries (Pauly *et al.*, 1998; EC, 2009; FAO, 2010; Froese *et al.*, 2010; Shephard *et al.*,2010) and climate change related effects (Caldeira and Wickett, 2003; Edwards and Richardson, 2004; Perry *et al.*, 2005). Although marine extinctions are not easily uncovered, it has been shown that regional ecosystems are rapidly losing species and functional groups (Worm *et al.*, 2006).

Global awareness that our ecosystems need protection was raised at the UN Convention of Biological Diversity in 1992 in Rio de Janeiro. The main outcome of this convention consists of two key documents: the Rio declaration and the Agenda 21. One of the objectives in Agenda 21 is to improve the conservation of biological diversity and the sustainable use of our biological resources. A significant reduction of biodiversity loss by 2010 was advocated at the 2002 Convention on Biological Diversity at The Hague (Butchart *et al.*, 2010). This target was further refined in 2006 (CBD, 2006). Specifically, at least 10% of each of the world's ecological regions should be effectively conserved and areas of particular importance to biodiversity should be protected (Toropova *et al.*, 2010). Targets focusing on specific biomes such as the conservation of marine and coastal areas were specified. Marine Protected Areas (MPA's) should be established to protect species, maintain productivity, and preserve nursing grounds for fishes or to protect complete ecosystems (CBD, 2006). A drawback however is that the degree of protection and the activities allowed in these MPA's are not well defined and may vary considerably (Toropova *et al.*, 2010). At a European level, the European Commission issued an action plan to halt the loss of biodiversity by 2012 for the marine environment (EC, 2006). Therefore, the European Commission agreed on two directives to oblige member states to designate MPA's in the frame of the Natura 2000 network:

- The European Birds Directive (2009/147/EC) aims to protect all European wild birds and the habitats of listed species, in particular through the designation of Special Protection Areas (SPA). In relation to the marine environment this is translated in the designation of habitats of sea birds.
- The Habitats Directive (92/43/EEC) focuses on creating a network of Special Areas of Conservation (SAC).

Other European strategies which can be deployed to protect the marine environment are:

- Integrated Coastal Zone Management (2002/413), which promotes the sustainable management of coastal zones, while balancing environmental, economic, social, cultural and recreational objectives all within the natural limits.
- The Water Framework Directive (WFD) (2000/60/EC), which obliges the member states to achieve good status of all water bodies, including marine waters up to 1.85 kilometre off coast by 2015.
- The Marine Strategy Framework Directive (MSFD) (2008/56/EC), which aims to protect the EU marine environment in an effective way. A good environmental status of the EU marine waters should be achieved by 2020. Within this directive European Marine Regions will be established based on geographical and environmental criteria. To meet the requirements of this directive each EU member State is committed to develop strategies for their marine waters.

The decisions taken in the framework of these legal instruments should be based on sound scientific knowledge and it has been advocated that in the marine management, habitat suitability maps, biological valuation maps and biodiversity maps are suitable instruments

for communication with policy makers and marine managers (Derous *et al.*, 2007; Degraer *et al.*, 2008; Fraser *et al.*, 2008; Willems *et al.*, 2008).

A world to discover

Whereas land biodiversity patterns and the factors influencing this biodiversity are known for numerous taxa, our understanding of global biodiversity in the sea is more limited (Tittensor *et al.*, 2010). The ocean surface hides a diverse world which can only be discovered by sampling or diving at sea. This is an expensive and labour intensive task since adequate equipment and research vessels are needed. Therefore, it is important not only to collect but also to preserve valuable scientific data. In the 20th century, a lot of effort was done to discover the unknown marine world: sampling campaigns were organised, thousands of species were described, and information on species diversity and communities in the oceans was gathered. This information was often fragmented, but recently a lot of effort has been put into compiling this information in databases (Table 1.1). This is achieved by gathering researchers in regional or global consortia such as MarBEF and Census of Marine Life. In this way a better insight in the biodiversity of the oceans and the world as a whole can be achieved. This wealth of information gives new opportunities to explore and study the biodiversity of the oceans.

Project name	Database content	Number of species*	Website
CBOL iBOL	DNA barcode sequences of marine and terrestrial species	93 543 Target 2015: 5 000 000	http://www.barcodeoflife.org/ http://ibol.org/
Encyclopedia of Life	Marine and terrestrial species	Target: 1 900 000	http://www.eol.org/
Catalogue of Life	Marine and terrestrial species	1 333 403	http://www.catalogueoflife.org/col/
GBiF	Marine and terrestrial species	919 873	http://www.gbif.org/
Itis	Marine and terrestrial species	518 498	http://www.itis.gov/
WoRMS	Marine species	207 762	http://www.marinespecies.org/
OBIS (Census of Marine Life)	Marine species	114 879	http://www.iobis.org/
ERMS	Marine species	31 000	http://www.marbef.org/data
NeMys	Marine and terrestrial species	14 945	http://nemys.ugent.be
MarBOL	DNA barcode sequences of marine species	6 199	http://www.marinebarcoding.org/
MANUELA	Marine meiobenthic species	1 250	http://www.marbef.org/projects/Manuela/

*Table 1.1. Examples of biodiversity databases available on the web (*on February 1st 2011).*

Need for accurate models to understand and protect the marine habitat

The disclosure of this large amount of data helps in understanding the complexity of marine ecosystems, and modelling may serve both in the understanding and protection of this habitat. Mathematical models can reveal the factors influencing the biodiversity and may in this way contribute to the understanding of the structure of marine communities. On the other hand, there is a need to delineate areas which need to be preserved to protect the diversity and resilience of the seas. The designation of these protected areas should be based on sound scientific models. Delineating a protected area involves spatial multi-criteria-analysis (Villa *et al.*, 2002; Pomeroy *et al.*, 2004) encompassing a vast number of criteria such as geological features, diversity and composition of benthic and pelagic communities, potential human use of the area (tourism and fisheries), the protection of specific species, pollution status of the area, and many more (Villa *et al.*, 2002). Biodiversity maps and habitat suitability maps should thus be considered during the decision process. Given the importance of these models, it is crucial that these models are beyond discussion and all potential modelling pitfalls should be tested for, and avoided. In this way, these models can really contribute to understanding and preserving biodiversity. In this thesis, we investigate how the construction of erroneous or non-significant models caused by spatial autocorrelation and preferential sampling can be avoided.

BIOLOGICAL DATA

Here, we will focus on the marine benthos, with emphasis on free-living nematodes and the distribution of two macrobenthic species: an ecosystem engineer, the polychaete *Lanice conchilega* and an invasive species for the North-East Atlantic Area, the bivalve *Ensis directus*. The marine benthos encloses all those species associated with the sea floor. This diverse community can be divided in different groups according to size, location and type. Generally, benthic organisms are grouped according to size: macrobenthos (i.e. benthic organisms retained on a sieve with 1 mm mesh size), meiobenthos (i.e. metazoan organisms passing a 1 mm sieve but retained on a sieve of 38 µm) and microbenthos, (i.e. microscopic benthos passing a 38 µm sieve). The lower size of 38 µm for the meiobenthos may vary: sizes of 32 µm, 44 µm and 63 µm are also applied (Giere, 2009).

Nematodes

In this thesis we focus on a phylogenetic group within the meiobenthos: the Nematoda. The nematodes or roundworms are the most diverse phylum of pseudocoelomates, and one of the most diverse of all animals. Over 26 000 species have been described, of which over 16 000 are parasitic and more than 4 000 are free-living marine nematodes (Hugot *et al.*, 2001) (Table 1.2). It has been estimated that the total number of nematode species might be approximately 500 000 (Hammond, 1992) or even 1 000 000 (May, 1988) and they could be the second most diverse group after the Arthropoda (Hugot *et al.*, 2001). Nematodes are not

only highly diverse, but are often complex and biologically specialised metazoans (De Ley, 2006). Free-living nematodes represent a high 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. In meiofaunal samples, nematodes are usually the dominant taxon both in abundance and in biomass (Giere, 2009).

Nematode communities are very useful as indicator organisms (Bongers and Ferris, 1999; Kennedy and Jacoby, 1999; Geetanjali *et al.*, 2002) in the assessment of sediment quality and pollution status of the environment (Schratzberger *et al.*, 2000a) for numerous reasons: they show a wide distribution from pristine to extremely polluted habitats, they do not rapidly migrate from stressful conditions; they respond rapidly to disturbance and enrichment; and they show a clear relationship between structure and function which can be deduced from the mouth cavity and the pharynx (Bongers and Ferris, 1999). In addition, 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; Schratzberger *et al.*, 2000a; Vanaverbeke *et al.*, 2011).

Life style	# of species
Free-living marine	4 070
Free-living terrestrial	6 610
Plant parasites	4 110
Invertebrate parasites	3 500
Vertebrate parasites	8 360
Total	26 650

Table 1.2. Number of described nematodes (Hugot et al*., 2001).*

Moreover, biodiversity loss of nematode communities might be associated with exponential reduction of the ecosystem function (Danovaro *et al.*, 2008). Free-living nematodes fulfil many different functions in the sediment. Wieser (1953) divided nematodes in four trophic groups according to the shape of their buccal cavity. Selective and non-selective deposit feeders, epigrowth feeders and predators/omnivores were discerned. However, this classification is a simplification of the complex and diverse feeding patterns in nematodes. Some species are facultative predators or feed on ciliates (Moens and Vincx, 1997); other species show switches in feeding behaviour, depending on the available food (Giere, 2009) or the ontogenetic age (Lorenzen, 2000). Most nematodes are nowadays considered selective feeders (Giere, 2009). They can selectively differentiate between prey organisms (Moens *et al.*, 2000) and even between bacteria (Moens *et al.*, 1999). It has been postulated that the high diversity within a nematode community is caused by this food partitioning and resulting niche separation (Heip *et al.*, 1985; Moens and Vincx, 1997; Moens *et al.*, 1999).

These characteristics make the Nematoda a perfect phylum to study community patterns on a broad and local scale.

Macrobenthic species

In this thesis we also focus also on two macrobenthic species: *Lanice conchilega* and *Ensis directus.*

Lanice conchilega (sand mason) is a tube building polychaete and is a species with a wide spread bathymetrical (0-1900 m) and geographical range (Hartmann-Schröder, 1996). Conservation of the species as such is therefore not the main issue here. However, the habitat built by dense aggregations of the species is considered to be a reef (Rabaut *et al.*, 2009). The species changes its direct environment considerably (Rabaut, 2009) as it is considered to be an important ecosystem engineer (*sensu* Jones *et al.*, 1994). The worm builds linear tubes consisting of coarse sand grains cemented with mucus (Jones and Jago, 1993) which can reach a diameter of 5 mm and a length of 65 cm (Ziegelmeier, 1952). The tube is located mainly in the sediment, and only one to four centimetres protrude in the water column. This species has the ability to build dense aggregates and patches with more than 1500 ind.m^{-2} are not uncommon (Zühlke, 2001). These aggregations change the local sedimentary and hydrodynamic environment and the tubes themselves compact the sediment and increase the rigidity (Jones and Jago, 1993). This altered habitat induces changes in the benthic community, resulting in an increase of both macrobenthic abundance and diversity (Callaway, 2006; Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008). Moreover, the *L. conchilega* reefs have a high functional value (Godet *et al.*, 2008) and are related to higher densities of juvenile flatfish such as *Pleuronectes platessa* (plaice) (Rabaut *et al.*, 2010). Recent research indicated the value of the species as a bio-irrigator, which pumps oxygen in the sediment. This mechanism can contribute to the mineralisation and denitrification process in the sediment (Braeckman *et al.*, 2010) and creates an extended habitat for nematodes (Braeckman *et al.*, 2011).

Thus, *L. conchilega* can be considered to be a valuable species in a conservation context (Van Hoey, 2006; Godet *et al.*, 2008; Rabaut *et al.*, 2009) and the development of habitat suitability models for this species has been advocated (Rabaut, 2009).

Ensis directus (Atlantic jackknife, American jackknife clam or razor clam) is a large edible bivalve species. It is indigenous to the Atlantic coast of Canada and North-America and prefers muddy, fine sand with small amounts of silt (Beukema and Dekker, 1995; Kennish *et al.*, 2004) and is found in the intertidal or subtidal zones (Mühlenhardt-Siegel *et al.*, 1983; Swennen *et al.*, 1985). It can burrow very quickly (Swennen *et al.*, 1985), and is also able to swim (Drew, 1907). It is probably introduced into Europe as larvae in ballast water of a ship crossing the Atlantic around 1978 (von Cosel, 1982). The first strong year class occurred in the German Bight in 1979 (von Cosel, 1982). Since then it has spread across the Dutch and Belgian coast. The first Belgian observations date from 1987 (Kerckhof and Dumoulin, 1987) and since its arrival, it has become the bivalve with the highest biomass in several areas along the coast (Tulp *et al.*, 2010). The species can occur in high densities (i.e. bivalve banks) and densities of 1000-2000 ind.m^{-2} are not uncommon (Armonies and Reise, 1999; Tulp *et al.*, 2010). These banks show a patchy distribution, but these patches are not permanent and

in Europe prominent events of mass mortality in late winter or early spring have been observed (Armonies and Reise, 1999). The potential distribution of this invasive species and its potential harmful effect on the natural community in Europe is largely unknown. This invasive species may compete for space and food with indigenous species and dense populations may change the community structure of the benthic fauna (Gollasch *et al.*, 1999). Therefore, habitat suitability modelling can be useful for creating insight in the ecology and the possible distribution of the species. Also, fisheries have shown interest in fishing this bivalve and there might be a link between sea ducks (i.e. Common Scoter (*Melanitta nigra*)) and high densities of *E. directus* (Houziaux *et al.*, 2010). Therefore, distribution maps reflecting densities will be created by applying geostatistics.

STUDY AREA

The study area changes throughout the thesis since different databases and different subsets of these databases have been used. However, the main areas of interest are the continental shelf area of the Southern Bight of the North Sea (referred to as SBNS) (Fig. 1.2) and the Belgian Part of the North Sea (BPNS)[1].

The North Sea is a shallow sea located between Great Britain, Scandinavia, Denmark, Germany, the Netherlands and Belgium. The SBNS is delimited to the North by the thermal stratification of the water column during summer (around 54°N) and to the South by the Strait of Dover between Great Britain and France. The area has a maximum depth of about 54 m (Fig. 1.2) and is characterised by strong semi-diurnal tidal currents (up to about 1 $m.s^{-1}$) and frequent strong winds. This results in a well mixed water column throughout the year (Lee, 1980). The area is characterised by a complex system of sand banks which follow the residual current and are therefore oriented parallel to the coast (Muylaert *et al.*, 2006). The net bottom shear stress is directed to the North (Pingree and Griffiths, 1979).

The seabed sediments consist mainly of fine to medium sands (125-500 µm) (Verfaillie *et al.*, 2006). The Eastern part of the Belgian coast is characterised by high concentrations of silt-clay (Fig. 1.2). The origin and the formation of these silt-clay deposits in front of the coast are explained by the neap-spring tidal cycles and the different sources of Suspended Particulate Matter (SPM). These sources are mainly the erosion of exposed clay layers and the import of SPM from the Strait of Dover (Fettweis and Van den Eynde, 2003).

The SBNS receives carbon and nutrients from river inputs (mainly from the Rhine, Meuse and Scheldt), atmospheric deposition, and exchanges with the Atlantic Ocean through the English Channel (Baeyens *et al.*, 2007). The sources of nutrients and carbon are mostly linked to anthropogenic activities (agriculture, industries, domestic wastewater) (Baeyens *et al.*, 2007). These nutrients give rise to pronounced phytoplankton blooms in spring and late summer. The spring bloom is dominated by two major phytoplankton groups, diatoms and *Phaeocystis* (van der Zee and Chou, 2005). Chl *a* deposition on the sea floor is directly

[1] In Chapter 2 the BPNS is referred to as the Belgian Continental Shelf (BCS)

related to these phytoplankton blooms (van Oevelen *et al.*, 2009). The organic matter in the water column can enter the sediment through physically mediated input (advective injection and passive deposition), or by benthic organisms which actively filter and deposit organic matter from the water column on and into surface sediments (Kautsky and Evans, 1987; Kotta *et al.*, 2005). Mineralisation of this newly settled organic matter can induce hypoxic or even anoxic conditions in the sediment (Graf, 1992).

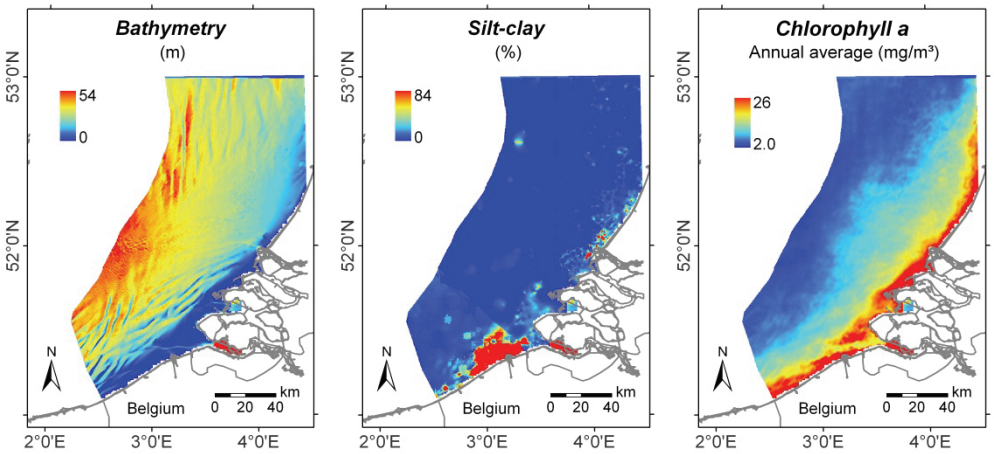


Fig. 1.2. Bathymetry, silt-clay fraction of the sediment, and average chlorophyll a *in the water column of the Southern Bight of the North Sea*[2]

BIODIVERSITY

Biodiversity indices

Biodiversity is defined by Glowka *et al.* (1994) as '*The variability among living organisms from all sources including, among other things, terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are a part; this includes diversity within species, between species and of ecosystems.*' Biodiversity can be expressed in many different ways: it can be linked to species richness, evenness, taxonomical diversity, genetic diversity, functional diversity or other features of the species community. The most common expression of biodiversity is the number of different species in a given area (species richness). However, this estimate of biodiversity is strongly influenced by sample size, and a number of statistical techniques have been developed to correct for this: estimators for total species richness (Chao, 1984; Chao, 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). Depending on the scale α-, β- and γ-diversity can be distinguished: α-diversity is the biodiversity within a particular area (e.g. within the

[2] source: Renard Centre of Marine Geology (RCMG, www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management for the oceanographic and sedimentological data.

area of a core sample). β-diversity is a measure of biodiversity which compares the species diversity between ecosystems. This involves comparing the number of taxa that are unique to each of the ecosystems and gives a view on species turnover across habitats. γ-diversity refers to the total biodiversity over a large area or region (Whittaker, 1972). The 'intrinsic' diversity of a community is given by its α-diversity. Thus, an area with higher α-diversity may be considered more important than one with lower α-diversity values, for conservation purposes (Hernández-Stefanoni and Ponce-Hernandez, 2004). But the contribution of an area to the overall γ-diversity is also depending on the β-diversity. Therefore, areas with lower α-diversity may still be important to conservation management because of their contribution to the total diversity of the area. Developing an additional map with an estimation of the β-diversity would be interesting from a conservational point of view (Samson and Knopf, 1982). However, available approaches to predict β-diversity are hampered by the twofold scale dependence of β-diversity, owing to the size of sampled units as well as their mutual distances (Feilhauer and Schmidtlein, 2009). Moreover, for the SBNS more species have a sample specific name (e.g. *Araeolaimoides* sp.1 MV, with MV the reference to the data supplier) than there are species with an accepted name. On average about 17% of the species in a sample are sample specific. Thus, these species can be used to estimate the α-diversity of a sample, but they cannot be used for estimating β-diversity. Moreover, the maps of the environmental variables should be transformed in gradient maps which will further increase the error on the model. For these reasons an error proliferation on the prediction of β-diversity might be expected. Therefore, to keep the error rate as low as possible, we focussed on α-diversity of the core samples.

Biodiversity patterns

Local or regional variation in biodiversity has fascinated many researchers. Which processes and factors may explain those differences in biodiversity? Darwin described evolutionary processes and natural selection as the driver behind species differentiation (Darwin, 1876). The distribution of these species is however not homogeneous and the spatial and temporal variation in benthic biodiversity reflects not only evolutionary processes but also ecological processes which operate at different spatial and temporal scales (Levin *et al.*, 2001). Many hypotheses have been postulated to offer an explanation for these spatial variations in biodiversity (Table 1.3). There is a considerable overlap between these theories, and some operate on a large temporal and spatial scale, some on a small scale. These small-scale processes are hierarchically embedded in processes taking place on a larger scale. The local processes include competition, exclusion, facilitation, resource partitioning, disturbance of the physical environment and physiological tolerance (Etter and Mullineaux, 2001). On a larger regional scale, environmental factors are important for structuring benthic communities (Fig. 1.3).

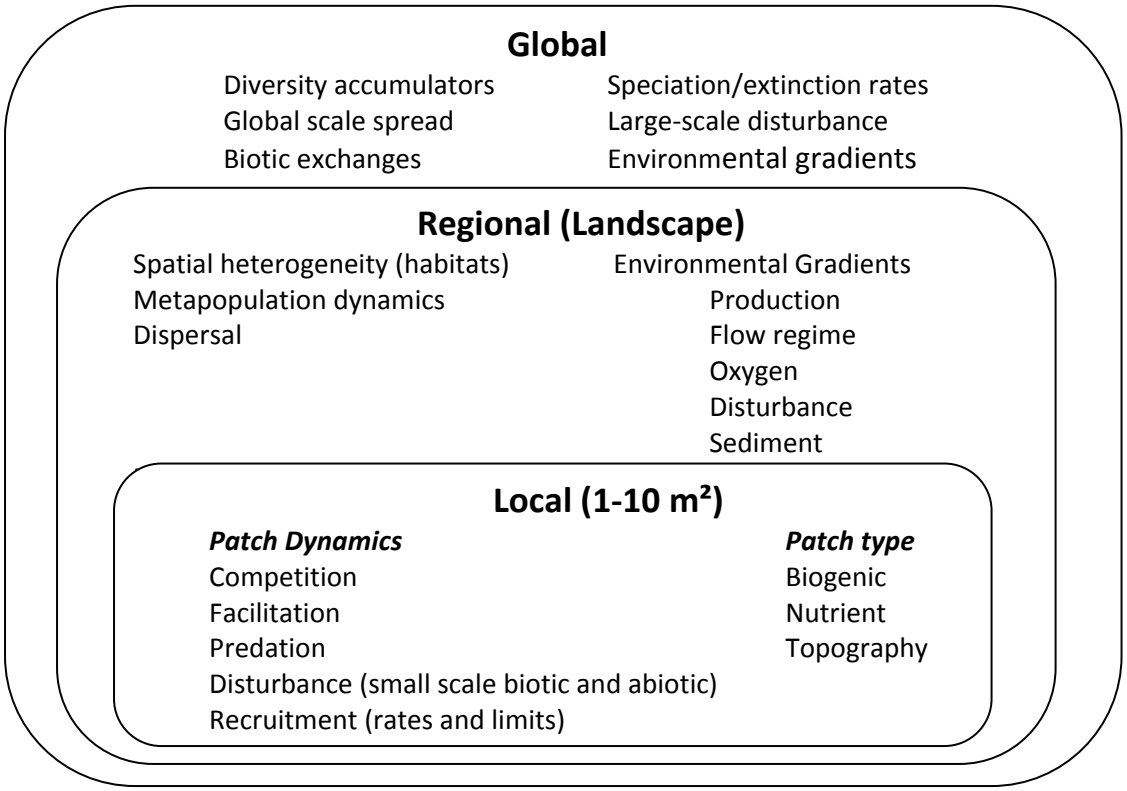


*Fig. 1.3. Processes regulating species diversity at local, regional and global scales (*from *Levin* et al.*, 2001)*

their response to environmental factors (Ricklefs and Shluter, 1993). On a large scale, metacommunities are formed: metacommunities are local communities connected by the dispersal of multiple species (Hubbell, 2001). Leibold *et al.* (2004) suggested four types of metacommunities: the species sorting, source-sink dynamics, the neutral model and patch dynamics type. Here, research on metacommunities is impaired, since little is known about the dispersal capacities of nematode species. Nevertheless, we touch upon two aspects in the discussion: patch dynamics and species sorting.

In this study, the importance of both local and regional factors on biodiversity will be addressed: local processes involve species interactions and will be investigated by null models involving 'species assembly rules'. On a larger scale, environmental factors are assumed to be important. The influence of these abiotic factors on biodiversity and species distributions will be investigated with techniques such as artificial neural networks, generalised least squares and habitat suitability modelling.

Species assembly rules

Species are often unevenly distributed in space and form distinct species communities. One of the fundamental questions in ecology is how unexpectedly strong species associations are created in nature. Evolutionary ecologists are concerned about the question: how do speciation and extinction form a species pool, while community ecologists focus rather on

the question: how are communities formed from this species pool (Weiher and Keddy, 2001). The latter question is addressed by the 'assembly rules'.

Hypothesis	Description
Competition (Dobzhansky, 1950; Dayton, 1971; Grassle and Sanders, 1973; Diamond, 1975)	Species evolve through competition, occupy their own niche and develop therefore specific morphology.
Biological disturbance hypothesis (Paine, 1969; Dayton and Hessler 1972)	Predator-controlled diversity reduces abundance of competitors, maintains resource availability and prevents competitive exclusion.
Intermediate disturbance hypothesis (Connell, 1978) and the dynamic equilibrium (Huston, 1979)	Diversity is maximised at an intermediate frequency and/or magnitude of disturbance
Habitat heterogeneity hypothesis (MacArthur and MacArthur, 1961; Bazzaz, 1975)	Spatial heterogeneity leads to faunal diversification. Specialisation arises through microhabitat exploitation and niche partitioning.
Productivity-diversity hypothesis (Connell and Orias, 1964; Yount, 1956; Rex, 1981)	Positive (Connell and Orias, 1964), negative (Yount, 1956) and optimum relations (Rex, 1981) between productivity and diversity have been found.
Island biogeography (MacArthur and Wilson, 1967)	Isolated islands foster the evolution of new species.
Stability-time hypothesis (Connell, 1978)	Stressed and young environments will have less species than old and/or stable environments.
Theory of climatic stability (Klopfer, 1959)	Regions with stable climates allow the evolution of finer specialisations.
Historical explanation (Jackson, 1992)	Past geological events (e.g. sea level rise) may explain present diversity patterns.
Latitudinal diversity hypotheses (Fischer, 1960; Rohde, 1992; Rex *et al.*, 1993)	Diversity decreases with increasing latitude. Several hypotheses have been postulated to explain this pattern.

Table 1.3. Selection of hypotheses and influences explaining spatial differences in biodiversity.

There are mainly two confronting ideas explaining the assembly of communities: the island paradigm and the trait-environment paradigm (Weiher and Keddy, 2001). The island paradigm relates the structure of communities to dispersal, competition, immigration and extinction, while the trait-environment paradigm considers the environmental factors as the main driving force in structuring communities. As mentioned before, the influence of the environment on diversity and the occurrence of species are in this thesis investigated on a larger geographical scale. Therefore, in this thesis, the term 'species assembly rules' will relate to the 'island paradigm' and thus species interactions. This idea was first introduced by Diamond (1975). He argued that interspecific competition between species occupying similar niches results in a non-random pattern of species distributions. More specifically, some species pairs may never be found together due to competitive exclusion, forming a perfect checkerboard pair. His rules have been strongly disputed and there has been a proliferation of studies promoting, refuting and testing these ideas (Connor and Simberloff

1979; Diamond and Gilpin, 1982; Weiher and Keddy, 1995; Bell, 2000; Weiher and Keddy, 2001; Hubbell, 2001; Bell *et al.*, 2006; Purves and Turnbull, 2010). It has been shown that neutral factors such as birth, death, random dispersal and the total number of organisms in the community may even result in these non-random patterns (Bell, 2000; Hubbell, 2001). However, recent work shows that neutral processes alone cannot explain the observed community patterns (Bell *et al.*, 2006; Purves and Turnbull, 2010).

Most of the community assembly research concentrated on terrestrial studies (Gotelli and McCabe, 2002; Ribichich, 2005) or marine macrobenthos (Pagliosa, 2005). In contrast, Nematoda have received considerably less attention. The main advantages of using nematodes lie in their resistance to disturbance and the vast number of species found in small volumes of sediment.

While environmental variables structure nematode communities on a large scale (Vanaverbeke *et al.*, 2011), species interactions may become important on a smaller scale (Joint *et al.*, 1982; Steyaert *et al.*, 2003). Therefore, we focussed our search for assembly rules on repeated samples at the same location at the same moment in time (further referred to as 'replicate samples'). Replicate samples are generally collected within a small area where differences in environmental variables are small relatively to the environmental differences on a larger scale. Thus, the presence of non-random communities is tested on replicate samples with the use of null models.

It is important to note that the goal of these theoretical models is only to recognise non-random community patterns (Gotelli and Graves, 1996). Revealing the cause of these non-random patterns can only be established by experimental set-ups (Gotelli and Graves, 1996; Gotelli and McCabe, 2002).

Biodiversity and the environment

Levin *et al.* (2001) described six main factors structuring biodiversity in marine benthic environments: boundary constraints, sediment heterogeneity, productivity and food supply, bottom-water oxygen concentrations, sea currents and catastrophic disturbance. Their overview focused mainly on deep-sea macrobenthic species, but studies on the nematode communities of shallow seas revealed that many of these factors are important as well for structuring nematode communities (Heip *et al.*, 1985; Vanreusel, 1990; Vincx *et al.*, 1990; Steyaert *et al.*, 1999; Vanaverbeke *et al.*, 2002; Vanaverbeke *et al.*, 2011).

As mentioned before, biodiversity can be expressed in different ways. In this study we focused on diversity indices related to species richness, evenness and taxonomic richness.

Choice of modelling technique

The relationship between the diversity aspects and the environment may be complex and difficult to model and commonly used mathematical models may be inadequate (Lek and Guégan, 1999). In modelling different approaches can be used depending on the aim of the

model. Multiple linear regression (MLR) is one the most frequently used predictive methods in ecology. The popularity of MLR and linear models in general lies in its ease of use and its capacity to give explanatory results, as the coefficients of the environmental variables provide straight-forward information about their relative importance and they can give some measures of confidence about the estimated coefficients. However, linear models are based on the *a priori* selection of suitable functions and algorithms, such as linear relations in traditional linear modelling or a link linear relation in generalised linear models (Park and Lek, 2005). Therefore, linear models cannot deal with non-linear relationships between dependent and independent variables, unless complex data transformations are performed (Gevrey *et al.*, 2003). Gevrey *et al.* (2003) stressed another drawback of linear regression techniques: only variables with statistically significant coefficients are analysed, for which it may lack resolution. That is why the use of more complex supervised learning techniques is justified in ecology where the relationships between variables are principally non-linear (Lek and Guégan, 1999; Gevrey *et al.*, 2003). In data mining, a large range of modelling techniques are available: support vector machines, classification and regression trees, random forests, artificial neural networks (ANNs) and many more (Izenman, 2008). ANNs are known as powerful computational tools. They can be used for many purposes: classification, pattern recognition and modelling based on empirical data. Most of these tasks can also be performed by conventional statistics. However, ANNs often provide a more effective way to deal with problems that are difficult, if not intractable, for traditional computation (Park and Lek, 2005). As mentioned before, linear models are based on the *a priori* selection of suitable functions and algorithms. For ANNs no such *a priori* selection is needed. ANNs with a single layer can approximate any mathematical function and they can treat complicated problems, even if the data are imprecise and noisy. Their superior modelling capacity is also apparent from the 35 papers we analysed: in 34 out of 35 papers ANNs outperformed conventional linear models. Moreover, their implementation is not precluded by the theoretical distribution shape of the data (Bishop, 1995). If enough data is available and the architecture is properly selected, ANNs provide optimal solutions for any relation between the dependent and the independent variables. In this way unknown or unsuspected relations can be revealed. The relation between the sediment characteristics and the diversity of nematode communities has been widely reported, however the relationship on a large scale was not yet established. Moreover, our goal was to find out which aspects of biodiversity are best explained by the large scale environmental variables. To allow a fair comparison between the different diversity indices, a powerful and flexible tool, which is not subject to modelling constraints, was needed. An additional advantage of using ANNs is that the optimisation process to find an optimal architecture can be automated. This is especially appropriate when several dependent variables need to be modelled. However, a drawback of the flexibility of the ANNs is that the models are prone to overfitting. Therefore, special attention was paid to find the optimal network which does not overfit. Moreover, ANNs are often seen as 'black boxes' which means that the contribution of each input variable to the model output is hidden in the model layers and is difficult to disentangle from the network

(Lek *et al.*, 1996a). However, to reveal the contribution of each environmental variable to the resulting models, we applied three methods: two known in literature: the Perturb method (Yao *et al.*, 1998; Gevrey *et al.*, 2003) and the Profile method (Lek *et al.*, 1995, 1996a, b; Gevrey *et al.*, 2003). A third technique was developed to check the validity of the previous two.

MAPPING

Mapping biodiversity of regions

Biodiversity may vary considerably across and within regions. At this scale, environmental influences contribute to these biodiversity differences. The relation found between the environmental variables and the biodiversity can be used to create diversity maps by applying geostatistics. Geostatistics is a branch of statistics that allows the estimation of the values of a variable of interest at non-sampled locations. Although this approach is related to interpolation methods, it extends far beyond simple interpolation problems (Van Meirvenne, 2007). It consists of a collection of mathematical techniques dealing with the characterisation of spatial phenomena. The technique was originally developed to predict probability distributions of ore grades for mining operations (Matheron, 1963). Currently, geostatistics are applied in diverse disciplines including petroleum geology, oceanography, geography, forestry, environmental control and landscape ecology. More recently, it is also applied in the marine environment to map distribution patterns of marine species (Mello and Rose, 2005; Rios-Lara *et al.*, 2007) and biodiversity patterns (Reese and Brodeur, 2006).

Particularly interesting are the hybrid interpolation techniques. They rely on two techniques (a) interpolation relying on point observations of the spatial variable; and (b) interpolation based on regression of the target variable on spatially exhaustive environmental variables (maps) (Hengl *et al.*, 2007). One of these techniques is known as regression kriging (Hengl *et al.*, 2004). First a regression of the dependent variable with the environmental variables is applied and then it uses simple kriging to interpolate the residuals from the regression model. This allows the use of any regression method to correlate the dependent variable with the environmental variables (Hengl *et al.*, 2007).

Mapping a species' potential distribution

The diversity of a taxon across a region gives an estimation of the species-rich areas. However, this does not supply any information on the community composition. The presence of a single species may be important and reveal underlying structures. Especially keystone, indicator or umbrella species may hold more information about the community. Other species such as *Lanice conchilega* contribute to the diversity of the area by modifying the habitat as a habitat engineer (Callaway, 2006; Rabaut *et al.*, 2007; Van Hoey *et al.*, 2008) and forming reefs, while the distribution of invasive species such as *Ensis directus* may be of

interest to nature conservationists and fisheries. In that case habitat suitability models (HSMs) reveal information on where the species can potentially be found.

Choice of modelling technique

Different modelling techniques have been developed to estimate the potential habitat of a species. Depending on the type of input data, two types of HSMs can be discerned: those based on presence/absence data and HSMs based on presence-only data. Presence/absence data are commonly used for HSMs, but this includes some presumptions about the information. Often the absence of a species is not 100% sure for different reasons 1) not all organisms in a sample are identified, 2) species show patchy or ephemeral distributions and may not be present at the time of sampling although the habitat is suitable for the species, or 3) the species may not have obtained its full range because of a disturbed environment or because it is an invasive species. In such cases presence-only modelling techniques are preferred. Several presence-only modelling techniques have been developed: Bioclim (Nix, 1986), Domain (Carpenter *et al.*, 1993), GARP (Stockwell and Peters, 1999), Maxent (Phillips *et al.*, 2004) and ENFA or Biomapper (Hirzel *et al.*, 2002). Maxent has proven its better predictive capacities compared to other presence-only modelling techniques in several independent cases (Hernandez *et al.*, 2006, 2008; Hijmans and Graham, 2006; Pearson *et al.*, 2007; Sergio *et al.*, 2007; Carnaval and Moritz, 2008; Ortega-Huerta and Peterson, 2008; Benito *et al.*, 2009; Roura-Pascual *et al.*, 2009) and may compete with or even outcompete presence/absence modelling techniques such as boosted regression trees (BRT), generalised additive models (GAM), generalised linear models (GLM) and multivariate additive regression splines (MARS) (Elith *et al.*, 2006; Wisz *et al.*, 2008). These good predictive capacities have been attributed to the ℓ_1-regularisation which prevents the algorithm from overfitting. Other models often do not apply any form of regularisation, and this can cause the observed difference in predictive performance (Gastón and García-Viñas, 2011). Moreover, Maxent is a generative approach, rather than discriminative. This can be an inherent advantage when the amount of training data is limited (Phillips *et al.*, 2006). Research pointed out that the technique can be applied with as little as 5 sampling points (Pearson *et al.*, 2007). For the nematode species in the database occurrence data is often scarce. A software related advantage is that it allows computerising the calculation of thousands of HSMs by running batch-files. Therefore, Maxent was applied in this thesis to create habitat suitability models. In spite of these promising features, Maxent models seem to have two major drawbacks: the models may fail to make general predictions (Peterson *et al.*, 2007) and the models may be inaccurate in the presence of biased data (preferential sampling) (Phillips *et al.*, 2009). In this research, these drawbacks are tackled in various ways.

Mapping a common species

Species with specific habitat requirements and a small spatial range are generally easier to model than common species (Segurado and Araújo, 2004; Evangelista *et al.*, 2008) because widely distributed species are not restricted to specific habitats. However, knowing where to find high densities of a common species may be of interest to fisheries (e.g. *Ensis directus*) or for conservation purposes (e.g. *Lanice conchilega* reefs). Application of regression kriging techniques could be an option, but at least 100, and preferably 144 observations are required for reliable kriging (Webster and Oliver, 2007). Clearly, this represents some limitation to the applicability of regression kriging. Therefore, an alternative was investigated: instead of using presence-only or presence/absence data, a threshold was applied to the density or relative density of the species. The applicability of this methodology was tested for 6 common nematode species and applied to *Lanice conchilega*.

MODEL OPTIMISATION

Creating a model is generally speaking not difficult. There are ample examples of software modules which create a model within seconds if data is supplied in the correct format. However, creating a significant model which provides correct insight in the ecology of species and communities needs careful consideration. There are many pitfalls when it comes to creating a model (Ülgen *et al.*, 1996; Rosemann, 2006a; Rosemann, 2006b): pitfalls concerning the data, the modelling technique and model interpretation. Counteracting these pitfalls will result in more accurate and reliable predictive models.

Pitfalls concerning the data

Ecological and environmental data are spatial data, and spatial data may be subject to two potential problems: spatial autocorrelation and preferential sampling.

A characteristic of most spatial data is that it shows spatial autocorrelation (SA). In other words, if two samples are taken closely together they are more likely to resemble each other than if the samples were taken at a larger distance from each other (positive SA). Consequently, both samples are not independent and the statistical assumption of independence is violated. SA may amplify or blur true ecological relations and incorporating spatial autocorrelation may even invert observed patterns (Kühn, 2007). Moreover, SA may inflate validation statistics since the test localities are not spatially independent from the training data (Hampe, 2004) (see Paragraph below for more information on training and test set). The presence of spatial autocorrelation is established by calculating Moran's I (Moran, 1950). Accounting for SA is often done by spatially separating training and test set (Pearson *et al.*, 2007; Murray-Smith *et al.*, 2009).

When spatial data is collected, preferential sampling may occur, i.e. some areas are more frequently visited than others. A model is more likely to deviate from a random model when such collection bias occurs (Raes and ter Steege, 2007). If the sampling points correlate with

some environmental characteristics, these environmental characteristics may erroneously 'explain' the patterns found in the data, while their contribution to the observed pattern may be negligible. Ignoring preferential sampling can thus negatively affect model parameter estimation and prediction. Here, we analyzed the presence of preferential sampling by either plotting the declustered mean as a function of cell size (Van Meirvenne, 2007) or by the use of null models (Raes and ter Steege, 2007).

Pitfalls concerning modelling

Overfitting occurs when the overly complex model describes the random error or noise instead of the underlying relationship. In that case, the model will fit very well the training data but it will perform poorly when applied to new, unseen data. The model is therefore not able to generalise (Izenman, 2008). Reducing model complexity will often result in better generalisation capability. However, if the model is overly simple it will not be able to explain the variation in the data. Consequently model complexity often has an optimum. This optimum can be found by applying cross-validation or by using an independent validation set which is solely used at the completion of the modelling exercise. In case of ANNs, three sets are created: a training set, a validation set and a test set. During the refining of the model weights, the error on the validation set is monitored. Once the error of the validation set increases, the iteration process is stopped. This is an internal validation, thus the validation set is not an independent dataset. Therefore, a third set, the test set, is created. k-fold cross-validation is a technique to assess if a model is capable to generalise to unseen data. The data is split in k partitions, each partition is once used as a validation set, while the other (k-1) partitions are used to train the model (training data) (Olson and Delen, 2008). Choosing the number of folds is always a trade-off between more samples for testing or more samples for training the model. 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, in case of data and time limitations a three-fold or a five-fold cross-validation may be more appropriate (Goethals, 2005). 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, both due to data and time limitations.

During the modelling, it is tempting to optimise and keep on optimising modelling parameters or look for further refinements. Increasing the complexity of the model, leads easily to being 'lost in best practice' (Ülgen *et al.*, 1996; Rosemann, 2006b). Moreover, if an immense amount of data is available, there is an almost unlimited choice in modelling possibilities. Setting clear targets and limitations should also be part of the modelling process (Ülgen *et al.*, 1996).

Pitfalls concerning the model output and interpretation

Statistical modelling involves finding a relation between the independent variables and a dependent variable. However, it is important to realize that correlation does not necessarily imply causality. In other words, if a mathematical relation is found between the dependent and the independent variable, this does not necessarily mean that the independent variable causes the changes in the dependent variable.

A model which can explain the variation in the data is only meaningful if the contribution of the environmental variables to the model output is known. More specifically, ANNs are generally known as 'black boxes'. Several techniques have been developed to reveal the contribution of each variable (Gevrey *et al.*, 2003). In this way the model can contribute to the ecological knowledge of the species. Having too much faith in the model output without knowing the limitations of the model is hazardous and may lead to false conclusions (Ülgen *et al.*, 1996). Coupling back to the ecological relevancy of the conclusions is a first step in counteracting this pitfall.

Explaining these limitations to potential end users should be part of the modelling process (Ülgen *et al.*, 1996).

DESCRIPTION OF THE DATA

Species data

The nematode data were retrieved from two databases: a database compiled within the UGent and the MANUELA database, which partly comprises the UGent database.

The UGent database was compiled within the framework of this thesis. In order to build a consistent database different quality controls were performed (Addendum 2). The database consists of historical data collected at Ghent University. It consists of 22 subsets with data collected in the framework of PhD research, BSc dissertations and funded research projects (Addendum 2). The complete database contains information on locations from all over the world, including the poles and tropical regions. However, the bulk of the data is collected at the SBNS. The database contains more than 206 000 species identifications and assembles data within the time frame 1971-2004. This database has the advantage that sampling and identification techniques are more similar for the data used in the analyses (see Addendum 2). This database is used in Chapters 3 to 6.

The MANUELA database was compiled within the EU Network of Excellence MarBEF. MANUELA was a Research Project focusing on the meiobenthic assemblages. The MANUELA database captures data on meiobenthos on a broad European scale (Vandepitte *et al.*, 2009). Table 1.4 gives an overview of the main characteristics of the UGent and the MANUELA database. The UGent database is included in the MANUELA database.

The macrobenthos data was retrieved from the MacroDat database and completed with new data in the case of *Ensis directus*. The MacroDat database is compiled at Ghent

University, with data of over 1275 stations and more than 640 000 species identifications collected between 1971 and 2008 (Degraer *et al.*, 2003a).

	# of subsets	# of stations	# of nematodes identified	# of nematode species	Time frame
UGent database	22	279	206 000	1975 (taxon level) 588 (species level)	1971-2004
MANUELA-database	82	1213	> 1 500 000*	1484 (taxon level) 951 (species level)	1966-2006

*Approximated value since counts were not always available

Table 1.4: Main characteristics of the UGent and the MANUELA database.

Environmental data

The environmental data was, if possible, retrieved from the databases. This is in fact the most accurate data, since it links the biotic data directly with the environmental data. However, in many cases this data is unavailable or incomplete and should be replaced by data from environmental maps. The advantage of maps is that a value can be drawn for each pixel within a given area that is mapped. A drawback is that these data are less accurate compared to data collected in the field. The maps have a resolution of 199 m x 199 m for the BPNS and 249 m x 249 m for the SBNS. Table 1.5 gives an overview of the available maps for both areas.

The available maps were acquired in two ways: by remote sensing or by interpolating field data and modelling. The data acquired by remote sensing covers data on total suspended matter and chlorophyll *a* in the water column (Park *et al.*, 2006). The data was collected by the MERIS spectrometer on board of the Envisat satellite within the Belcolour project. Eighty chlorophyll *a* maps and 90 total suspended matter maps were gathered in the time frame 2003-2005. These maps were reduced to three biologically relevant maps: the minimum, maximum and average values.

The second group of maps was derived from point sampling at sea or from modelling. The water current properties were modelled by the Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM). The sediment characteristics, such as median grain size and the silt-clay fraction were supplied by the Renard Centre of Marine Geology (RCMG) at Ghent University (Verfaillie *et al.*, 2006) and TNO Built Environment and Geosciences-Geological Survey of the Netherlands (TNO). The bathymetrical data were provided by the Ministry of the Flemish Community Department of Environment and Infrastructure, Waterways and Marine Affairs Administration and completed with data from the Hydrographic Service of the Royal Netherlands Navy and by the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management (RNLN). The only map representing anthropogenic effects on the sea

floor reflects information on the intensity of sand extraction. This data was supplied by the Federale Overheidsdienst Economie (FOE).

Variable type	variable	BPNS	SBNS	Source
Anthropogenic	Intensity of sand extraction	x		FOE
Biochemical	Average total suspended matter	x	x	Belcolour
	Maximum total suspended matter	x	x	Belcolour
	Minimum total suspended matter	x	x	Belcolour
	Average chlorophyll content	x	x	Belcolour
	Maximum chlorophyll content	x	x	Belcolour
	Minimum chlorophyll content	x	x	Belcolour
	Average salinity	x		Belcolour
	Maximum salinity	x		Belcolour
	Minimum salinity	x		Belcolour
Current prop.	Minimum bottom shear stress	x		MUMM
	Mean bottom shear stress	x		MUMM
	Maximum bottom shear stress	x		MUMM
	Size of the residual currents	x		MUMM
	Maximum depth-averaged current velocity	x		MUMM
	Magnitude of the residual transports	x		MUMM
	Residual currents	x		MUMM
	Residual transports	x		MUMM
	Tidal amplitude	x		MUMM
	Maximum current velocity at the bottom layer	x		MUMM
	Average current velocity at the bottom layer	x		MUMM
Oceanographic	Water depth	x	x	RCMG & RNLN
	Slope of the sea bottom	x		RCMG
	Bathymetric Position Index (1600 m range)	x		RCMG
	Bathymetric Position Index (240 m range)	x		RCMG
	Rugosity of the bottom	x		RCMG
	Orientation of the slope of the bottom	x		RCMG
Sediment	Median grain size	x	x	RCMG & TNO
	Gravel content	x		RCMG
	Sand content (63 µm - 2 mm)	x		RCMG
	Silt-clay content (0-63 µm)	x	x	RCMG & TNO

Table 1.5. Environmental variables available for the BPNS and the SBNS.

AIM AND OUTLINE OF THE THESIS

The overall aim of this research is to develop powerful statistical models which cope with modelling pitfalls which are typical for spatial data and data assembled from various sources, but are often ignored (Dormann, 2007). In this respect issues such as spatial autocorrelation, sampling bias (preferential sampling) and sampling effort are addressed. The aim of this thesis is two-fold: 1) adapting the currently used modelling process in such a way that potential pitfalls for a given dataset are revealed and circumvented; 2) getting insight in the ecology and biodiversity of the taxa under study on a small and large spatial scale. More specifically the diversity of nematode communities and the factors contributing to this diversity are modelled. Besides, diversity and community modelling this research focuses on habitat suitability modelling of nematode species and two macrobenthic species.

In this thesis, modelling focuses on statistical modelling. Therefore, different statistical modelling techniques such as artificial neural networks, geostatistics and maximum entropy modelling were used for the reasons mentioned before. These modelling techniques were combined with techniques such as cross-validation and null models to improve model quality and to address the previously mentioned pitfalls in order to create significant models which are able to generalise to unseen data (Segurado *et al.*, 2006; Dormann, 2007; Raes and ter Steege, 2007).

In Chapter 2 the potential role of species assembly rules on a small scale are investigated by applying null models to the original dataset. More specifically, we addressed the question whether the species composition in replicate samples is different from a random species distribution based on the local species pool and what might cause any possible non-random patterns. A routine was developed in Matlab to evaluate if the species composition between replicate samples is significantly different from a random distribution from the local species pool.

Chapter 3 explores the relation between biodiversity indices and environmental variables. Artificial neural networks are often considered to be 'black boxes', not revealing insight in the relation between dependent (i.e. biodiversity indices) and independent (i.e. environmental variables) variables. Therefore, two existing (Perturb and Profile) and one new technique (Modified Profile) were applied to reveal these relationships. As such, we were able to investigate the factors related to the observed diversity patterns.

In Chapter 4 different modelling and kriging techniques are compared to find the best map of nematode diversity for the SBNS. The use of replicate samples in geostatistical modelling is explored. These replicate samples may give a good insight in the local variation of the biodiversity caused by sampling errors and variability at small distances and this information may contribute to the quality of the final model.

Chapter 5 explores the influence of preferential sampling, spatial autocorrelation and overfitting on habitat suitability models. Null models were applied before to reveal the presence of preferential sampling (Raes and ter Steege, 2007), but they are equally useful in

revealing spatial autocorrelation and overfitting. As a result habitat suitability models for more than 100 nematode species were developed. These resulting models can be found in Addendum 3.

Well performing habitat suitability models are generally difficult to develop for common or generalist species with no specific habitat requirements. However, in some cases it is relevant to know where these species can be found in high densities. Therefore, we investigated the influence of the relative abundances on the outcome of habitat suitability models of six nematode species in Chapter 6. Do the high relative abundances relate to specific habitats?

Chapter 7 illustrates two applications of these modelling techniques to *Lanice conchilega* and *Ensis directus*.

The final chapter, Chapter 8, holds general conclusions and possible opportunities for future research. A theoretical introduction to artificial neural networks, geostatistics and maximum entropy modelling is supplied in Addendum 1. The codes developed in Matlab and R can be found in Addendum 5.

REMARK

Apart from the introduction, Chapter 7 and the general discussion, this thesis is a compilation of research articles which have been published or will be submitted to peer reviewed journals. This may result in overlapping information about the data and modelling techniques in the different chapters, but it also means that the chapters stand on their own and can be read individually.

CHAPTER 2

REVEALING SPECIES ASSEMBLY RULES IN NEMATODE COMMUNITIES

Bea Merckx, Maaike Steyaert, Tim Ferrero, Andrea McEvoy, Tom Gheskiere, Michaela Schratzberger, John Lambshead, Ann Vanreusel, Magda Vincx, Jan Vanaverbeke. Revealing species assembly rules in nematode communities.

CHAPTER 2

REVEALING SPECIES ASSEMBLY RULES IN NEMATODE COMMUNITIES

ABSTRACT

Species assemblages are not randomly assembled from a local species pool; they often show segregated or aggregated distribution patterns. These patterns may be attributed to both biotic and abiotic factors. On a large scale abiotic factors may be important, while on a smaller scale other factors such as species interactions may become essential. Here we will focus on small-scale patterns in nematode communities. Species patterns are generally revealed by null models based on presence/absence data. Since there is an increasing chance of falsely rejecting the null hypothesis of a random assembled community with increasing matrix size, we used an algorithm generating independent null matrices and applied a large number of swap attempts to build a null matrix. Moreover, we applied an additional test to reveal the susceptibility of the analyses of checker and the C-, T- and V-score to a Type I error for randomised data. To minimise the influence of the abiotic environment, we restricted the swapping algorithm of the null model to the replicate samples of one sampling event. Since stronger species interactions are expected for species of the same functional type, the nematode data was split according to the four feeding types defined by Wieser (1953). Our data indicate that species tend to aggregate and co-occur more often in some replicate samples than would be expected from a random species distribution of the local species pool. This is in accordance with the patchy distribution patterns known for nematode species. These aggregated patterns are also found for the different feeding types. The factors causing these aggregated patterns cannot be established since they are not included in the data, but the data do indicate that competitive exclusion is unlikely at the scale of a sample core.

Keywords

Null models, Nematoda, species assembly rules, aggregated pattern, patchiness

INTRODUCTION

Community ecology searches for repeated community patterns to investigate how communities are assembled from species pools (Wilson, 2001). The mechanisms behind these patterns can be mainly attributed to two factors: the abiotic environment and assembly rules. The latter focuses on community patterns due to interactions between species, such as competition, facilitation, mutualism and other biotic interactions (Wilson, 2001). Species assembly rules were first formulated by Diamond (1975). These rules imply that interspecific competition between species with similar niches results in non-random patterns of species distributions where certain species are competed out of the community. As a result, some species pairs may never be found together which leads to less species co-occurrences than expected by chance. These rules have been strongly disputed and there has been a proliferation of studies promoting, refuting and testing these ideas (Connor and Simberloff, 1979; Diamond and Gilpin, 1982; Weiher and Keddy, 1995; Bell, 2000; Hubbell, 2001; Weiher and Keddy, 2001; Bell *et al.*, 2006; Purves and Turnbull, 2010). In addition, it has been shown that neutral processes considering only birth, death, random dispersal and species richness may result in non-random patterns as well (Bell, 2000; Hubbell, 2001; Whitfield, 2002). However, it has been realised that neutral processes alone cannot explain the observed community patterns (Bell *et al.*, 2006; Purves and Turnbull, 2010). Moreover, the theory of Diamond (1975), stating that closely related species with similar attributes are more likely to outcompete each other, is contradicted by an alternative theory, which postulates that closely related species might have similar tolerances to environmental stressors, and would thus rather co-occur within the same communities (Webb, 2000). This theory has found support in marine communities, where species assemblages tend to be more closely related than would be expected by chance (Mouillot *et al.*, 2007; Somerfield *et al.*, 2009). This may indicate that environmental and evolutionary factors are determining marine community composition, rather than species interactions.

The observations of non-randomness in species assemblages have been tested extensively for terrestrial and freshwater ecosystems (Gotelli and McGabe, 2002). In marine habitats, null model analyses have been applied only recently (Mouillot *et al.*, 2007; Carranza *et al.*, 2010; Semmens *et al.*, 2010). Here we focus on the free-living marine nematodes. Many studies have investigated the biologic interactions between nematodes and other taxonomical groups (Schrijvers *et al.*, 1995; Debenham *et al.*, 2004; Kristensen, 2008; Braeckman *et al.*, 2011). However, here we will focus on interactions generating non-random distribution patterns of nematode species. Nematode communities are characterised by a high local diversity (Heip *et al.*, 1985) and within samples several species may belong to the same trophic group and may even be congeneric. Segregated patterns can be expected, since interspecific interactions between nematode species have been reported. In natural conditions, competition within nematode assemblages has been suggested (Heip *et al.*, 1985; Yodnarasri *et al.*, 2008) and significant interactions between species of the same

feeding type have been observed (Alongi and Tietjen, 1980; Joint *et al.*, 1982; Steyaert *et al.*, 2003). In experimental conditions, nematode species reveal complex interactions such as competition (Alongi and Tietjen, 1980), inhibition (De Mesel *et al.*, 2006a), facilitation (dos Santos *et al.*, 2009) and predation (Moens and Dos Santos, 2010). Since competitive interaction is mostly expected within groups of species feeding on the same food type, non-random patterns indicating segregation of species are often sought for within feeding types or guilds (Fox and Brown, 1993; Fox and Fox, 2000; Heino, 2009). Here, the data was subdivided into four feeding types (Wieser, 1953): selective deposit feeders (1A), non-selective deposit feeders (1B), epigrowth feeders (2A), and predators and omnivores (2B).

The aim of this study is to investigate species co-occurrence patterns within the nematode community, while (largely) excluding the effect of the environmental variables on the distribution patters. When these patterns are actually observed, they may then be attributed to species assembly rules (i.e. as a result of species interactions) rather than by the environmental conditions. To minimise the effect of the environment on the outcome of the analysis, the swapping algorithm was restricted to repeated samples taken at the same location at the same moment in time (further referred to as 'replicate samples'). Our null hypothesis states that for all the sampling stations, the species in the replicate samples are randomly assembled from the local species pool, with the local species pool being all the species found in the replicate samples from one sampling event. The presence of non-random distribution patterns are revealed by null models. In null model analysis, co-occurrence indices derived from the real species-samples matrix are compared with the indices derived from randomly assembled matrices. These random matrices can be assembled in many different ways (Gotelli, 2000). Here, we applied two null models for presence/absence data, the fixed-fixed and fixed-equiprobable model (Gotelli, 2000). Since, there is an increasing chance of falsely rejecting the null hypothesis of a random assembled community with increasing matrix size (Fayle and Manica, 2010), we 1) developed an algorithm generating independent null matrices in contrast to the generally used 'sequential' swap algorithm (Gotelli and Entsminger, 2003) and 2) applied a large number of swap attempts to build a null matrix. Moreover, we applied an additional test to reveal the susceptibility of the different analyses to a Type I error.

MATERIALS AND METHODS

Species data

To investigate assembly rules within nematode assemblages, data was drawn from the MANUELA database. 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).

The challenge in this study is to find non-random community patterns on a small spatial scale, but in such a way that these patterns are little or not influenced by environmental conditions. Replicate samples are obtained within a small spatial scale and at a certain moment in time. These replicate samples exhibit local variations in species composition, for similar environmental conditions. Thus, they can provide information on nematode assemblages on a limited time and spatial scale. Consequently, only those samples with at least two replicate samples are extracted from the MANUELA database. Time series were considered as different samples, since seasonal fluctuations may alter the species composition significantly (Vincx, 1989b; Vanaverbeke *et al.*, 2004a; Franco *et al.*, 2008).

In this way 911 replicate samples belonging to 338 sampling events (with each 2 to 4 replicate samples) were selected from the database (Fig. 2.1 and Fig. 2.4). Only those species found in more than one replicate sample are considered, resulting in a final dataset consisting of data on 450 different nematode species. The surface area of the samples varied between 3.8 cm² and 23.76 cm²: 44% of the samples had a surface area of 10 cm² and 24% had a surface area of 23.76 cm². A small proportion of the samples (about 4 %) had a surface area smaller than 6 cm² and for 28 % of the samples the surface area was not exactly known.

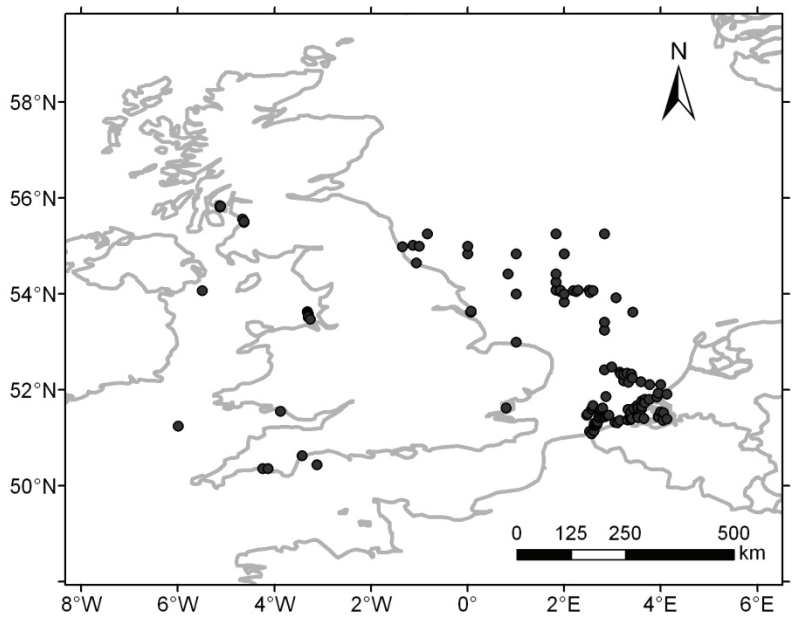


Fig. 2.1. Location of the sampling stations (•).

Tests for species assembly rules

There are two aspects in a null model test: the test statistic and the null model. The test statistic can be any parameter summarising a community aspect which might relate to species interactions, such as species aggregation or segregation. The main aspect in assembly-rule work is framing a valid null model (Wilson, 2001). The null model tests if the null hypothesis is valid. In this study, we test community patterns against a null hypothesis of random community assembly. Our null hypothesis states that for all the samples the species

in a core sample (i.e. a replicate sample) are randomly assembled from the local species pool with the local species pool being the species found in the replicate samples from one sampling event.

The null model tests whether the test statistic is significantly different from random, and should be chosen carefully to include every feature of the observed and thus realistic data, except the tested feature (Tokeshi, 1986). A systematic comparison of the dataset with different null models may help to reveal random or non-random distribution patterns in the species data. Therefore, two null models were developed using the Matlab software package.

Indices for revealing assembly rules

Several indices have been developed to summarise patterns in species distributions. Here, 4 indices based on presence/absence data are considered: Checker (Diamond, 1975; Gotelli, 2000), the C-score (Stone and Roberts, 1990), the T-score (Stone and Roberts, 1992) and the V-statistic (Pielou and Robson, 1972 *in* Schluter, 1984).

The C-score and Checker are commonly used indices (Gotelli and Rohde, 2002; Ulrich, 2004; Sanders *et al.*, 2007; Tomašových, 2008; Carranza *et al.*, 2010; Semmens *et al.*, 2010; Kamilar and Ledogar, 2011). Checker is the number of species pairs which never co-occur and form perfect checkerboard patterns (Diamond, 1975; Gotelli, 2000). The C-score indicates how species tend to avoid each other ('checkerboardness') (Stone and Roberts, 1992) and is a measure of species segregation (Stone and Roberts, 1990):

$$C - score = \frac{2\sum_{i=1}^{S}\sum_{j=i+1}^{S}(r_i - r_{ij})(r_j - r_{ij})}{S(S-1)} \qquad \text{(Eq. 2.1)}$$

where S is the number of species, r_i is the number of sites where species i occurs and r_{ij} is the number of sites where both species i and j occur. In a competitively structured community, the C-score should be significantly larger than expected by chance. The T-score on the other hand measures how species tend to aggregate ('togetherness') (Stone and Roberts, 1992):

$$T - score = \frac{2\sum_{i=1}^{S}\sum_{j=i+1}^{S} r_{ij}(N + r_{ij} - r_i - r_j)}{S(S-1)} \qquad \text{(Eq. 2.2)}$$

where N is the total number of samples.

The V-score is as an index for species association in samples. V is calculated as follows (Schluter, 1984):

$$V = \frac{\sum_{j=1}^{N}(T_j - \bar{T})^2}{N.\sum_{i=1}^{S}\left(1 - \frac{r_i}{N}\right).\frac{r_i}{N}} \qquad \text{(Eq. 2.3)}$$

with T_j the number of species in replicate sample j and $\bar{T}$ the observed mean number of species per replicate sample. A value larger than 1 indicates that the species co-vary positively, while if V is smaller than 1 the species co-vary negatively (Schluter, 1984).

Checker is most prone to measurements errors since a single occurrence can destroy a perfect checkerboard pair, while the C-score and the V-ratio are more robust and patterns can still be detected in noisy datasets (Gotelli, 2000).

The Null Models

Null models provide tools for testing non-standard hypotheses about patterns in ecological data (Gotelli, 2000). The general idea is that the original index of the real data matrix is compared with indices calculated from randomised data. There is always a trade-off between generalism and realism of the null model (Gotelli, 2000). A general null model without any constraints easily rejects the random null hypothesis. Thus, the null model fails to include obvious community features and reveals a non-random pattern. This is a Type I statistical error and should be avoided at all times. By introducing more of the original structure into the model, the model becomes ecologically more realistic. However, simulations will closely reflect the observed data and the null hypothesis will never be rejected if too much structure is incorporated in the model. In other words, the test is too conservative and produces a Type II error. Thus, the community is in fact different from random, but the null model is not capable of revealing this pattern (Wilson, 2001).

On a large scale, a pattern will always be discerned in the data due to environmental drivers (Chapter 1, Fig.1.3). Consequently, the influence of the environment should be excluded as much as possible when these patterns are investigated. This is achieved here by using replicate samples; replicate samples are obtained from the same station and are assumed to reflect similar environments. Thus, differences in species composition between replicate samples will be less influenced by the environmental conditions, but more likely by other factors.

Since more competitive interaction is expected within feeding groups, the original dataset was split according to the four feeding groups defined by Wieser (1953).

Due to data limitations the null model used in this study is based on the assumption that the volume of the individual replicate sample is small enough to allow species interactions, while the distance between the replicate samples is large enough to exclude species interactions and is still small enough to reduce the differences in environmental conditions. Because of the small size of the nematodes, this is an important consideration. Assumptions for interactions within a replicate sample cannot be deduced for this data since the nematode community is identified for the whole replicate sample and not for patches within the replicate sample. For some replicate samples data is available on the vertical species distribution at different depths in the sediment (slices). It is well known that nematode communities change with sediment depth: this may be attributed to both environmental conditions (Soetaert *et al.*, 1994; Steyaert *et al.*, 1999) and competitive or predatory interactions (Joint *et al.*, 1982; Steyaert *et al.*, 2003). Thus, these vertical environmental gradients will confound the distinction between the influence of abiotic and biotic factors on

the species distribution, while it is the purpose of this study to find patterns which are little or not influenced by environmental gradients.

The swapping algorithm

Null model analyses are based on binary presence/absence matrices where each row represents a species and each column a site, and each value indicates presence (1) or absence (0). Gotelli (2000) described nine different swapping algorithms, we applied two of them which are commonly used (Sanders *et al.*, 2007; Tomašových, 2008): the Fixed-Fixed (Swap1) and Fixed-Equiprobable (Swap2) approach.

For the first approach the matrix elements are reshuffled, but row and column totals of the original matrix are preserved (Connor and Simberloff, 1979; Gotelli and Entsminger, 2003; Kamilar and Ledogar, 2011) (Swap1). This has the following ecological background: setting a fixed row total will ensure that the number of occurrences of each species in the null communities is the same as in the original dataset (Gotelli, 2000). Thus, rare species will remain rare and common species remain common. Setting a fixed column total ensures that species poor sites, will remain species poor. Since the values of the number of species in a replicate sample (T_j) and the number of sites where species i occurs (r_i) remain the same with this algorithm, the V-score (Eq. 2.3) of the null models will have the same value as the V-score of the original data (Gotelli, 2000) and no pattern can be revealed. Therefore, the V-score was not considered for Swap1.

For the Fixed-Equiprobable approach, only the row totals are kept constant, and the replicate samples are considered to be equiprobable which can eliminate observed differences in species richness of replicate samples (Swap2). This null model approach is thus less conservative than the first one and it seems to produce good results for sample data collected in the field (compared to island data) (Gotelli, 2000).

For Swap1 the presence/absence data were transposed with the swap algorithm suggested by Gotelli (2000). In the data matrix submatrices of the form:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

are randomly chosen and zeros and ones are swapped. In this way row and column totals are kept constant, but the overall matrix changes. Here, the submatrices are chosen within the replicate samples of a sample. Thus, species can only shuffle between replicate samples and species from one sampling site cannot move to another sampling site. For each of the 338 samples (with each 2 to 4 replicate samples) 1000 swapping attempts were applied. Thus to create one null model 338 000 swapping attempts have been made. This is well above the commonly used value of 5000 (Gotelli and Entsminger, 2005) and the recommended value of 50 000 for large matrices (Fayle and Manica, 2010). Increasing the number of swaps decreases the Type I error (Fayle and Manica, 2010). Moreover, in our case every null model is developed independently, while the null models generated by the 'sequential' swap algorithm (Gotelli and Entsminger, 2003) are non-independent since each new null matrix is

generated based on the one before and differs from it by only 4 matrix elements (Gotelli and Ulrich, 2010). This implies that more null matrices are needed in case of a 'sequential' swap algorithm to obtain unbiased estimates of significance values (Fayle and Manica, 2011).

The above described indices were then calculated for each null matrix. This was repeated for 999 null matrices (Gotelli, 2000; Carranza *et al.*, 2010).

For Swap2, the swapping algorithm is somewhat easier. Only the row totals have to remain the same, thus a species can move from one replicate sample to another and the species richness of the replicate samples can change, but the number of observations for each species remains constant. One restriction is added to this swapping algorithm: if the swapping would result in replicate samples with no species, then the swap is not allowed, since degenerate matrices may increase the frequency with which the null hypothesis is rejected (Gotelli, 2000), and may thus enhance Type I error.

Gotelli (2000) showed that both swapping algorithms have good power for detecting non-random patterns in noisy datasets for Checker, the C-score and the V-ratio and have a low chance of falsely rejecting the null hypothesis. These swapping algorithms were applied to the complete dataset and to the datasets for the four feeding types.

An additional advantage of restricting the swapping algorithm to the replicate samples is that replicate samples were processed according to the same methodology. Thus, differences in sampling techniques between researchers and institutes will only have a minor influence on the final result.

Null Model Check

Recent research by Fayle and Manica (2010) showed that the probability of incorrectly detecting a signal in truly random data (Type I error) for the different indices increases with matrix size. They reviewed 47 publications, and it seems that our database is larger than any of the databases analyzed in these publications. Therefore, we developed an additional test, to check for Type I errors related to the structure of our database or related to the swapping algorithm. A reliable null model should only reveal significant differences caused by non-random distributions of species. Thus, if an artificial data matrix is supplied to the swapping algorithm, the null model should not reveal significant differences for the test statistics based on the randomised data matrix. Otherwise these significant differences should be attributed to other factors aside non-random structure in the species assemblage, such as errors inherent to the swapping algorithm or due to the specific structure of the data matrix. To keep the test within realistic constraints the artificial data matrix was composed by keeping the same number of species in a replicate sampling, but by randomly assigning species to a replicate sample. Thus, species aggregations or segregations should not be revealed for this matrix. For this artificial dataset, the test statistics are calculated and compared with the results of null models created with the swapping algorithm described above. To check the reproducibility of this test, it was repeated for three artificial data

matrices. Since the test is repeated three times and the computation of null models is time consuming, the number of null models was restricted to 500. No significant differences should be detected by this test.

RESULTS

Null model check

The null model check for the artificial datasets for the first swapping algorithm with presence/absence data (Swap1, Fixed-Fixed), reveals no significant difference for any of the test statistics (Table 2.1). The second swapping algorithm (Swap2, Fixed-Equiprobable) however, reveals significant differences for the C-, T- and V-score for the artificial data matrix (Table 2.1). The C-score is larger than expected for a random community, while the T- and V-score are smaller than expected by chance; indicating segregated species patterns. The Checker-index shows no unequivocal response. In one case it is significantly larger than expected by chance, thus indicating that the number of species pairs forming perfect checkerboard patterns is larger than expected for a random distribution, while in the two other cases it cannot be distinguished from random.

Swapping algorithm	Test statistic	Results for 3 artificial data matrices		
Fixed-Fixed	C-score	2260.0	2259.4	2259.1
(p/a)	T-score	2828.2	2827.6	2827.3
	Checker	2724.0	2730.0	2785.0
Fixed-Equiprobable	C-score	**2259.8** (>)	**2259.3** (>)	**2259.4** (>)
(p/a)	T-score	**2828.0** (<)	**2827.5** (<)	**2827.6** (<)
	V-score	**6.97** (<)	**6.97** (<)	**6.97** (<)
	Checker	**2803.0** (>)	2794.0	2800.0

Table 2.1. Values of the test statistic for the artificial data matrices; values in bold indicate that the test statistic of the artificial data matrix is significantly different from the null models ($p<0.05$ for a two-sided confidence interval), (>) and (<) mean that the test statistic of the artificial matrix is respectively significant larger or smaller than expected by chance.

Results for the real data

Presence/absence data

The two swapping algorithms applied for the presence/absence data are thoroughly investigated by Gotelli (2000). He found that these two swapping algorithms have the best properties concerning Type I and Type II errors. However, Fayle and Manica (2010) showed that the algorithm may be prone to Type I errors for large datasets. The most conservative

null models (Fixed-Fixed) reveal no significant difference for the C-score and the T-score (Fig. 2.2) indicating that no species segregation and aggregation is apparent in the data. On the other hand, the number of perfect checkerboard pairs is significantly higher compared to the

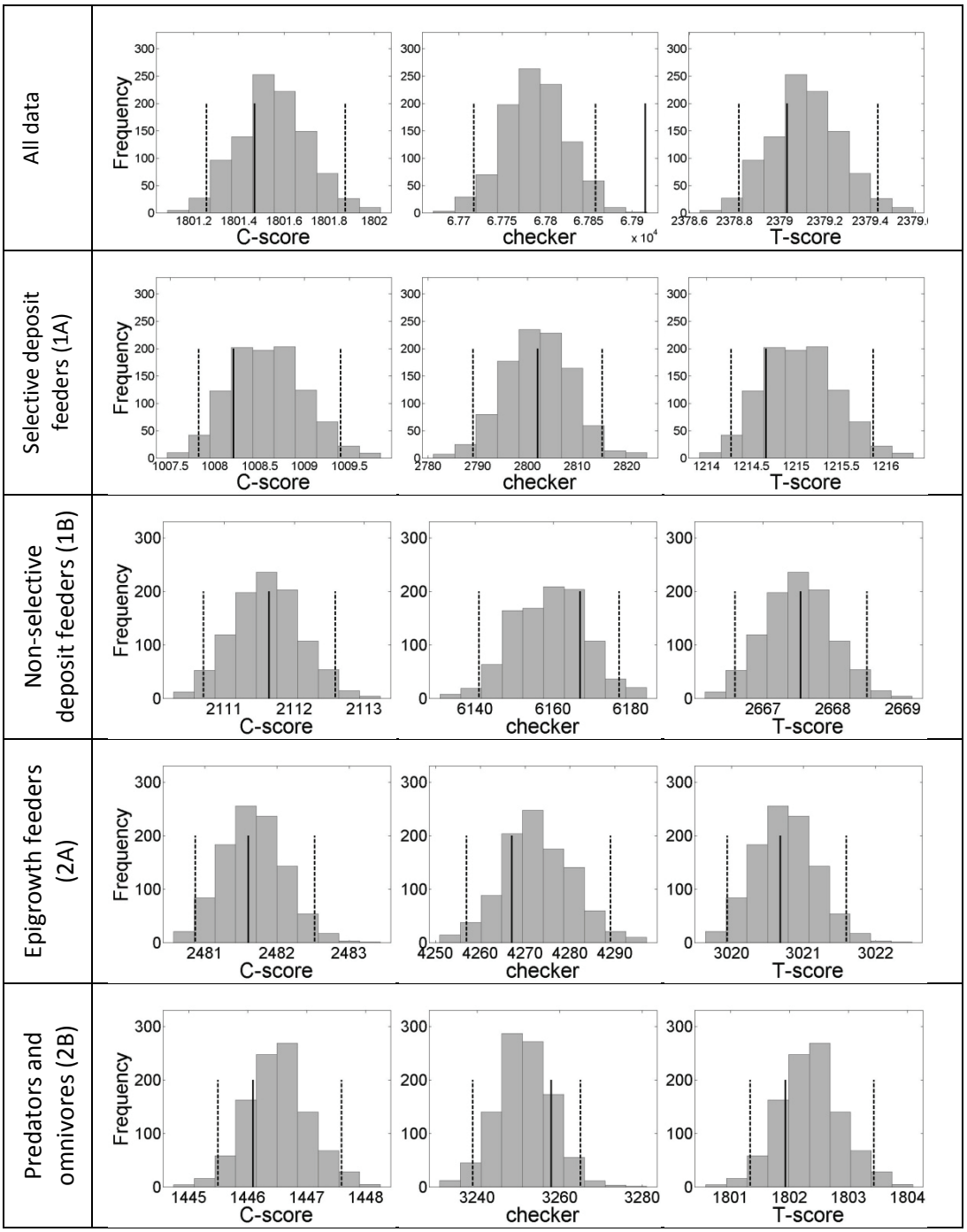


Fig. 2.2. Comparison of the two sided 95% confidence interval of the 999 null models (dotted lines) based on Swap1 (Fixed-Fixed) with the original value (full line) for the three community parameters (C-score, Checker and T-score) for all the data and the four feeding types.

null models. Thus, there are more checkerboard pairs in the real data matrix than would be expected by chance. Checker is a parameter which may be prone to Type II error (Gotelli, 2000), but it has good Type I characteristics and significant differences should be reliable (Gotelli, 2000; Carranza *et al.*, 2010).

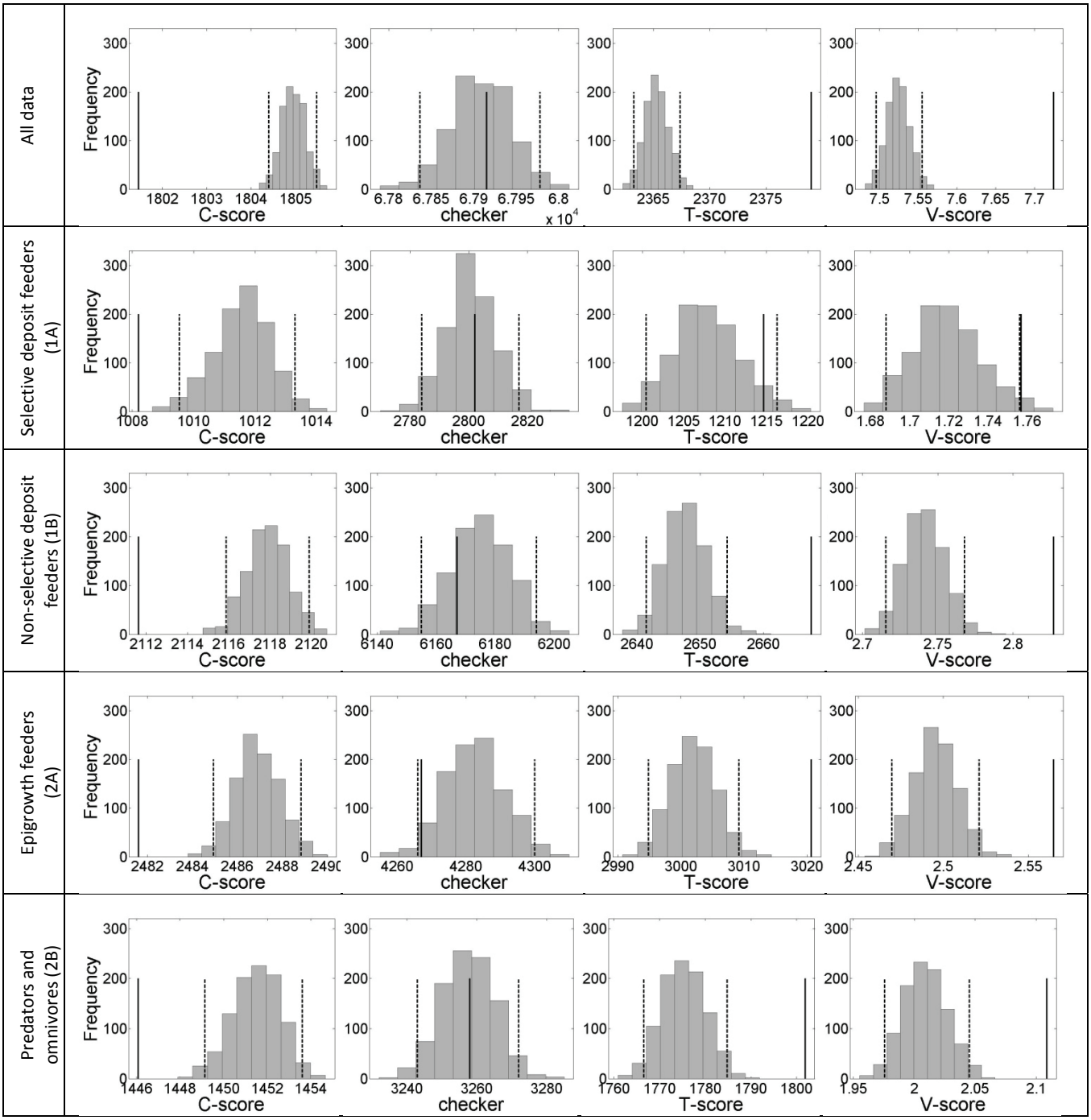


Fig. 2.3. Comparison of the two sided 95% confidence interval of the 999 null models (dotted lines) based on Swap2 (Fixed-Equiprobable) with the original value (full line) for the four community parameters (C-score, Checker, T-score and V-score) for all the data and the four feeding types.

The less conservative test (Fixed-Equiprobable) shows significant differences for the C-score, the T-score and the V-score, but not for Checker (Fig. 2.3): The C-score of the real data matrix is significantly smaller than the random values, while the T-score and V-score are higher than expected by chance. As demonstrated in the previous paragraph, this swapping algorithm results in Type I error for the C-, T- and V-score, thus caution is necessary when interpreting these results. However, for the real data matrix the opposite pattern is found compared to the artificial data matrices. Only the T-score for the 1A feeding type reveals no significant pattern.

The C-score and Checker indices quantify co-occurrence and can produce the same results (Gotelli, 2000), which is clearly not the case here. The C-score with Swap2 indicates that the species tend to aggregate more than expected by chance. In contrast, the Checker index with Swap1 indicates that there are more checkerboard pairs than would be expected by chance. However, Stone and Roberts (1992) found that a high 'checkerboardness' may be the result of aggregated species. To resolve this apparent contradiction they developed the T-score which in our case confirms the presence of aggregated communities. The overall V-score for the entire area is larger than one indicating that species co-vary positively.

DISCUSSION

Algorithm

The Fixed-Fixed algorithm for presence/absence behaves well for the artificial datasets: the indices calculated for the artificial data matrix are not significantly different from the indices calculated for the null models. However, the Fixed-Equiprobable algorithm for presence/absence data reveals that species tend to co-occur less than expected by chance, while such patterns are not supposed to be present in the artificial dataset. The presence of this Type I error may be due to the large size of our datasets (Fayle and Manica, 2010). The large amount of data triggers thus some unexpected problems. Fayle and Manica (2010) attributed these problems to 1) the 'sequential' swapping algorithm (resulting in non-independent null matrices) generally used in null model analysis and 2) the use of too few swappings to construct one null matrix. In our research we did not apply the 'sequential' swap: each null matrix was built independently from the previous null matrix and the number of swaps to construct one null model was increased from 5000 to 338 000 swapping attempts. This resulted in a time-consuming null model analysis, which we expected to be less prone to a Type I error for the different indices. Nevertheless, a significantly smaller T-score and a significantly larger C- and V-score are found for the randomised data. Remarkably, the opposite pattern, a higher co-occurrence than expected by chance, was found for the real data matrix. Thus, notwithstanding the bias of the swapping algorithm and the matrix structure towards segregated communities, the real data overrules this bias and indicates that species tend to co-occur in some replicate samples, forming aggregated patterns (Fig. 2.4).

This pattern is not found for the first swapping algorithm: when the total number of species in the replicate sample is kept constant, no co-occurrence patterns are revealed. This can be related to the fact that the algorithm is too conservative to reveal any non-random distribution patterns. However, it is possible that the co-occurrence patterns revealed by Swap2 are caused by the differences in species richness between the replicate samples and less by the presence of specific species pairs.

The V-score of the presence/absence data for the complete dataset is larger than 1 indicating that the species co-vary positively, which is not surprising since the species form distinct communities over the studied area, which can be ascribed to environmental gradients (Vincx *et al.*, 1990; Vanreusel, 1990; Vanaverbeke *et al.*, 2011). This is also reflected in the null models, where V is larger than one as well. It is evident that the null models also have V-scores larger than one, since swapping is only allowed within replicate samples. Thus, species will not appear in regions where they are not observed.

The analyses based on data concerning the feeding types generally display the same patterns as for the entire dataset. Hence, there is no evidence that species within feeding types interact stronger with each other. For the Fixed-Fixed algorithm for presence/absence data a significant difference for Checker was found for the overall data but not for the four feeding types apart. This may indicate that the checkerboard pairs are formed by species belonging to different feeding types. Yodnarasri *et al.* (2008) observed competitive interactions between epigrowth feeders (2A) and non-selective deposit feeders (1B). However, checking for patterns within the combined group 1B and 2A with the Fixed-Fixed algorithm resulted in a random pattern for all the test statistics (results not shown).

The tendency of species to aggregate does not necessarily imply that specific species pairs co-occur more often than expected by chance. Appointing individual species pairs which often co-occur could be an interesting starting point to set up future experiments. However, this is a statistical challenge (Sfenthourakis *et al.*, 2006) because even a small number of species results in a high number of species pairs. Many pairs will be significantly different from random just by chance at the 5% or 1% error threshold (Ulrich *et al.*, 2009). In our case, the entire dataset contains data on 450 species, which is an unusual high number, resulting in 101 025 unique species pairs. Thus, appointing non-random species pairs is statistically very precarious.

Sample size and patchiness

The effect of the sample size on the result of the null model is minimised by restricting the swapping algorithm to the replicate samples of one sampling event, which all have the same size. The most probable effect of the different sample sizes on the outcome of the analysis is an increase of a Type II error: although a non-random distribution pattern is present, it cannot be derived from the data. For instance, in case species tend to form aggregated patterns (Fig. 2.4), this pattern might be obscured by large samples because species might

be sampled at the edges of the large replicate sample. The different sample sizes might thus blur the distribution patterns.

It is a well-known phenomenon that nematode communities tend to form patchy distributions (Li *et al.*, 1997; Somerfield *et al.*, 2007; Gingold *et al.*, 2010a). On a small scale the horizontal distribution of nematodes shows a strong patchiness; within a range of a few centimetres nematode densities can drop with a factor 3 (Arlt, 1973). Horizontal patch size of meiofauna can vary between 0.3 to 700 cm² (Heip and Engels, 1977; Findlay, 1981; Blanchard, 1990). The distribution of most nematode species show strong aggregations (Blanchard, 1990) and species may show repeating patterns in densities of 8, 10 or 12 cm depending on the species (Blome *et al.*, 1999). The meiofauna sampling cores in our study have mostly a diameter of 3.6 cm (44 %) or 5.5 cm (24 %) and it is thus evident that the cores may sample at the middle of a dense nematode community or between these communities (Fig. 2.4). The swapping algorithm is based on presence/absence data, and patchiness is often associated with higher nematode densities. To validate the hypothesis that patchiness and thus higher densities are linked with higher species richness an additional test was done: for each sample (with more than 2 replicate samples) the Spearman rank correlation coefficient between the total density of the nematode community and the species richness in the replicate samples is calculated. This could only be done for the samples where the total density of the replicate sample is known (216 samples): 50% of the 216 samples have a Spearman rank correlation coefficient larger than 0.5 and only 12% have a Spearman rank smaller than -0.5. For small samples, it is easy to produce a strong correlation by chance and caution should be paid when interpreting these results, but it is clear that there is a strong tendency to find more species in replicate samples with higher densities (as represented in Fig. 2.4).

Ecology

The previous analyses indicate that the communities in the replicate samples are not randomly structured: species tend to aggregate in some replicate samples within a station and not in others. However, the mechanisms explaining the non-random pattern are difficult to assess. Even if the null hypothesis is rejected, it is impossible to leap to the conclusion that species interactions have led to these patterns (Simberloff and Connor, 1981). The actual mechanism behind the non-random patterns should be revealed by experiments (Gotelli, 2001). However, if competition and facilitation are at work, these mechanisms are expected to leave different signatures in the pattern of species co-occurrence.

Competition may result in a given species pair co-occurring less often than expected by chance, whereas facilitation may result in a given species pair co-occurring more often than expected by chance. Previous research of assemblages in marine environments show that any pattern can be found: random patterns for gastropods (Carranza *et al.*, 2010), strongly aggregated patterns for reef fish assemblages (Semmens *et al.*, 2010) or strongly segregated patterns for brachiopods (Tomašových, 2008).

On a large scale of meters to kilometres environmental gradients structure nematode communities (Soetaert *et al.*, 1994; Li *et al.*, 1997). When reducing the scale of observation, other factors may become more important, such as species interactions and patch dynamics (Levin *et al.*, 2001). According to our results nematode species tend to aggregate in some replicate samples more than would be expected by chance: this may be attributed to both a

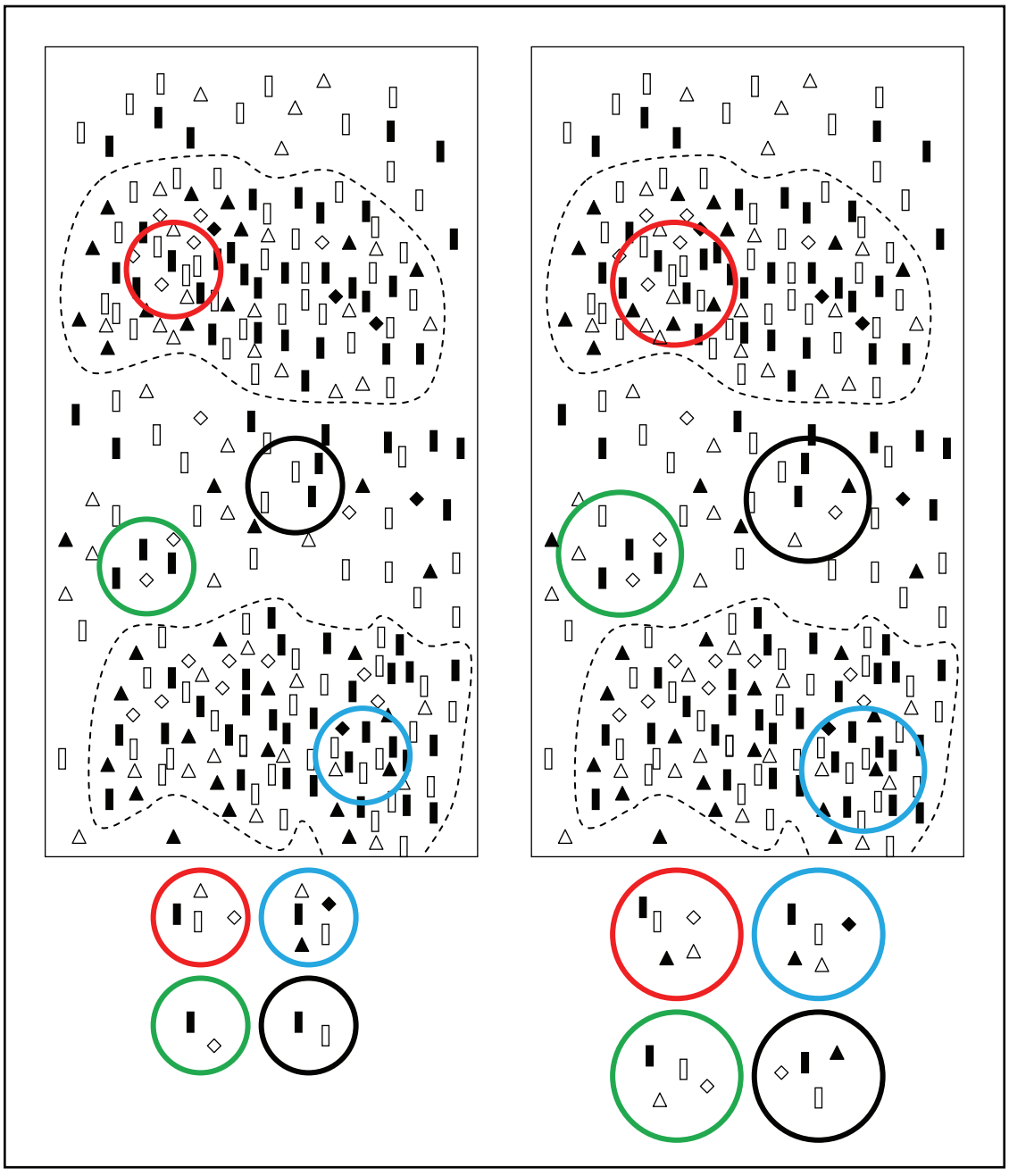


Fig. 2.4. Schematic representation of a sampling event where 4 replicate cores are taken with a small sampling core and a larger sampling core in the same community. Patches with higher densities of the nematode species are delineated by a dashed line. The species found in the replicate samples are represented in the circles at the bottom. The sampling design at the left is more likely to reveal the aggregated pattern of the species (blue and red core), while the sampling design at the right samples better the total species richness.

similar response to the environment (Sanders *et al.*, 2007) or to species interactions. Environmental differences between replicate samples may lead to different communities in these samples. The replicate samples are obtained at a certain moment in time at a certain sampling station. Therefore, the environmental circumstances in the replicate samples should be comparable. However, information about the actual distances and physical differences between the replicate samples is unavailable and it is possible that our original assumption of similarity of abiotic factors between replicate samples is idle.

Nematode communities often show a patchy distribution, a pattern which is confirmed here. Many factors may contribute to the origin of a patchy distribution: microtopography (Hogue and Miller, 1981; Sun *et al.*, 1993; Blome *et al.*, 1999), the presence of biogenic structures and macrofauna (Reise, 1981; Reidenauer, 1989; Braeckman *et al.*, 2011) or patches in food sources (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). Our study confirms the presence of small scale aggregations, with some replicate samples holding more species than others. However, the mechanisms behind this non-randomness cannot be unravelled with these null models.

The same aggregation patterns were found for the different feeding types. The current classification of nematodes by feeding groups is rather coarse and species belonging to the same feeding type may express different adaptations in the buccal cavity (Wieser, 1953; Deutsch, 1978). There has been some debate on this subdivision and more refined subdivisions have been suggested (Moens and Vincx, 1997; Moens *et al.*, 2004) but these differentiations are currently unknown for most nematode species. Thus, this refined subdivision could not be applied to our data. The aggregated patterns of the species belonging to the same feeding type may also be explained by the theory of Webb (2000) which postulates that closely related species are more likely to co-occur due to a similar response to the surrounding environment. This theory is supported by observations of co-existing closely related meiofaunal species (Heip *et al.*, 1985; De Mesel *et al.*, 2006b), and has been attributed at the time to the presence of microhabitats. Somerfield *et al.* (2009) suggested that in an open dynamic system such as the marine environment competition is most probably only operating on short time scales and small spatial scales. Indeed, species segregations have been found on a small vertical spatial scale between sediment slices of 1 mm or 5 mm (Joint *et al.*, 1982; Steyaert *et al.* 2003) where species interactions between two epigrowth feeders (Joint *et al.*, 1982), between predator and prey nematodes (Steyaert *et al.*, 2003) have been observed. But other factors such as food availability, oxygen distribution, physical disturbance and compaction of the sediment (Arlt, 1973; Joint *et al.*, 1982; Steyaert *et al.* 2003) may also contribute to these segregated patterns.

CONCLUSIONS

The results of our analysis are not unequivocal. Large databases may reveal non-random community patterns while they are not present (Fayle and Manica, 2010). This is also

supported by our analyses: randomizing the data revealed a Type I error for the different indices. When applying the swapping algorithms to large databases we therefore recommend an additional test which investigates the Type I error properties for the indices and the algorithms under study. It is clear that further research is needed to find the factors causing these errors and further adjustment of the algorithm is needed in such a way that co-occurrence patterns can be unequivocally revealed from large databases.

Nevertheless, our analyses indicate the presence of non-random community patterns at the level of replicate samples within a sampling station, suggesting locally aggregated nematode communities. Our analyses also indicate that patches with higher nematode densities generally have higher species richness. This is in accordance with previous research describing the patchy distribution of nematode assemblages which has been attributed to a variety of biotic and abiotic factors.

However, many questions remain unresolved: which factors contribute to this non-random distribution pattern? Is this pattern a general pattern for the entire area or do some regions or samples contribute strongly to the observed pattern?

Drawing conclusions regarding species interactions is impossible based on the algorithm; this is only achievable by carefully monitored experimental set-ups. However, our analyses do not suggest the presence of competitive interactions. Other factors may contribute to the observed aggregated pattern, such as the coarse subdivision of the feeding types, the large scale of the replicate samples compared to the interaction scale of nematodes, the unknown environmental differences between the replicate samples and the patchy distribution of the nematodes. Besides, in an open system such as the marine environment competitive interactions may only be present on a small temporal and spatial scale (Somerfield *et al.*, 2009) and due to environmental stressors closely related species may even co-occur more than expected by chance (Webb, 2000).

ACKNOWLEDGEMENTS

This research is funded by the Fund for Scientific Research (FWO) of the Flemish government (FWO07/ASP/174). The authors thank all the nematode data providers: Dr. Maaike Steyaert, Dr. Tim Ferrero, Andrea McEvoy, Dr. Tom Gheskiere, Dr. Michaela Schratzberger, Prof. Dr. John Lambshead, Prof. Dr. Ann Vanreusel, Prof. Dr. Magda Vincx and Dr. Jan Vanaverbeke. Special thanks to the Flanders Marine Institute (VLIZ, www.vliz.be) for help in building the biological database. This research was conducted within the MANUELA framework (www.marbef.org/projects/Manuela), which is a Responsive Mode Project undertaken as part of the MarBEF EU Network of Excellence 'Marine Biodiversity and Ecosystem Functioning', which is funded by the Sustainable Development, Global Change and Ecosystems Programme of the European Community's Sixth Framework Programme (Contract No. GOCE-CT-2003-505446). This research was also supported by the GENT-BOF Project 01GZ0705 Biodiversity and Biogeography of the Sea (BBSea).

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).

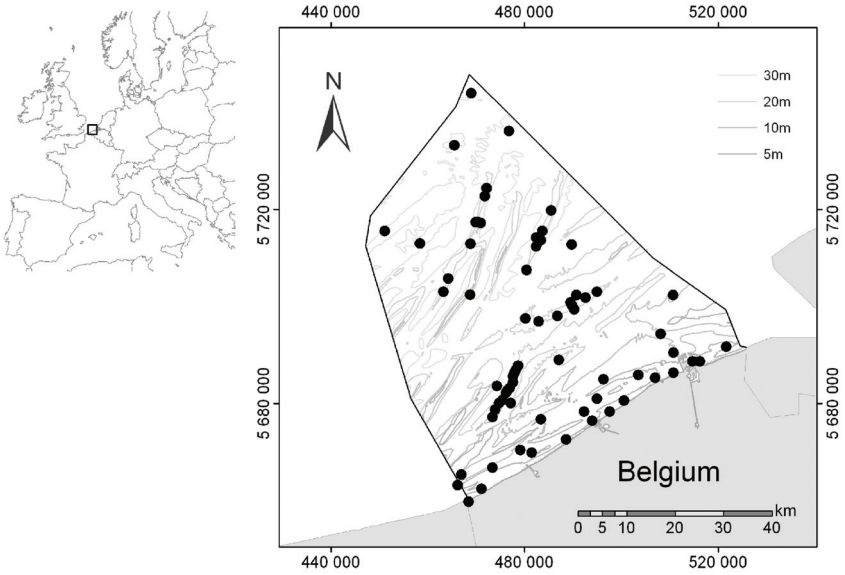


Fig. 3.1. Sampling stations on the Belgian Continental Shelf (•) (UTM31N-WGS84 coordinates).

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.

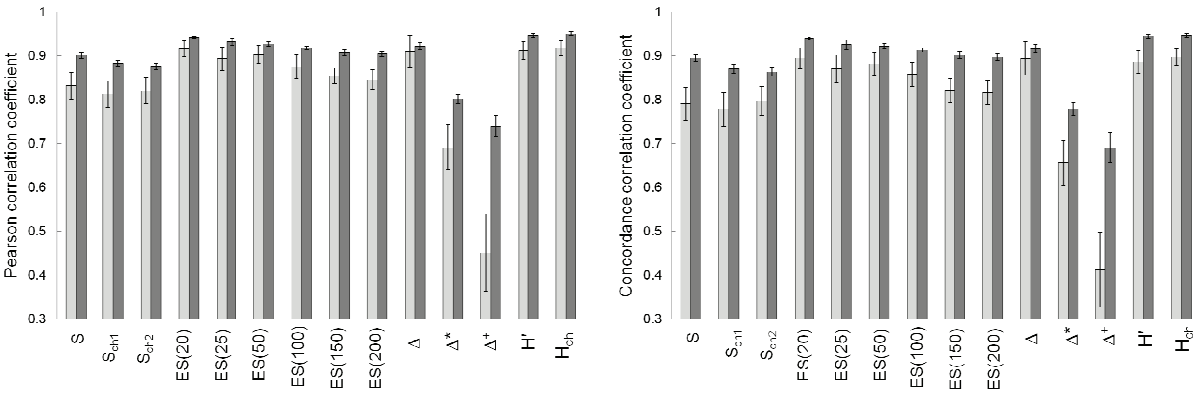


Fig. 3.2. Pearson product-moment correlation coefficient (left) and concordance correlation coefficient (CCC) (right) for observed and predicted values for the independent test set (light grey bars) and the learning set (dark grey bars) for the 10 selected models (±S.E.).

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.

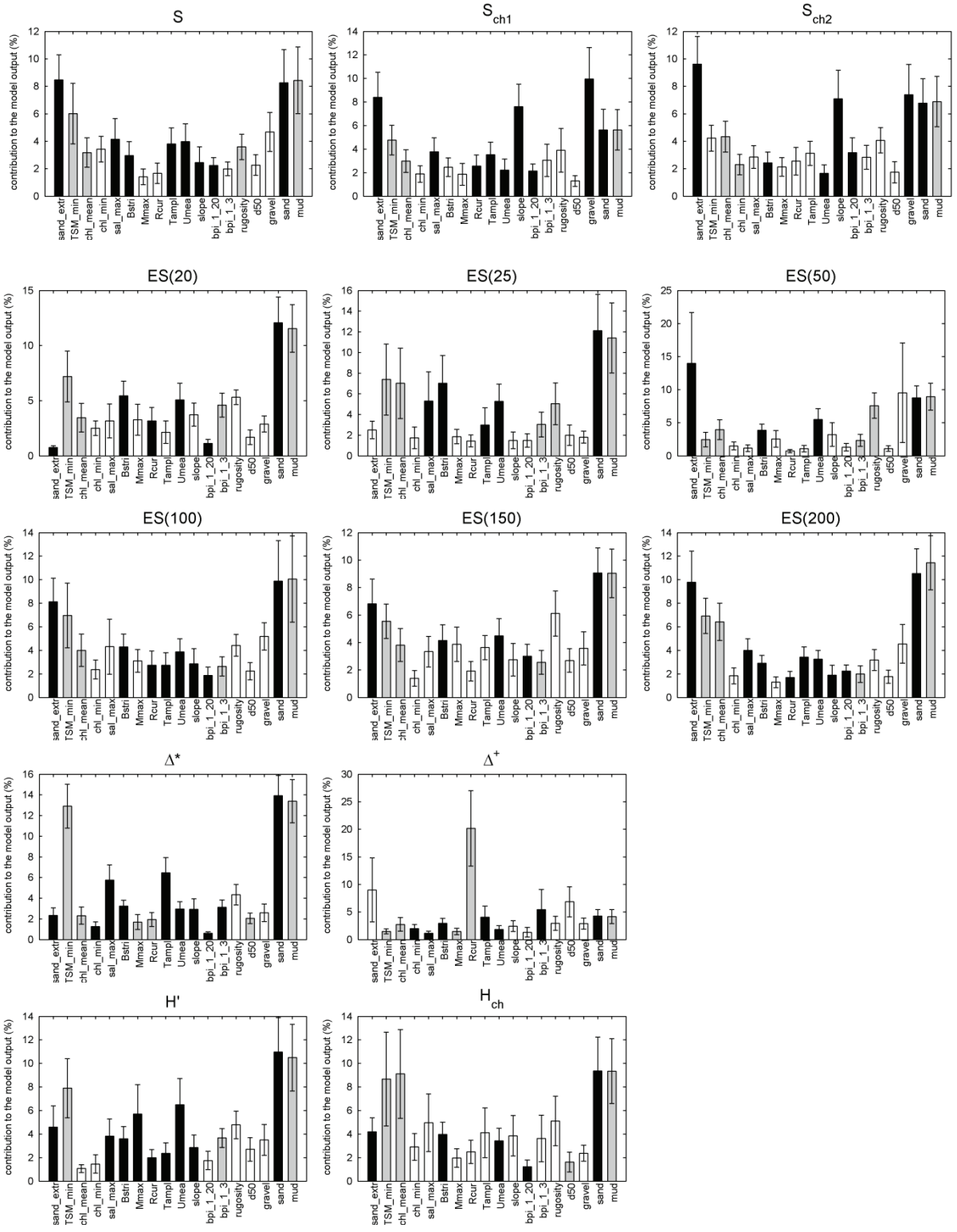


Fig. 3.3. Contribution of the environmental variables to the diversity indices according to the Perturb method averaged over the 10 selected models (±S.E.). The black bars represent those variables which are positively correlated with the diversity index for at least 75% of the models according to the 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.

CHAPTER 4

MAPPING NEMATODE DIVERSITY IN THE SOUTHERN BIGHT OF THE NORTH SEA

Adapted from: Merckx, B., Van Meirvenne, M., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2010. Mapping nematode diversity in the Southern Bight of the North Sea. Marine Ecology Progress Series 406, 135-145.

CHAPTER 4

MAPPING NEMATODE DIVERSITY IN THE SOUTHERN BIGHT OF THE NORTH SEA

ABSTRACT

In order to protect the biodiversity of the seas from, e.g. overexploitation, the spatial distribution of biodiversity and the mapping of biodiversity hotspots are of great importance. In the present paper we discuss different methods to develop full coverage biodiversity maps of free-living marine nematodes in the Southern Bight of the North Sea. A database with sampling data, gathered over 3 decades (1972 to 2004), combined with exhaustive environmental data, was employed to predict species richness and the expected number of species by different methods: ordinary kriging (OK) and regression kriging (RK) with ordinary least squares (OLS) and generalised least squares (GLS). The predictive value of these methods was evaluated by an independent validation set. Replicate samples were used to make an accurate estimation of the nugget variance, since replicates reveal local variability. Accordingly, it was feasible to find a spatial pattern in the residuals of the regression models. Our analysis pointed out that GLS improved the OK models substantially, while RK only slightly improved the GLS model. The diversity of marine nematodes is substantially influenced by the silt-clay fraction and the amount of total suspended matter, which is also reflected in the resulting map with a species-poor area near the coast line, especially near the south of the mouth of the Scheldt estuary. Off coast diversity and evenness are generally higher.

Key words

Nematoda, diversity, geostatistics, North Sea, mapping, generalised least squares

INTRODUCTION

The ocean floor is a heterogeneous environment with marine biota patchily distributed. Therefore, diversity and densities of marine communities tend to be higher in localised areas (Reese and Brodeur, 2006). Moreover, the marine seabed is increasingly disturbed and degraded by bottom trawling, sand extraction, dredging and dumping, which inevitably reduces biological diversity. Identification of diversity hotspots is a major concern in diversity conservation, and the identification of these marine, biological hotspots is a growing area of research (Malakoff, 2004; Reese and Brodeur, 2006). Biodiversity indices are often used to describe these areas of biological interest. However, studies investigating diversity in large areas are scarce, due to the high costs involved in such a labour-intensive process, and most studies result in point observations, while there is a growing need for full coverage maps.

Geostatistical interpolation techniques offer a powerful and cost-effective alternative; based on available point observations of communities and full coverage maps of relevant environmental data, full coverage maps of diversity can be constructed. Kriging has been developed for spatially structured mining data (Matheron, 1963) and is widely used in the terrestrial environment to create maps of chemical properties of soil and air (Hengl *et al.*, 2004; Hoek *et al.*, 2008; Van Meirvenne *et al.*, 2008) and more recently it has been applied to model faunal (Walker *et al.*, 2008) and floral (Hernández-Stefanoni and Dupuy, 2007) distributions. In the marine environment, it has been employed to map soil characteristics (Verfaillie *et al.*, 2006), distribution patterns of marine species (Mello and Rose, 2005; Rios-Lara *et al.*, 2007) and, to a lesser extent, diversity (Reese and Brodeur, 2006). In the present study, 2 conceptually different approaches were used: (1) interpolation relying only on point observations of the diversity index, known as ordinary kriging (OK) and (2) interpolation based on both point observations and full coverage environmental maps, known as regression kriging (RK).

In the present study, we focused on free-living marine benthic nematodes, a taxon within the meiofauna, comprising metazoans passing through a 1 mm mesh sieve but retained on a 38 µm mesh sieve. To our knowledge, these geostatistical techniques have not been applied before on Nematoda. This is surprising since these free-living roundworms represent the highest metazoan diversity in many benthic environments in terms of species numbers (Heip *et al.*, 1985): >50 species are commonly found in a single 10 cm^2 core. Owing to their interstitial lifestyle, properties of the sediment, such as grain size distribution, the silt-clay fraction and food availability, have a strong influence on the diversity and composition of nematode assemblages (Heip *et al.*, 1985; Vanreusel, 1990; Vincx, 1990; Steyaert *et al.*, 1999). Nematode communities seem to be resilient to disturbance, and their restoration occurs easily after temporal, low impacts (Kennedy and Jacoby, 1999; Schratzberger *et al.*, 2002), making them a perfect community to model based on long-term environmental and full coverage data. Previous research on the predictability of nematode diversity did indeed

yield accurate predictive models (Merckx *et al.*, 2009); yet, these were not area-covering models.

The research area is the Southern Bight of the North Sea. The seafloor is not at all homogeneous as it is characterised by sand dunes and a wide range of sediment types, varying from muddy to sandy environments (Lanckneus *et al.*, 2002). The coastal area is characterised by a high amount of total suspended matter, chlorophyll *a* (chl *a*) and silt-clay fraction, especially near the Belgian coast. The primary objective of the present study was to create accurate biodiversity maps of the nematode diversity of the Southern Bight of the North Sea.

MATERIALS AND METHODS

Study area

The research area, with a total surface of about 18 000 km^2, is situated in the Southern Bight of the North Sea, near the Belgian and the Dutch coastal area (latitude: 51°6'2'' to 52°59'19'' N; longitude: 2°14'39'' to 4°30'43'' E).

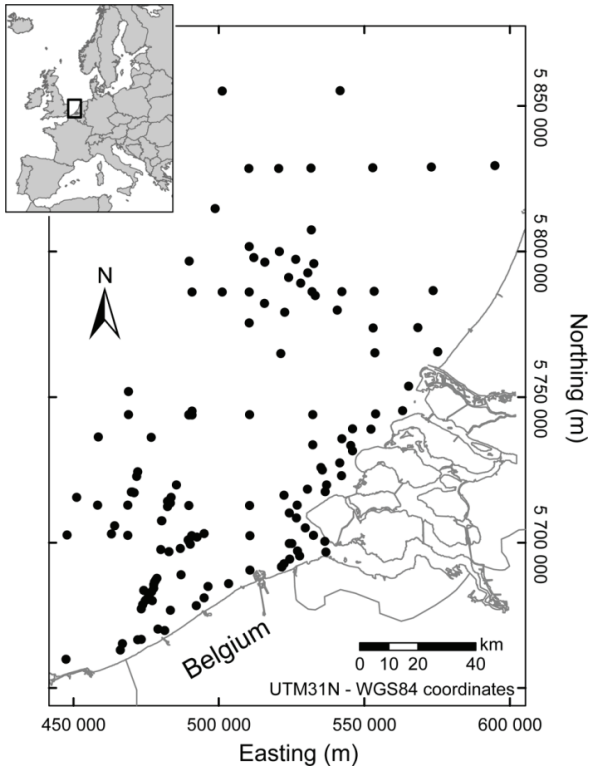


Fig. 4.1. Study area and location of the sampling stations (•).

Nematode data

The nematode data were retrieved from the MANUELA database. Within the EU Network of Excellence MarBEF, MANUELA is a Research Project focusing on meiobenthic assemblages. The MANUELA database was compiled capturing the available data on meiobenthos on a

broad European scale (Vandepitte *et al.*, 2009). For the present paper, the area of research was restricted to the Southern Bight of the North Sea, since, firstly, full coverage environmental maps were available for the entire region and, secondly, results were not biased by sampling strategy because all data were collected by a single institute, the Marine Biology Research Group of Ghent University. The resulting dataset consisted of 562 samples belonging to 153 different stations (Fig. 4.1). These data included information on 99 966 nematodes identified to species level and collected in the time frame from 1972 to 2004. Different sampling gears were used to collect these data: 49% of the samples were taken with a Reineck box corer, 31% with a Van Veen grab, 19% with a NIOZ box corer and 1% by divers. All subsamples were taken with Perspex cores, with a surface area of 10 cm^2.

Environmental data

The source of the environmental data can be divided into 2 groups: from maps acquired by remote sensing and from maps derived from data sampled in the field.

The first group of data was derived from remote sensing by the MERIS spectrometer on board the Envisat satellite of the ESA and comprises data on total suspended matter and chl *a* in the water column (Park *et al.*, 2006). For the time frame from 2003 to 2005, 80 maps of chl *a* and 90 maps of total suspended matter were available. These maps were reduced to 3 maps for each variable, containing biologically relevant information: the minimum, maximum and average values. This data reduction technique is often applied in ecological modelling (Loiselle *et al.*, 2008; Cunningham *et al.*, 2009; Echarri *et al.*, 2009). Through sedimentation and degradation, chl *a* and total suspended matter enrich the bottom organic matter (Druon *et al.*, 2004). This input of organic matter is known to influence nematodes directly, as it serves as a food source (Vanaverbeke *et al.*, 2004b; Franco *et al.*, 2008), or indirectly, as microbial degradation often results in oxygen-stressed sediments (Graf, 1992), which can have a strong adverse effect on nematode diversity (Steyaert *et al.*, 1999).

The second group comprised data on sediment characteristics and bathymetry. The sediment was described by the median grain size and the silt-clay fraction. These maps were supplied by the Renard Centre of Marine Geology, Ghent University (Verfaillie *et al.*, 2006), and by the TNO Built Environment and Geosciences 'Geological Survey of the Netherlands'. The bathymetrical data were provided by the Ministry of the Flemish Community Department of Environment and Infrastructure, Waterways and Marine Affairs Administration, Division Coast, Hydrographic Office and completed with data from the Hydrographic Service of the Royal Netherlands Navy and by the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management. The silt-clay fraction and the median grain size are important factors determining meiobenthic diversity (Heip *et al.*, 1985; Steyaert *et al.*, 1999; Vanaverbeke *et al.*, 2002; Merckx *et al.*, 2009). Depth in shallow waters does not directly affect the nematode community, but it modifies the effects of other factors, such as trophic conditions, sediment properties and current properties.

An overview of the range of the environmental data in the dataset is shown in Table 4.1. The range is calculated for both the database, used to build the model, and for the full coverage maps, used for model application.

		Database			Maps		
Parameter	Description	Min.	Max.	Median	Min.	Max.	Median
S	species richness	1	77	30			
ES(25)	expected species richness	1	19	13			
d50	median grain size (µm)	99	541	261	4	692	317
Silt-clay	silt-clay fraction (%)	0.01	95	1.3	0	84	0.053
TSM_mean	average total suspended matter ($g.m^{-3}$)	1.9	24	8.1	1	24	2.6
TSM_max	maximum total suspended matter ($g.m^{-3}$)	3.8	50	28	2.3	66	7.3
TSM_min	minimum total suspended matter ($g.m^{-3}$)	0.55	10	1.2	0.2	14	0.8
Chl_mean	average chlorophyll *a* ($mg.m^{-3}$)	2	12	4.9	1.3	26	3.2
Chl_max	maximum chlorophyll *a* ($mg.m^{-3}$)	4.3	35	22	2.7	39	12
Chl_min	minimum chlorophyll *a* ($mg.m^{-3}$)	0.04	2.3	1.3	0.04	20	1.1
Depth	depth of the water column (m)	2	44	15	-1.3	53	26
Year	year of sampling	1971	2004	1985			

Table 4.1. Range and median values of diversity indices (first 2 parameters) and environmental variables of the dataset used to build the models (database), and of the environmental variables in the maps (maps)

Diversity indices

As nematodes can occur in large numbers, nematode identification is generally carried out on a subsample. Subsamples consist mostly of 200 individuals, although the exact number of identified nematodes varies between samples. Therefore, we used a diversity measure that is independent of sampling effort: the expected number of species. The ES(*n*) is the expected number of species if the sample were of the smaller size n (Hurlbert, 1971). When individuals are independently sampled with similar probability from a small sample, the expected species richness is (Sanders, 1968; Hurlbert, 1971; Simberloff, 1972):

$$\mathrm{ES(n)} = \sum_{i=1}^{S}\left[1 - \frac{\binom{N-x_i}{n}}{\binom{N}{n}}\right] \qquad \text{(Eq. 4.1)}$$

where N is the total number of individuals in the sample, S is the total number of species (i.e. species richness), x_i is the number of individuals of species i in the sample and n is the number of individuals in the subsample. The term inside the summation sign is the probability that a sample of n individuals will contain species i (Gotelli and Graves, 1996). Previous research on the predictability of nematode diversity showed that models developed for ES(25) yielded good predictions (Merckx *et al.*, 2009).

The species richness (S), although not independent of sampling effort, was included in the analysis as well, since it is the most commonly used index and is representative for most of

the samples because the sampling methodology remained unchanged over the years. The range of diversity indices in the dataset is shown in Table 4.1.

Geostatistical modelling

Geostatistics offers powerful interpolation methods for spatial analyses, especially in a patchy environment. The cornerstone in geostatistics is the modelling of the variogram (Webster and Oliver, 2007). It represents the average variance between observations separated by a distance $\boldsymbol{h}$ and has a strong descriptive and interpretative power for the structure of the spatial variability of a variable. The variogram is estimated by (Journel and Huijbregts, 1978; Goovaerts, 1997):

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{\alpha=1}^{N(\boldsymbol{h})} \{z(\boldsymbol{x}_\alpha + \boldsymbol{h}) - z(\boldsymbol{x}_\alpha)\}^2 \qquad \text{(Eq. 4.2)}$$

with $\gamma(\boldsymbol{h})$ being the variogram for a distance vector (lag) $\boldsymbol{h}$ between observations $z(\boldsymbol{x}_\alpha)$ and $z(\boldsymbol{x}_\alpha + \boldsymbol{h})$ of the diversity at the locations $\boldsymbol{x}_\alpha$ and $\boldsymbol{x}_\alpha + \boldsymbol{h}$ and with $N(\boldsymbol{h})$ being the number of pairs separated by $\boldsymbol{h}$.

A variogram is represented as a graph and reveals the underlying spatial pattern of variables, having more similar values when they are spatially closer. The experimental variogram is a plot of the calculated $\gamma(\boldsymbol{h})$ values versus the lag $\boldsymbol{h}$, while the theoretical variogram is the curve fitted through these points, yielding a continuous function of $\gamma(\boldsymbol{h})$ (Fig. 4.2). This curve fitting procedure is a crucial step in variogram analysis (Webster and Oliver, 2007). Four important variogram characteristics can be derived: the sill, the range, the nugget and the model type. The 'sill' represents the total variance of the variable and is the maximum of the variogram model. The 'range' is the maximal spatial extent of spatial correlation between

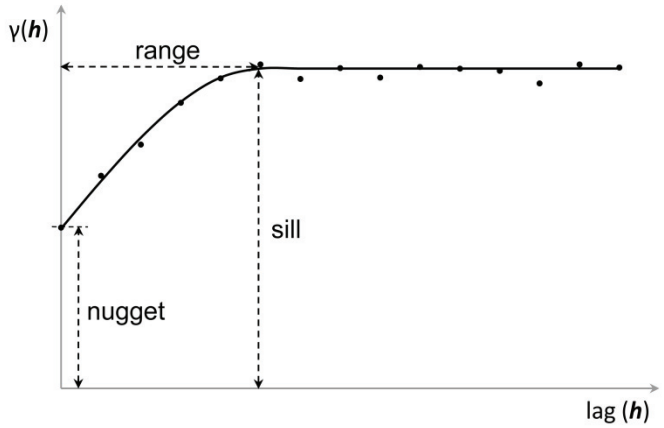


Fig. 4.2. Experimental variogram (dots) and the fitted theoretical variogram (line). The latter is a function describing the degree of spatial dependence of a spatial variable. The plot shows the relation between the semivariance $\gamma(\boldsymbol{h})$ of the variable and the distance $\boldsymbol{h}$ between paired data. The nugget (or nugget variance) is the variance at the limit as the lag tends to zero; the sill is the limit of the variogram at infinite lag distances and the range is the distance at which the difference between the variogram and the sill becomes negligible.

observations of the variable. At lags larger than the range, the expected difference between observations is maximal (being the sill) and independent of the distance. The extrapolation of the variogram model to lags of 0 is called the 'nugget variance' and represents sources of random noise, such as sampling errors and variability at distances closer than the smallest sampling lag. The relative structural variance, i.e. the proportion of total variance that can be attributed to the spatial autocorrelation, can be derived from the variogram: it is the total variance minus the nugget, divided by the total variance. The theoretical variogram can be composed of nested models or structures. Common models are the spherical, exponential, Gaussian and power model. If the variable shows anisotropic variability, directional variograms can be derived (Journel and Huijbregts, 1978; Verfaillie *et al.*, 2006). All variograms were calculated and constructed with the software Variowin 2.2 (Panatier, 1996).

In biological applications it is common to collect replicate samples at nearly the same sampling point to check for local variation in species communities. These replicates have the same geographical coordinates, but are in reality some metres apart, since they result from different drops. Due to limitations of the software, it is impossible to use these replicates in variogram modelling since they have the same coordinates. However, by randomly adding a realistically small variation, within a range of metres, these replicates can be used to accurately estimate the nugget effect.

Two geostatistical interpolation techniques were used in the present paper: OK and RK. OK does not need auxiliary information. However, when such information is available and it is related to the variable of interest, RK often outperforms OK because it exploits this additional information (Hengl *et al.*, 2007).

The OK algorithm uses a weighted linear combination of sampled points situated around the location $\boldsymbol{x_0}$, where the interpolation is conducted. Observations closer to $\boldsymbol{x_0}$ get a higher weight than observations further away. An underlying assumption for OK is that the mean value is locally stationary; thus, it has a constant value inside the interpolation window. The algorithm can be written as:

$$Z^*_{OK}(\boldsymbol{x}_0) = \sum_{\alpha=1}^{n(\boldsymbol{x_0})} \lambda_\alpha \, Z(\boldsymbol{x}_\alpha) \qquad \text{(Eq. 4.3)}$$

with $n(\boldsymbol{x_0})$ being the total number of observations in the interpolation window around $\boldsymbol{x}_0$ and λ_α being the weight attributed to the observation $Z(\boldsymbol{x}_\alpha)$. The weights λ_α are obtained by solving a set of equations involving knowledge of the variogram, and they are chosen in such a way that the prediction error variance is minimised (Webster and Oliver, 2007).

When exhaustive secondary information is available, RK can be used alternatively. Predictions by RK involve 2 steps: first, the relationship between the primary variable and the secondary environmental variables at the sampling locations are modelled by a linear regression, and this model is then applied to the unsampled locations using the environmental variables at this location. Second, the residuals of this linear model are subjected to simple kriging (SK) with an expected mean of 0 (Deutsch and Journel, 1992).

The linear model can be written as a linear combination of the environmental variables:

$\hat{Z}(\boldsymbol{x}_0) = \sum_{k=0}^{p} \hat{\beta}_k \, q_k(\boldsymbol{x}_0) \qquad q_0(\boldsymbol{x}_0) \equiv 1$ (Eq. 4.4)

where $q_k(\boldsymbol{x}_0)$ is the value of the independent variable k at the location $\boldsymbol{x}_0$, $\hat{\beta}_k$ is the estimated regression coefficient of the variable k and p is the number of dependent variables. In the current study, the regression coefficients are estimated in 2 ways: by ordinary least squares (OLS) and generalised least squares (GLS). The latter is an iterative technique (Carroll and Rupert, 1988), which takes the spatial correlation between observations into account (Cressie, 1993). Different steps were carried out for both linear regression techniques. First, variables where standardised to a mean of 0 and a standard deviation of 1, to understand the relative importance of each environmental variable (Schroeder *et al.*, 1986). Secondly, multicollinear variables with a Pearson product-moment correlation coefficient of >0.8 were removed, and, ultimately, by backwards selection, only the highly significant terms were retained from a second-order full model. Normality of the residuals was checked at a significance level of $p < 0.05$. In case no normality was found, the primary variable was log-transformed.

As mentioned before, RK combines 2 approaches: linear regression (the first term in Eq. 4.5) and simple kriging with an expected mean of 0 for the residuals of the linear model (the second term in Eq. 4.5). Thus, the complete model can be written as (Hengl *et al.*, 2007):

$\hat{Z}(\boldsymbol{x}_0) = \sum_{k=0}^{p} \hat{\beta}_k \, q_k(x_0) + \sum_{\alpha=1}^{n(x_0)} \lambda_\alpha \, e(\boldsymbol{x}_\alpha) \qquad q_0(\boldsymbol{x}_0) \equiv 1$ (Eq. 4.5)

where $e(\boldsymbol{x}_\alpha)$ is the residual at location $\boldsymbol{x}_\alpha$.

Validation

A quality control of the different models was performed using a validation dataset containing 30% of all the samples, leaving 70% of the data for the training set. Samples were randomly assigned to the validation set; however, replicates were kept in the same dataset. Replicate samples are, due to spatial autocorrelation, more alike than other samples. If replicate samples are distributed over both sets, the values of the validation set will be predicted accurately, since similar values are present in the training set. This will result in overly optimistic model statistics. As a consequence, keeping replicates in the same dataset will give a more realistic estimation of the accuracy of the geographic interpolation. This validation dataset was used exclusively at the completion of the analysis to compare the performance of the different modelling techniques. Therefore, 5 statistics were calculated: the mean estimation error (MEE) (Eq. 4.6), the root mean-square estimation error (RMSEE) (Eq. 4.7), the mean absolute estimation error (MAEE) (Eq. 4.8), the Pearson product-moment correlation coefficient and the Spearman rank correlation coefficient.

$MEE = \frac{1}{n} \sum_{\alpha=1}^{m} \left(z^*(\boldsymbol{x}_\alpha) - z(\boldsymbol{x}_\alpha) \right)$ (Eq. 4.6)

$RMSEE = \sqrt{\frac{1}{n} \sum_{\alpha=1}^{m} \left(z^*(\boldsymbol{x}_\alpha) - z(\boldsymbol{x}_\alpha) \right)^2}$ (Eq. 4.7)

$MAEE = \frac{1}{n} \sum_{\alpha=1}^{m} |z^*(\boldsymbol{x}_\alpha) - z(\boldsymbol{x}_\alpha)|$ (Eq. 4.8)

where m is the number of validation points, $z(\boldsymbol{x}_\alpha)$ is the measurement and $z^*(\boldsymbol{x}_\alpha)$ is the estimation at the same location. The MEE determines the degree of bias in the estimates; the RMSEE, like the MAEE, evaluates the magnitude of the average error; however, the latter is less sensitive to outliers. The Pearson product-moment correlation coefficient indicates the strength of the linear relationship between the predicted and the observed values of the validation set, and the Spearman rank correlation coefficient is the non-parametric estimation of the correlation between the observed and predicted values. Another way to analyse the validation error is by applying Chebyshev's inequality theorem. According to his theorem the proportion of normalised errors should be ≤1/9 (Hengl, 2007).

Practically, 5 different models were compared with this validation set: the model obtained by OK, the linear models determined by OLS and GLS without kriging (the first term in Eq. 4.5), and both linear models combined with kriging (both terms in Eq. 4.5). Based on the results of this validation set, the models with the best values for the test statistics were selected, and the whole dataset was used to create the final maps.

RESULTS

Linear regression with OLS and GLS

For each diversity index, 2 linear models are constructed (Table 4.2). After selecting the most significant variables, the final models include only 2 or 3 variables, namely the silt-clay fraction, minimum values of total suspended matter and the year of sampling. It is clear from Table 4.2 that the silt-clay fraction has the strongest explanatory power of all models. At least 10 observations are recommended per predictor to prevent overfitting (Hengl *et al.*, 2007). This condition is clearly met: in the training set there are 406 observations and 5 predictor variables. The coefficients of the 2 regression techniques are similar, indicating that no strong spatial clustering is present between the points (Hengl *et al.*, 2004).

Model		Intercept	Silt-clay	(Silt-clay)²	TSM_min	Year	R²
Log(S)	OLS	2.76	-0.81	0.34	-0.33	0.15	0.68
	GLS	2.75	-0.87	0.33	-0.28	0.1	0.66
ES(25)	OLS	9.53	-5.36	1.73	-1.57		0.73
	GLS	9.49	-5.33	1.59	-1.36		0.73

Table 4.2. Estimates of the regression coefficients of the linear models and coefficient of determination (R^2) of the models. The logarithm of the species richness, log(S), and the expected species richness, ES(25), were modelled by ordinary least squares (OLS) and generalised least squares (GLS). See Table 4.1 for definitions of parameters.

Variogram analysis

For each diversity index 3 variograms are modelled: 1 used for OK and 2 for RK. The latter 2 are inferred from the residuals of both linear regression techniques, OLS and GLS (Fig. 4.3). Directional variograms that were apparent in earlier research on the Belgian Continental Shelf (Verfaillie *et al.*, 2006) do not improve model performance and are omitted from the results. The variograms for OK reveal a strong spatial structure for both diversity indices, with a range of >40 km and a relative structural variance of almost 90%, which indicates that a large fraction of the total variance can be linked to spatial processes (Table 4.3). For this database, replicate samples were taken within ranges of metres, while the total area has a maximum cross section of 250 km; thus, the variation between replicate samples is at the same time an accurate estimate of the nugget. The variograms used for RK and inferred from the residuals of the linear regression models show a significant decrease in range and sill for both diversity indices. This reflects the effect of the linear regression: a considerable amount of variation in the data is explained by the linear regression, and the extent of the spatial dependency of the residuals is much smaller than that of the original diversity index. For instance, the initial relative structural variance is about 90% for both parameters and decreases to about 65-69% with GLS. The range decreases by 84% for ES(25) and with >90% for S.

For ES(25) both modelling techniques result in the same decrease of the relative structural variance. For the species richness, however, there is a marked difference between GLS and OLS: GLS is able to explain more of the relative structural variance than OLS, indicating a better performance of this modelling technique.

			Nugget	Range (km)	Sill	Relative structural variance (%)	Model
S	OK	S	44	42.7	320	88	spherical
	RK	residuals OLS	37	3.2	90	71	spherical
	RK	residuals GLS	34	2.6	76	69	spherical
ES(25)	OK	ES(25)	4.3	47.5	36.5	89	spherical
	RK	residuals OLS	3.1	13.6	5.8	65	spherical
	RK	residuals GLS	3	10.9	5.6	65	spherical

Table 4.3. Variogram parameters (see Fig. 4.3 for description). OK: ordinary kriging; RK: regression kriging; OLS: ordinary least squares; GLS: generalised least squares.

Independent validation

The independent validation set enables us to compare the efficiency of the different modelling techniques. Both regression techniques improve the model for ES(25) and S considerably compared to OK (Table 4.4). Therefore, the relation between the diversity and the environmental variables exists and explains a considerable amount of the variation in diversity. Nevertheless, there is still some spatial pattern present in the residuals; hence,

other unknown or fine-scaled factors may contribute to the geographical distribution of nematode diversity.

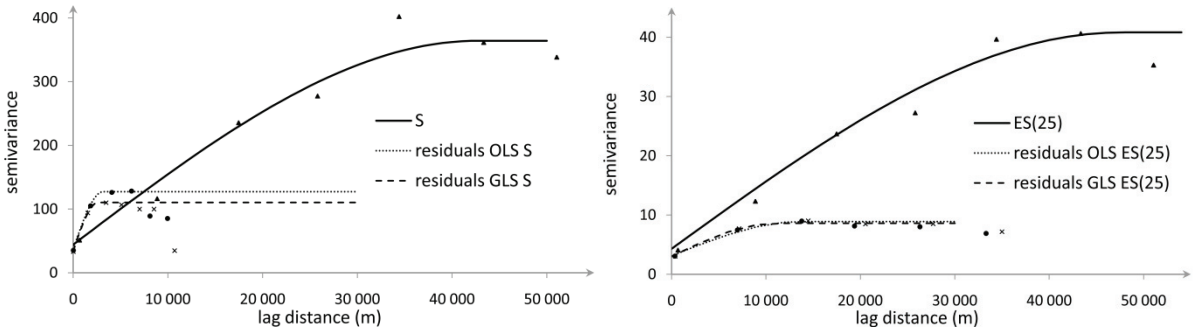


Fig. 4.3. Theoretical variograms of S (left) and ES(25) (right) (see Table 4.1) fitted to the experimental values of the diversity index and used for ordinary kriging (▲), to the residuals of the ordinary least squares (OLS) model (●), and to the residuals of the generalised least squares (GLS) model (×). The latter 2 are used for regression kriging.

Comparison of both regression techniques for both diversity indices shows that GLS offers the best models, kriging with GLS enhances some, but not all, of the model performance indices. The major improvement is found for the MEE; thus, kriging can account for the remaining bias of the linear models. For species richness, there is a clear difference between the Pearson product-moment correlation and Spearman rank correlation coefficients, indicating that there are some strong outliers present in the residuals. This is less pronounced for ES(25). Overall, higher correlation values were found for ES(25). All models meet Chebyshev's inequality condition, implying that there are not an unusually high number of locations where the errors are much higher than at other stations.

Biodiversity index	Linear model	Kriging technique	MEE	RMSE	MAEE	Pearson	Spearman
S	No model	OK	0.39	14.99	11.35	0.37	0.83
	OLS	No kriging	1.17	12.41	10.08	0.61	0.85
	OLS	RK	0.56	12.35	9.86	0.60	0.85
	GLS	No kriging	-0.53	**11.76**	9.46	**0.63**	0.86
	GLS	RK	**-0.37**	11.84	**9.42**	0.62	**0.87**
ES(25)	No model	OK	**0.02**	3.63	2.52	0.57	0.87
	OLS	No kriging	-0.26	2.67	2.14	0.80	0.84
	OLS	RK	0.04	2.56	2.04	0.80	0.87
	GLS	No kriging	-0.51	2.65	2.14	**0.81**	0.85
	GLS	RK	-0.06	**2.55**	**2.03**	0.80	**0.88**

Table 4.4. Statistics of predicted and observed values of the independent validation set. Best values for each diversity index are in bold. MEE: mean estimation error; RMSEE: root mean-square estimation error; MAEE: mean absolute estimation error; OK: ordinary kriging; RK: regression kriging. See Table 4.1 for further definitions.

Final maps

Final maps were constructed with all the available data (Fig. 4.4), resulting in 2 similar charts: near the Belgian coast there is a very low diversity. On average, only 9 species per sample were found in this region and ES(25) is about 4.4, while, for the whole region, an average number of 30 species per sample were found, yielding an average ES(25) of 11.8. Further offshore, the diversity increases considerably. Within this diverse area, there are small patches with high and low diversity, resulting from individual sampling points with higher or lower diversity than the surrounding samples. The range of these patches is larger for ES(25), since the range of the spatial dependency of the residuals is larger for this index.

DISCUSSION

Linear regression

The final linear regression functions all comprise a linear and quadratic function of the silt-clay fraction, which results in a positive parabola with a minimum diversity situated around 60% silt-clay. Consequently, the influence of silt-clay is not unequivocal; when the silt-clay fraction exceeds this threshold, the influence becomes less detrimental. This is contradictory to the general belief that the silt-clay fraction has a purely adverse effect on nematode diversity (Heip *et al.*, 1985; Vanreusel, 1990; Vanaverbeke *et al.*, 2002). However, in the case of strongly oxidised sediments, a positive relation between the silt-clay fraction and nematode diversity has been reported before (Steyaert *et al.*, 1999). A full coverage map of the redox potential was not available; however, organically enriched benthic environments are often encountered in areas with a high load of total suspended matter (TSM). High TSM values result in a reduced environment, and low values may permit highly oxidised sediments. Consequently, the negative correlation of species diversity with TSM may account for this effect.

The linear models indicate that in recent years the observed species richness (S) has increased. But this ‘effect’ is observed because in the last decade only environments with <20% silt-clay, thus with high species richness, were sampled. However, the relationship between the year of sampling and ES(25) is not significant. Similarly to S, low values of ES(25) were not found during the last decade, but the maximum values stayed almost the same over the whole period, remaining at a value of 20. This is due to the fact that ES(25) is a standardisation technique and is bound to an upper limit of 25.

Both diversity indices represent different aspects of the nematode assemblages: ES(25) is strongly influenced by the evenness of the nematode assemblage and to a lesser extent by species richness. Earlier research (Merckx *et al.*, 2009) already pointed out that evenness results in the best predictive models. Moreover, ES(25) is not dependent on sampling effort if sample area (cross-section) is the same for all samples and can therefore more readily be applied to heterogeneous data, originating from different sources.

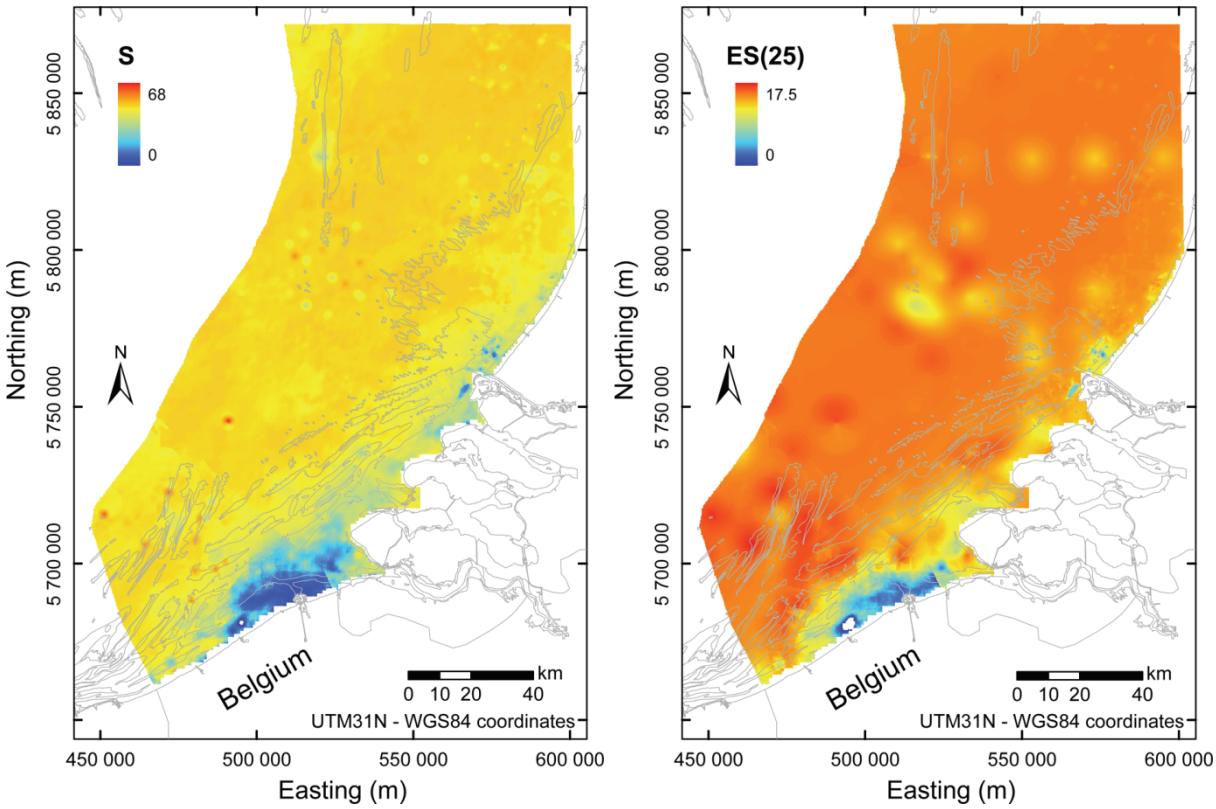


Fig. 4.4. Maps of the generalised least squares models, predicting the nematode diversity after kriging of diversity indices S (left) and ES(25) (right). Grey lines in water are bathymetric lines.

Model comparison

According to the results of the independent validation set (Table 4.4), RK performs better than OK in all cases. Therefore, the environmental variables explain a substantial part of the variation in the diversity of the nematode assemblage. The different variogram parameters underpin this result: the sill and range are much smaller for the variograms of the residuals.

The nugget is accurately estimated by including the sample replicates in the analysis. Since the purpose of replicates is indeed to assess the variance between samples, it is, at the same time, an excellent estimate for the local variation between samples. Instead of lumping or averaging these data, it is useful to keep them apart. In this way there are 2 913 station pairs within the smallest range found (2.6 km for the residuals of GLS of S), rather than only 23 pairs within this lag class, if replicates were to be averaged, which is less than the recommended 30 to 50 pairs (Journel and Huijbregts, 1978). In this case, the remaining spatial pattern in the residuals of the species richness would remain undetected and only a nugget effect would have been observed. As a consequence, the best model would have been the linear model without kriging.

Comparing both regression methods points out that, especially for S, GLS outperforms OLS. This is not surprising, since the iterative process of GLS minimises spatial autocorrelation and estimates the regression coefficients more accurately. However, according to Kitanidis

(1993), OLS may often be satisfactory because the iterative process of GLS, after an initial OLS, results in a negligible difference in the regression parameters, which was the case for ES(25) in our research.

Kriging improves, although to a lesser extent, the linear regression models. For data which are unevenly distributed, e.g. the samples on Kwintebank (51°15' N, 2°40' E), kriging has a declustering effect, because it takes both the distance to the interpolation point and the sampling configuration into account. Therefore, it is preferable to non-declustering techniques, such as linear regression (Verfaillie *et al.*, 2006).

Final maps

The resulting maps of species richness S and expected species richness ES(25) look quite similar, although both indices represent different aspects of diversity; S gives an indication of the number of species that are expected to be found in a 10 cm^2 sample in a certain area, while ES(25) expresses both species richness and evenness. Thus, it seems that, in nematode samples in the North Sea, both species richness and evenness increase in offshore regions. This is in strong contrast with the coastal region. Especially the region south of the Scheldt estuary has a very low diversity. This area is characterised by sediments with a high amount of silt-clay and a water column with elevated concentrations of chl *a* and TSM. Low nematode diversity in oxygen-stressed, fine sediments has been described before (Vincx, 1990; Steyaert *et al.*, 1999), but a link, although indirect, between nematode diversity and water column characteristics has never been shown before. Indeed, oxygen stress in marine sediments is caused by the microbial mineralisation of water column-derived organic matter (Graf, 1992), and our model indicates the link between water column processes and benthic diversity patterns.

Limitations to this research

Hengl *et al.* (2007) pointed out some limitations to RK concerning data quality, undersampling and extrapolation. Our data are historical data, supplied by different researchers within the Marine Biology Section of Ghent University. This has the advantage that sampling and identification techniques are similar. However, different types of sampling effort (e.g. small subsamples or complete cores identified) may have been applied depending on the intention of the original research. These differences can influence S, but will only slightly affect ES(25).

The predictive maps are created for a large area and are based on the data from 153 different stations and 562 samples. Variograms are typically derived from 100 to 200 observations, and, the larger the number of stations, the more precise the estimation is (Webster and Oliver, 2007). The results of the validation set indicate that kriging only slightly improves the model, which is probably due to the large average distance between the sampling points. The distance between the sampling points is often larger than the range of spatial autocorrelation of the residuals, so kriging will not alter the values of these points.

Including new data points will result in intersecting ranges for the residuals as well, and kriging will then result in better estimates.

Extrapolation of the model outside the feature area can be interpreted in 2 ways: extrapolation outside the geographical area and extrapolation for unknown environments. Concerning the geographical extrapolation, special caution should be taken when deriving data near the border of an area or in regions where few samples were taken. Regarding the environmental extrapolation, clearly the model is only suitable for known environments. Particularly for this research, all samples were taken with cores in soft sediments. Consequently, in environments where this sampling technique is not applicable, no data are available; therefore, the model is only valid for well-known sandy environments and cannot extrapolate for, e.g., hard substrates. It is clear from Table 4.1 that the data in the dataset cover nearly the complete range in the maps of the environmental variables. Only the most extreme values are not represented in the dataset. For these data, as well as unrepresented combinations of the environmental data, the model should be interpreted with caution.

Another potential issue is the limited variation in the environmental variables for the offshore region and the large distance between the sampling points. Since no environmental parameter could be identified that explains the differences in diversity in this region, the best model is the average value of the diversity indices for this area.

The kriging algorithm is based on the assumption that the measurements at a certain point are error-free, which is usually acceptable given the much larger spatial variability. The station values obtained by the GLS regression are corrected by the kriging algorithm with the *in situ* measured values. Consequently, the stations appear to be spots on the map. To optimise these maps, more relevant environmental variables and more sampling points would be needed in this area.

CONCLUSIONS

The growing need for detailed maps of biodiversity hotspots can be successfully fulfilled by regression and interpolation techniques, such as GLS and RK. When data are assembled from different sources, it is advisable to use diversity indices that are not dependent on sampling effort. In our case, ES(25) resulted in the best models: highest correlations and no outliers. The diversity of marine nematodes is substantially influenced by silt-clay and TSM, which is also reflected in the resulting map with a species-poor area near the Belgian coast.

ACKNOWLEDGEMENTS

This research is funded by the Fund for Scientific Research (FWO) of the Flemish government (FWO07/ASP/174). The authors thank all the data providers! The environmental data were gathered from different institutes: the 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 and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management, for the oceanographic and sedimentological data. Special thanks to the Flanders Marine Institute (VLIZ, www.vliz.be) for help in building the biological database. This research was conducted within the MANUELA framework (www.marbef.org/projects/Manuela), which is a Responsive Mode Project undertaken as part of 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-09033 of MarBEF. This research was also supported by the GENT-BOF Project 01GZ0705 Biodiversity and Biogeography of the Sea (BBSea). We also thank the reviewers for their in-depth questions and helpful suggestions to improve the quality of this manuscript.

CHAPTER 5

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

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

CHAPTER 5

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

ABSTRACT

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

Keywords:

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

INTRODUCTION

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

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

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

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

MATERIALS AND METHODS

Data

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

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

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

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

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

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

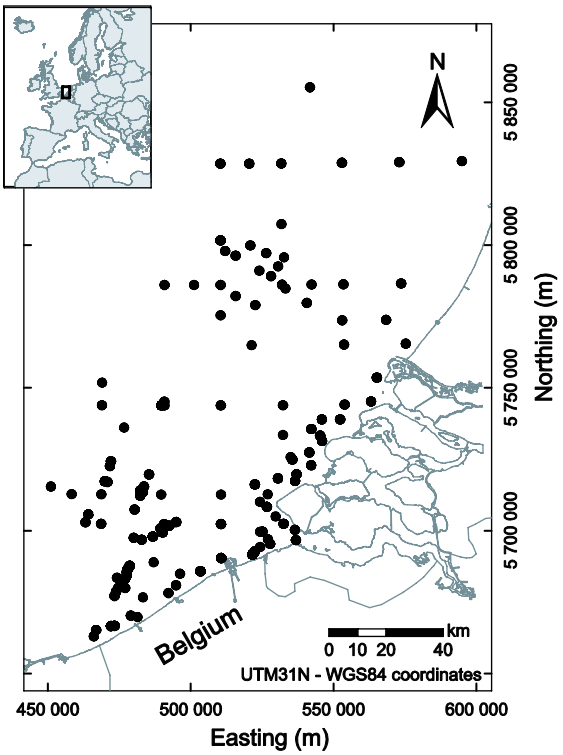


Fig. 5.1. Study area and location of the sampling stations (•).

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

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

Habitat suitability modelling

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

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

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

Validation of the models

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

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

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

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

Null models

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

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

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

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

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

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

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

Species AUC

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

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

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

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

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

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

RESULTS

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

Average of randomisations

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

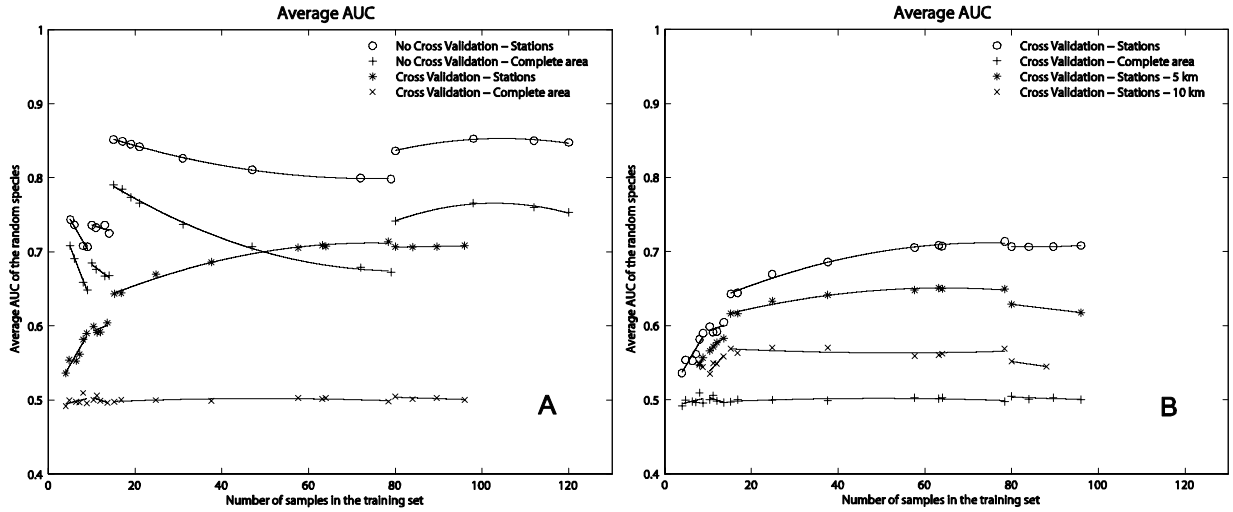


Fig. 5.2. Average AUC of the random models: (A) random samples are selected from the total area or from the sampled stations, both with and without cross-validation and (B) random samples are selected without any restrictions to the sampling distance between the different subsets and with at least 5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

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

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

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

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

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

95% CI of randomisations

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

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

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

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

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

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

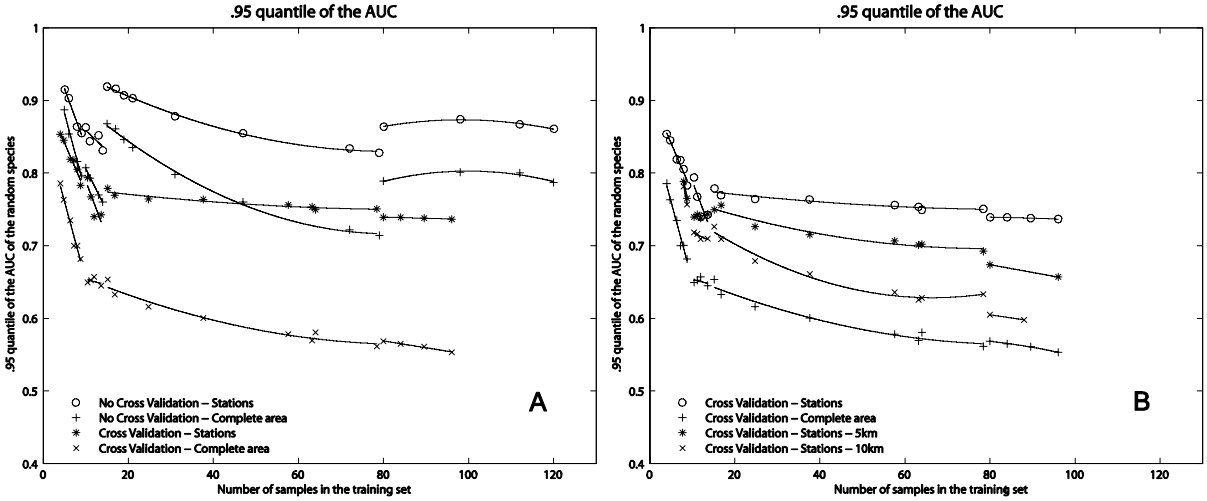


Fig. 5.3. 95% quantile of the random models: (A) random samples are selected from the total area or from the sampled stations, both with and without cross-validation and (B) random samples are selected without any restrictions to the sampling distance between the different subsets and with at least 5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

Selecting non-random species models

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

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

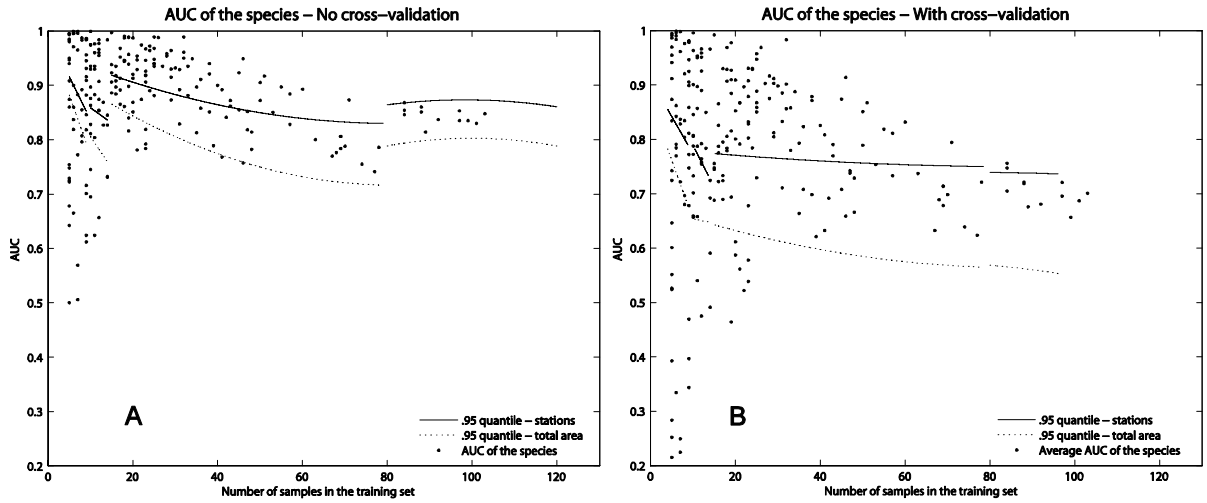


Fig. 5.4. (A and B) Fitted curves of the 95% quantile CI of the null models sampled from the stations and sampled from the total area, with (B) and without (A) cross-validation. AUC values of the species models with (B) and without (A) cross-validation are plotted against these fitted lines (•). Dotted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

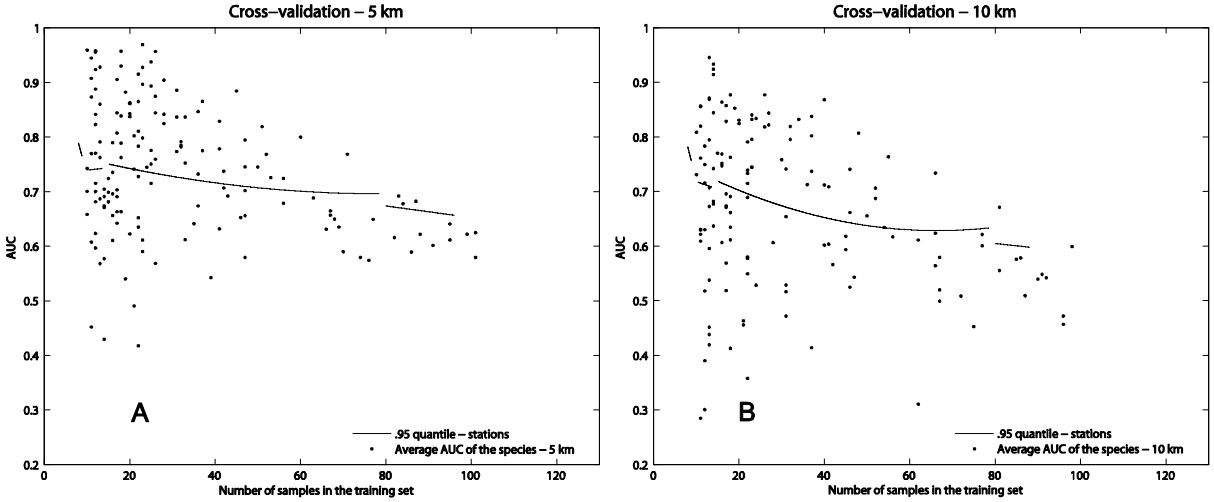


Fig. 5.5. (A and B) Fitted curves of the 95% quantile CI of the null models sampled from the environmentally biased sampling stations with cross-validation and test and training set at least 5 km (A) or 10 km (B) apart. AUC values of the species models with test and training set 5 km (A) or 10 km (B)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

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

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

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

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

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

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

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

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

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

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

DISCUSSION

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

Average of randomisations

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

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

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

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

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

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

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

95% Quantile of the randomisations

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

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

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

Selecting non-random species models

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

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

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

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

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

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

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

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

Drawbacks

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

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

CONCLUSIONS

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

ACKNOWLEDGEMENTS

This research is funded by the Fund for Scientific Research(FWO) of the Flemish Government (FWO07/ASP/174). The authors wish to thank all the data providers! The environmental data was gathered from different institutes: ESA and MUMM/RBINS are acknowledged for providing and processing MERIS data (chlorophyll *a* and TSM data, http://www.mumm.ac.be/BELCOLOUR), the Renard Centre of Marine Geology (RCMG, http://www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of

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

CHAPTER 6

HABITAT SUITABILITY MODELLING OF COMMON SPECIES

Bea Merckx, Maaike Steyaert, Ann Vanreusel, Magda Vincx, Jan Vanaverbeke. Habitat suitability modelling of common species. Submitted to Journal of Sea Research.

CHAPTER 6

HABITAT SUITABILITY MODELLING OF COMMON SPECIES

ABSTRACT

Habitat suitability models get increasing attention in conservation management. Rare and specialist species with specific habitat requirements are generally easier to model than common and generalist species. Since habitat requirements of common species are less stringent, these species have been less considered for species level conservation. However, recent research emphasises the importance of common species which appear in high densities to the structure, function and service provision of terrestrial, freshwater and marine ecosystems. Moreover, separating optimal from suboptimal regions for these species may be interesting for other purposes, such as fisheries. Since habitat suitability models are generally based on presence/absence or presence-only data, they are not able to model densities or relative abundances of a given species of interest. However, maps giving an indication of species relative abundances or total densities can be interesting tools for decision makers. Therefore, we constructed habitat suitability maps of marine nematode species including information on species densities. To reach this goal, we used two approaches: 1) the relative abundances of the species are considered to be separate observations of the species. Thus, the number of observations increased with increasing relative abundance; 2) a species was only considered to be present if its relative abundance was higher than a certain threshold (i.e. 1%, 5% and 10%). We show that implementing a threshold on the relative abundances results in most cases in better models which are capable of identifying the habitats where species occur in higher relative abundances.

Keywords

Habitat suitability modelling, relative abundances, Maxent, common species, Nematoda, North Sea

INTRODUCTION

Species with specific habitat requirements are generally easier to model than generalist species (Segurado and Araújo, 2004; Evangelista *et al.*, 2008; Merckx *et al.*, 2011). Identifying suitable habitats for endangered species gets a lot of attention in conservation management. However, recent work emphasises the importance of common species, species which appear frequently in the environment, for ecosystems too (Gaston and Fuller, 2008). If these common species appear in high abundances (commonness), relatively small declines in their relative abundances may result in large declines in individuals and biomass and may affect ecosystem functioning and provisioning of services (Gaston and Fuller, 2008) such as reduced productivity and higher vulnerability to invasions in plant communities (Smith and Knapp, 2003; Smith *et al.*, 2004). In marine benthic environments simulations show that ecosystem functioning, such as organic matter decomposition and the regeneration of nutrients vital for primary productivity, may be seriously impaired when abundant and common macrobenthic species disappear (Solan *et al.*, 2004).

One step in taking targeted protection measures is the understanding and prediction of species requirements to their habitat. Habitat suitability models (HSMs), as the name reveals, give an indication of which habitats are suitable for a species and which are less suitable. Traditional HSMs based on presence/absence or presence-only data may result in a too broad range of habitat characteristics for a species, reflecting both optimal and suboptimal habitats for the species under consideration (Hutchinson, 1957). Narrowing down the suitable habitat to optimal regions with potentially high densities of harvestable species may also be of interest to fisheries (Houziaux *et al.*, 2010).

HSMs are built mainly based on presence/absence or presence-only data. However, this huge data reduction results in neglecting the information about the densities of a species in a sample. These densities may differ strongly according to the habitat, even for common species. Indeed, species will not necessarily thrive in all occupied habitats. In this paper, we incorporate the information on relative abundances of the species by adding presences in areas where the species has been found in high relative abundances or by using minimum thresholds on species relative abundances.

MATERIAL AND METHODS

Data

The research area, with a total surface of about 18 000 km² is situated in the Southern Bight of the North Sea, near the Belgian and the Dutch coast (latitude 51°6’2’’ - 52°59’19’’ N; longitude 2°14’39’’ - 4°30’43’’ E) (Fig. 6.1). For this area full coverage environmental maps are available.

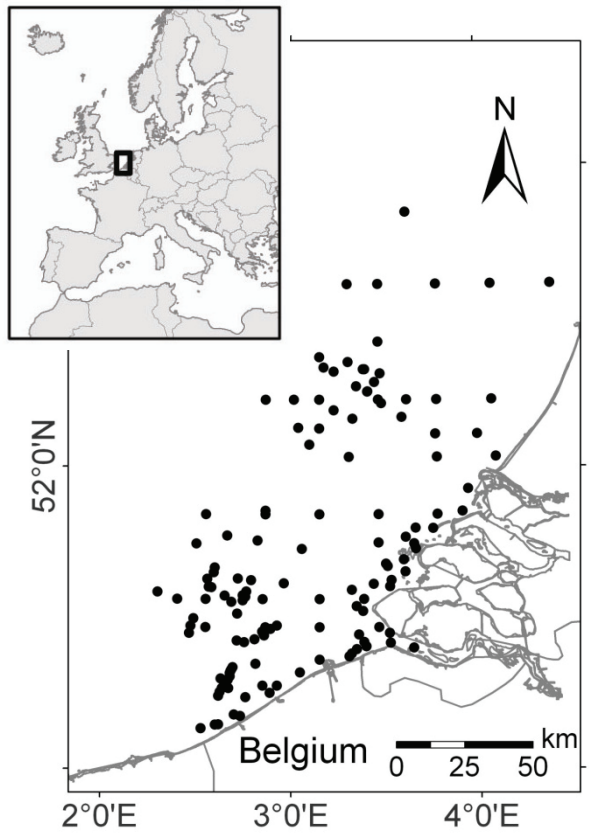


Fig. 6.1. Location of the 140 sampling stations (•).

The performance of the methods applied here, is evaluated by the use of a dataset of free-living marine benthic nematodes from the Southern Bight of the North Sea. Nematodes are usually the dominant taxon within the meiofauna, comprising metazoans passing through a 1 mm mesh sieve but retained on a 38 µm mesh sieve. These free-living roundworms represent the highest metazoan diversity and densities in many benthic environments (Giere, 2009). The nematode data were retrieved from the MANUELA database. MANUELA is, within the EU Network of Excellence MarBEF, a Research Project focusing on the meiobenthic communities. The MANUELA database contains data of meiobenthic species on a broad European scale (Vandepitte *et al.*, 2009). In this case, the data was restricted to the research area (Fig. 6.1). The total densities of the species in a sample are known for 65% of the stations. Leaving out the data of the 35% remaining stations would seriously reduce the available data. Therefore, the relative abundances, which were known for all the stations, were used. In fact, this means that the potential habitat of a species to reach high relative abundances or dominance in a certain area is modelled. Six nematode species which appear in more than 25% of the sampling stations and which appear at least 5 times in high relative abundances (>10%) were selected from the database (Table 6.1). The six species are found across the entire sampling area. All species, except *Dichromadroa cucullata* and *Onyx perfectus*, are known to appear in high relative abundances near the coastal area (Vincx, 1989a; Vanreusel, 1990), a region characterised by low diversity and low evenness (Merckx *et al.*, 2010), hence with dominance of certain nematode species.

Species	% of the stations where the species is observed
Daptonema tenuispiculum (Ditlevsen, 1918)	27
Dichromadora cucullata Lorenzen, 1973	71
Enoploides spiculohamatus Schulz, 1932	72
Onyx perfectus Cobb, 1891	69
Sabatieria celtica Southern, 1914	63
Sabatieria punctata (Kreis, 1924)	33

Table 6.1. Selected nematode species and the percentage of stations where the species is observed.

The seafloor, the habitat of the nematode species, is not at all homogeneous in this area; it is characterised by sand dunes and a wide range of sediment types, varying from muddy to sandy environments (Lanckneus *et al.*, 2002; Verfaillie *et al.*, 2006). The coastal zone is characterised by a high amount of total suspended matter and chlorophyll *a* in the water column and a high silt-clay fraction in the sediment, especially near the mouth of the Scheldt Estuary and the Eastern side of the Belgian coast (Fig. 6.2) (Fettweis and Van den Eynde, 2003; Eleveld *et al.*, 2004).

The environmental variables were retrieved from maps acquired by remote sensing and maps interpolated from data sampled in the field (Fig.6.2). The first group of maps summarises data on total suspended matter and chlorophyll *a* in the water column (Park *et al.*, 2006). The data is collected by remote sensing by the MERIS spectrometer on board of the Envisat satellite of the ESA. Eighty chlorophyll *a* maps and 90 total suspended matter maps were gathered during the time frame 2003-2005. These maps were reduced to three biologically relevant maps revealing the minimum, maximum and average values. This data reduction technique is often applied in ecological modelling (Loiselle *et al.*, 2008; Cunningham *et al.*, 2009; Echarri *et al.*, 2009). Satellite data are restricted to the water column but are of relevance for seafloor inhabiting organisms as sedimentation of chlorophyll *a* and total suspended matter enrich the bottom organic matter (Druon *et al.*, 2004; Franco *et al.*, 2008). The second group contains maps derived from point sampling at sea. It comprises data on sediment characteristics, such as median grain size and the silt-clay fraction, and bathymetry. These maps were supplied by the Renard Centre of Marine Geology, Ghent University (Verfaillie *et al.*, 2006) and TNO Built Environment and Geosciences-Geological Survey of the Netherlands. The bathymetrical data were provided by the Ministry of the Flemish Community Department of Environment and Infrastructure, Waterways and Marine Affairs Administration and completed with data from the Hydrographic Service of the Royal Netherlands Navy and by the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management. The silt-clay fraction and the median grain size are important factors determining the meiobenthic community (Heip *et al.*, 1985; Steyaert *et al.*, 1999;

Vanaverbeke *et al.*, 2002; Merckx *et al.*, 2009). Depth in shallow waters does not directly affect the nematode community, but it modifies effects of other factors such as sea surface temperature, phytoplankton blooms, light penetration, trophic conditions of the benthos and changing water currents. It is clear from Fig. 6.2 that some of the variables have a similar distribution, especially the maps concerning TSM and silt-clay show a strong resemblance. This is not surprising since the silt-clay deposits in front of the Belgian coast can be explained by the combined effect of neap-spring tidal cycles and the presence of TSM in the water column (Fettweis and Van den Eynde, 2003).

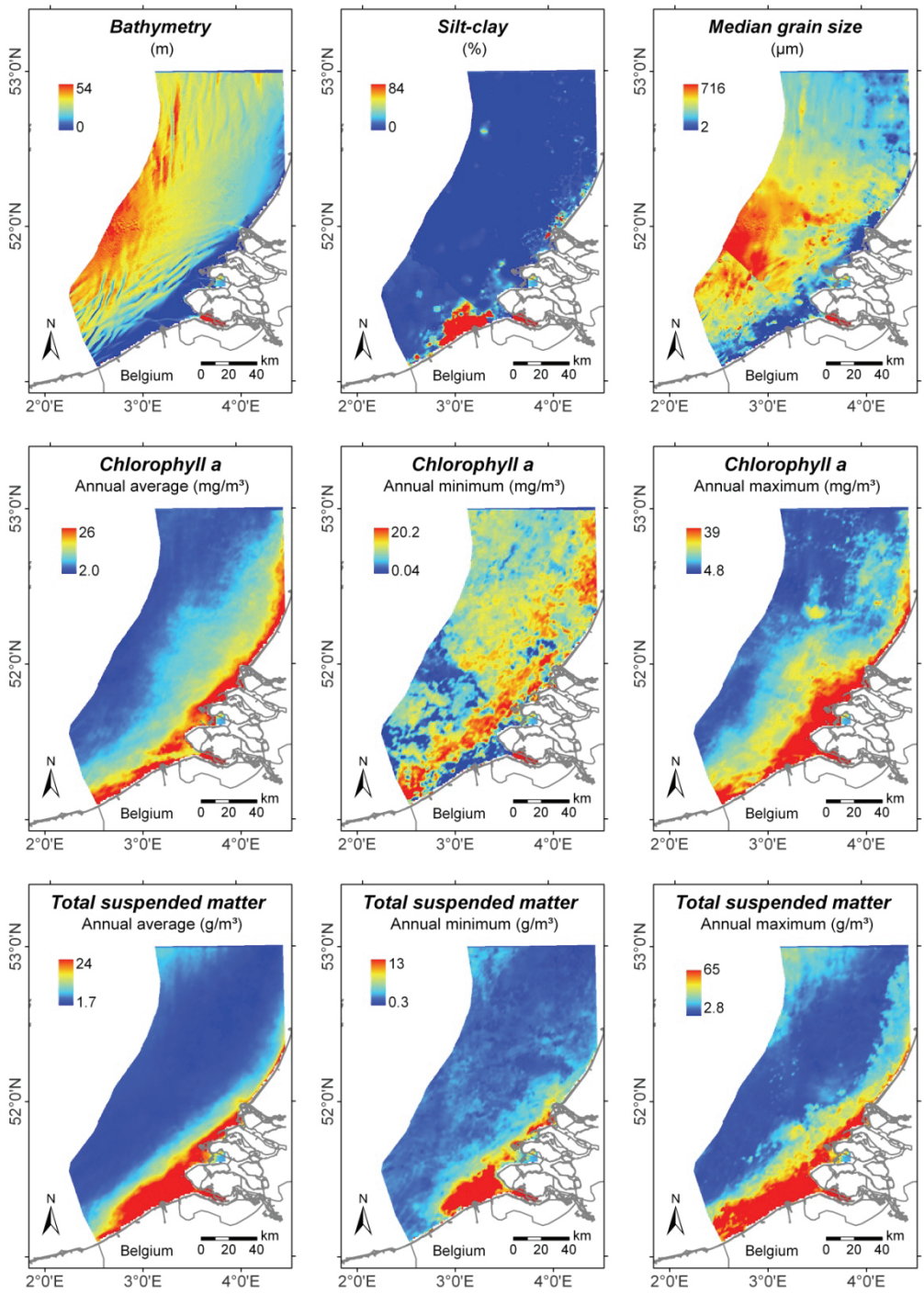


Fig. 6.2. Maps of the environmental variables (sources: see text).

Habitat suitability modelling

Habitat suitability models, attempt to correlate ecological niche elements with species presence and then project this relation into the geographical space to create predictive maps of locations with similar conditions. This approach has been cited by various names such as 'ecological niche modelling', 'species distribution modelling', 'habitat suitability modelling' and 'bioclimatic envelope modelling' (Pearson, 2007). We applied a habitat suitability technique (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005) based on presence-only data because absence of nematode species is rarely 100% certain. Generally, only a subsample is analysed and the presence of a species may not be ascertained. When a species is found presence, it is assumed to be present, notwithstanding the fact that sampling artefacts or erroneous determinations may result in false occurrences. Moreover, species may not have occupied their full niche due to unsaturated populations and inter- and intraspecific interactions (Fielding and Bell, 1997). Furthermore nematodes are known to show a patchy distribution (Li *et al.*, 1997; Somerfield *et al.*, 2007; Gingold *et al.*, 2010a). Earlier research pointed out that Maxent is a reliable presence-only modelling technique and it performs well compared to other presence-only modelling techniques (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Ortega-Huerta and Peterson, 2008; Wisz *et al.*, 2008). These good predictive capacities have been attributed to the ℓ_1-regularisation (see further) which prevents the algorithm from overfitting. Other models often do not apply any form of regularisation, and this can cause the observed difference in predictive performance (Gastón and García-Viñas, 2011). Moreover, Maxent is a generative approach, rather than discriminative. This can be an inherent advantage when the amount of training data is limited (Phillips *et al.*, 2006).

Maxent combines presence-only data with the information of environmental layers using the maximum entropy approach. This algorithm searches the probability distribution which maximises entropy within the constraints of the given data (Phillips *et al.*, 2006). The distribution π assigns a non-negative probability $\pi(x)$ to each point x within the area X, and these probabilities sum to 1. The approximation of π is also a probability distribution, and is denoted as $\hat{\pi}$. The entropy of the set of probabilities $\hat{\pi}(x)$ is defined as $H(\hat{\pi}) = -\sum_{x \in X} \hat{\pi}\,(x).\log_{\mathrm{e}}\,(\hat{\pi}(x))$. H reaches the maximum in the most uncertain situation when a species shows 'maximum entropy' and has the same likelihood across the whole region (Shannon, 1948).

The environmental variables or functions thereof are called the 'features'. These features set limitations to the choice of the probability distribution. The feature types which are used in this study are linear features, quadratic features, product features, threshold features and hinge features. The product features incorporate interactions between predictor variables. Threshold features equal one, once a certain threshold is passed (Phillips *et al.*, 2006). Hinge features, namely the forward hinge feature and the backward hinge feature, are recently introduced features. The forward hinge feature is 0 while the variable is smaller than a threshold h and then increases linearly to 1 at the maximum value of the variable. In a

similar way, the reverse hinge feature is 1 at the minimum value of the variable and drops linearly to 0 at the threshold h (Phillips and Dudík , 2008).

In order to reduce overfitting Maxent applies a penalty term which penalises models with many features (ℓ_1-regularisation). In this way models with fewer features are favoured. Such models are less likely to overfit (Phillips *et al.*, 2006). However, former research indicated that overfitting is still present (Merckx *et al.*, 2011). Therefore we still performed a backward and forward selection of the environmental variables and a feature selection.

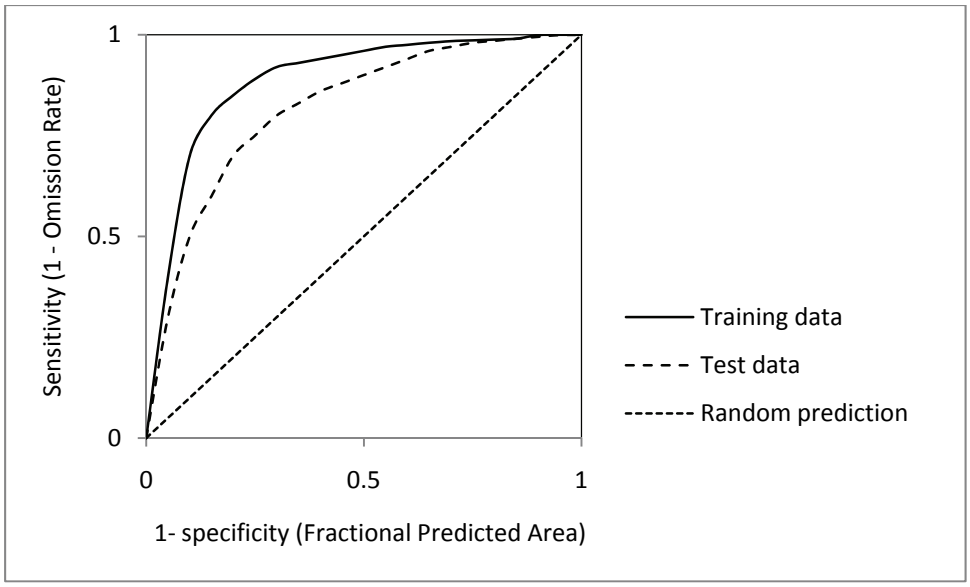


Fig. 6.3. ROC plot for test and training set. The AUC is the area under the ROC plot.

An important quality parameter in habitat suitability modelling is the area under the curve (AUC) (Fig. 6.3). It is a threshold independent measure. For different values of the false positive fraction (1-specificity), the sensitivity values are calculated (true positive fraction). These values are represented in a Receiver Operating Characteristic curve (ROC). The area under this plot equals the AUC. It indicates the overall performance of the model. An AUC of 0.5 indicates that the model predictions do not differ from a random prediction while a maximum value of 1 indicates a perfect model (Fielding and Bell, 1997). However, in presence-only modelling, as is the case for Maxent, there is no absence data available and therefore it is impossible to calculate the false positive fraction. The false positive fraction is the fraction where the species is predicted present, where it is in fact absent. However, this problem is circumvented by distinguishing presence from random, instead of presence from absence. For each AUC-analysis 10 000 pixels are drawn randomly from the study region (Phillips *et al.*, 2006). Then, the true positive fraction is replaced by the 'fractional predicted area', the fraction of the total study area for which the species is predicted to be present. Thus for an ideal model all presences are correctly predicted for an infinitely small predicted area. Consequently, it can be seen that the maximum achievable AUC is less than 1 (Wiley *et al.*, 2003). An AUC of 0.5 still corresponds with a random prediction: predicting x percent of the area as suitable for the species, will result in x percent correctly classified occurrences.

Given the calculation method of the AUC, it is clear that the AUC values of a common or generalist species will rarely be high. Since the species appears on a large fraction of the area, the sensitivity can only be large when a large fraction of the area is predicted. These difficulties of predicting generalist species is not only restricted to presence-only modelling. Evangelista *et al.* (2008) showed that all the modelling techniques they tested, including models based on both presence/absence data and presence-only data, showed difficulties in predicting generalist species. In habitat suitability modelling, data is always reduced to presence-only or to presence/absence data, even if densities are known. Nematode data is mainly obtained from sampling by cores pushed into the sediment. From these cores a subsample of 100 to 200 nematodes is taken and identified. In the database the total nematode densities of the core samples are only known in 65% of the cases. Therefore, the relative abundances of the species were calculated. In this paper, we investigate how the information of the relative abundances can be used to create HSMs which give an indication of the relative abundance of the species. In order to introduce these relative abundances in this presence-only modelling technique, we applied two methodologies: (1) the relative abundances are translated into separate observations of the species in this area (RA). Thus, if the relative abundance of a species is 5 percent, this is translated into 5 observations (Phillips and Pearson, pers. comm.); (2) we constructed 3 presence-only models based on different relative abundance thresholds: a species is considered to be present when it represents at least 1 percent of the local community (T1). The same was done for a threshold of 5 percent (T5) and 10 percent (T10). The performance of these 4 HSMs was then compared with the HSM based on the original presence-only data (PO). Thus, in total for each species 5 models were developed.

Validation of the models

Validation was done in two ways (Fig. 6.4): 1) to find the best model for each of the different datasets, a four-fold cross-validation was applied to 80% of the original data. 2) The remaining 20% of the data was used exclusively at the completion of the analysis to compare the performance of the 5 different habitat suitability models of each species.

The four-fold cross-validation on 80% of the original data is necessary to reduce overfitting. Since overfitting is not completely prevented by the ℓ_1-regularisation (Merckx *et al.*, 2011) a backward and forward selection for each environmental variable combined with a selection of the best feature combination was done. So, the original 80% of the data was used to create four cross-validation models: 75% of this data was used to create the model and the AUC was calculated for the remaining 25% of the data (i.e. the test data). Each quarter of the data was once used as a test set, thus this resulted in four models. The 4 AUC-values of the test sets were averaged and the model with the highest average AUC is selected. This resulted in 5 final models for each species.

These 5 models were then tested against the remaining unseen 20% of the species data, the test set. Two test statistics are calculated: the AUC and the Spearman rank correlation between the relative densities of the test set and the model output.

For the AUC, the test set was reduced to samples with high relative densities. These HSMs were trained for specific thresholds, thus these thresholds were also applied on the test set. This makes sense, since the calculation of the AUC is based on presence data. Suppose that stations with low relative abundances, not reaching the threshold, are kept in the test set and these stations have low probabilities according to the model, this will be considered as a misclassification (false negative), while it is in fact a correct prediction for this threshold. Thus these samples should be removed from the test set. However, a smaller predicted area will generally entail a larger AUC. Therefore, this statistic is only of secondary importance in the model selection.

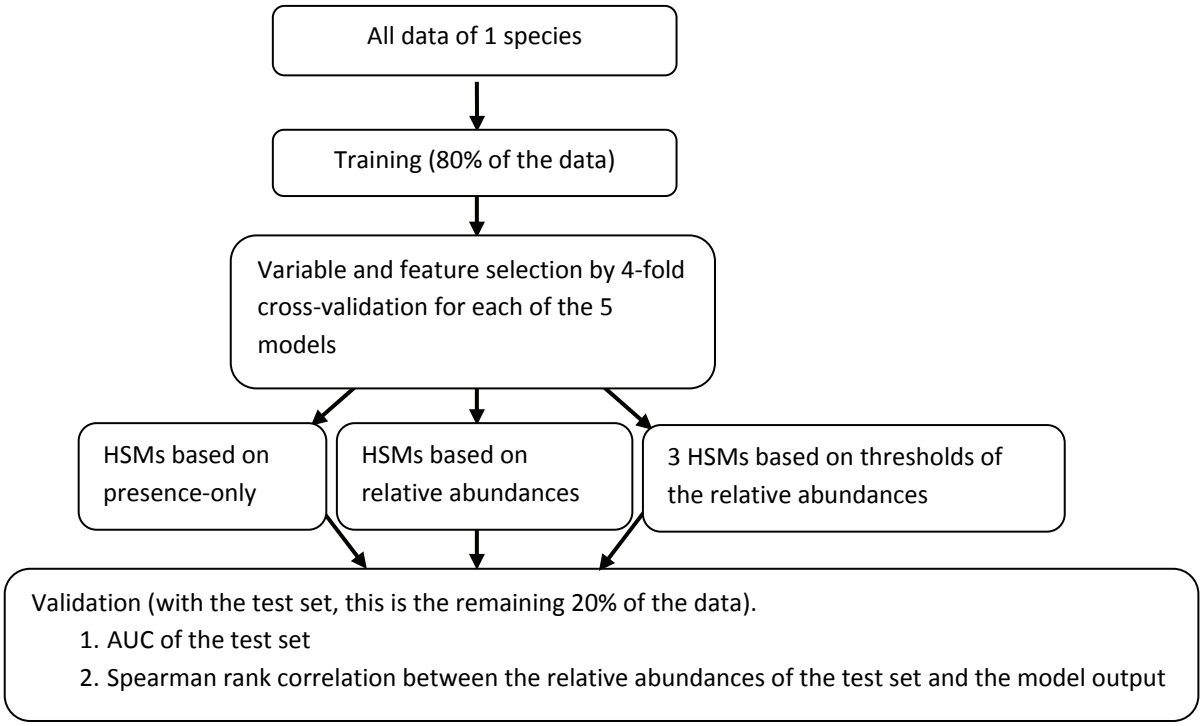


Fig. 6.4. Model validation scheme for 1 species.

More relevant is the Spearman rank correlation between the relative abundances of the test set and the model output. In this case no threshold was applied to the test set. This makes sense since the threshold-based models should predict lower probabilities when lower relative abundances are present and higher probabilities when higher relative abundances are found. Thus, for each species the original test set with all the samples is used to calculate the Spearman rank correlation for each of the 5 models.

Final maps

Based on the results of the final validation, two maps are created for every species: a reference map based on the model for the original presence-only dataset (PO) and a second map for the model showing the highest Spearman rank correlation for the test set.

RESULTS

Model selection

Table 6.2 gives an overview of the correlation between the observed relative abundances of the species in the test set and the output of the selected HSM. It is clear that when all the observations are used to build the model (i.e. in the case of the PO model) the correlation between the relative abundance and the output of the HSM is low and sometimes even negative.

The correlations generally increase with increasing threshold (Table 6.2). Thus it seems that the model is capable in identifying regions with higher relative abundances although less data is furnished to the modelling algorithm. Introducing the relative abundances (RA) as separate observations, results in better correlations than the original PO model in 50% of the cases. When thresholds are applied, the correlation increases five out of six times.

Species	Presence-only	1% threshold	5% threshold	10% threshold	Relative abundance
Daptonema tenuispiculum	-0.02	0.07	0.14	**0.31**	-0.07
Dichromadora cucullata	0.00	0.03	0.45	**0.59***	-0.1
Enoploides spiculohamatus	0.32	0.27	0.68*	**0.70***	0.39
Onyx perfectus	0.31	0.40	**0.55***	0.39	0.45*
Sabatieria celtica	0.41	0.30	0.28	0.11	**0.69***
Sabatieria punctata	0.68*	**0.71***	0.61	0.54	0.59

Table 6.2. Spearman rank Correlation coefficients between relative abundances of the test set and the predicted values for the samples of the test set. Values in bold indicate the highest correlation coefficients for the species. Significant correlations ($p<0.05$) are indicated with an asterisk.

	Presence-only	1% threshold	5% threshold	10% threshold	Relative abundance
Daptonema tenuispiculum	0.86	0.89	**0.94**	0.92	0.74
Dichromadora cucullata	0.58	0.68	0.51	**0.84**	0.48
Enoploides spiculohamatus	0.67	0.67	0.95	**0.97**	0.74
Onyx perfectus	0.66	0.59	**0.92**	0.92	0.8
Sabatieria celtica	0.79	0.76	0.85	**0.9**	0.88
Sabatieria punctata	0.94	0.94	0.95	**0.95**	0.95

Table 6.3. AUC of the test set. Values in bold indicate the highest AUC for the species.

The AUC of the test set also reveals higher values when thresholds are applied (Table 6.3). The AUC of the presence-only models of *Dichromadora cucullata*, *Enoploides spiculohamatus*

and *Onyx perfectus* is too small to be considered as an informative model (Merckx *et al.*, 2011). The presence-only models are depicted in Fig. 6.5 and 6.6, but only to compare the output with the models based on the relative abundances. Moreover, the comparison of the AUC values of the different models should be interpreted cautiously; the interpretation is not as straightforward as for the Spearman rank correlations. The number of samples in the tests set decreases as the threshold increases. It is the purpose of this modelling exercise to restrict the modelled suitable habitats to the actual habitats where the species can thrive at high relative abundances. As mentioned before, a restricted area will often result in an increase in the AUC. Indeed, since there is a reduction in the number of observations, there is an increasing chance that these observations are found within a restricted area, resulting in a small fractional predicted area containing all the observed presences, and thus having a high specificity. The increase in AUC does not necessarily mean a better performance of the model. Thus, in this case the Spearman rank correlation will give a better indication of the performance.

For the four-fold cross-validation it does make sense to use the AUC as a quality parameter, since the models which are compared are constructed with the same dataset, which have obviously the same distribution pattern. Thus if different models are based on the same dataset, a higher AUC will indeed indicate a better overall performance of the model.

	Model	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Daptonema tenuispiculum	10%							100% ↗		
Dichromadora cucullata	10%	23% ↘		8% -		35% ↗		30% ↘		
Enoploides spiculohamatus	10%		100% ↗							
Onyx perfectus	5%	19% ∩	6% ↗			54% ∩		17% ∩		
Sabatieria celtica	RA			5% ↗				37% ∩	56% ∩	
Sabatieria punctata	1%				15% ↘				81% ↗	

Table 6.4. Estimate of the relative contributions of the environmental variables to the final models. Only variables contributing more than 5% to the model are shown. Positive (↗), negative (↘) and optimum (∩) correlations are represented.

Final models

The variable contributions of the thresholded models are shown in Table 6.4. Average TSM contributes strongly to the model of *Daptonema tenuispiculum*. *Sabatieria celtica* and *Sabatieria punctata* seem to be strongly influenced by the maximum TSM level. *Enoploides spiculohamatus* shows a strong positive relation with maximum chlorophyll *a*. While species

found in high abundances off-shore such as *Dichromadora cucullata* and *Onyx perfectus* show a relation with water depth. As mentioned in the materials and methods section: the selected variables may represent proxies for other variables. The λ-values describing the thresholded models can be found in Addendum 4.

Fig. 6.5 and 6.6 show the resulting maps. It is clear that the models resulting from the data with the relative abundances thresholds result in narrower distribution patterns. This was to be expected: the number of samples in the model decreased, most probably resulting in smaller ranges of environmental variables.

DISCUSSION

Model selection

According to Table 6.2 the habitat where the species thrives is not favoured by the models based on the occurrence data. This makes sense since this model reflects which habitat is potentially suitable. A single occurrence of a species is considered equally important as high relative abundances of a species. The model does not differentiate between optimal and suboptimal habitats.

Using RA generally increases the correlation between the model output and the relative abundances in the test set. Thus, this methodology may identify regions where the species is found in higher relative abundances, but needs to be examined for each independent case. Applying thresholds seems to have more potential in differentiating between habitats where species occur in high and low relative abundances.

Final maps

In order to analyse if the resulting patterns are also realistic patterns, the habitat preferences of each species were compared with literature sources. It should be noted that some of the data of the cited literature sources (Vincx, 1989a; Vincx *et al.*, 1990; Vanreusel, 1990; Vanaverbeke *et al.*, 2007; Vanaverbeke and Vincx, 2008; Vanreusel, 1991) are actually used for building the models and we are aware that circle reasoning should be avoided when interpreting results. However, it is the first time that the data of 17 different studies are combined in one single analysis and finding the same patterns may reconfirm and strengthen the findings in the literature sources.

Daptonema tenuispiculum is a non-selective deposit feeder (1B) (Wieser, 1953; Vincx and Heip, 1987). The species is often found in stressed environments and can survive in sediments with low oxygen content (Boyd *et al.*, 2000). It is common in the mouth of Western Scheldt (Vincx and Heip, 1987), in unstable sediment due to fluctuating current velocities (Vanreusel, 1990) and dredged material disposal sites (Boyd *et al.*, 2000). *Daptonema tenuispiculum* is found in high densities in fine medium sand with a high amount of silt (44.6%) and organic carbon (Vincx, 1989a; Vanreusel, 1990; Schratzberger *et al.*,

2000b). It is dominant at the mouth of the Western Scheldt and at the Belgian coast, except for the Western coast (Vincx, 1989a). It is clear from Fig. 6.5 that the literature sources are confirmed by the model with the 10% threshold: the regions with high TSM-values are highlighted. TSM is a proxy for silt-clay and may indicate elevated levels of organic carbon. The original model with the presence data describes the area where the species is found well (Table 6.3), but it is less clear where high relative abundances can be found. The modified model with the 10% threshold results in more or less the same distribution of the species, but highlights the region where the species is found in high relative abundances.

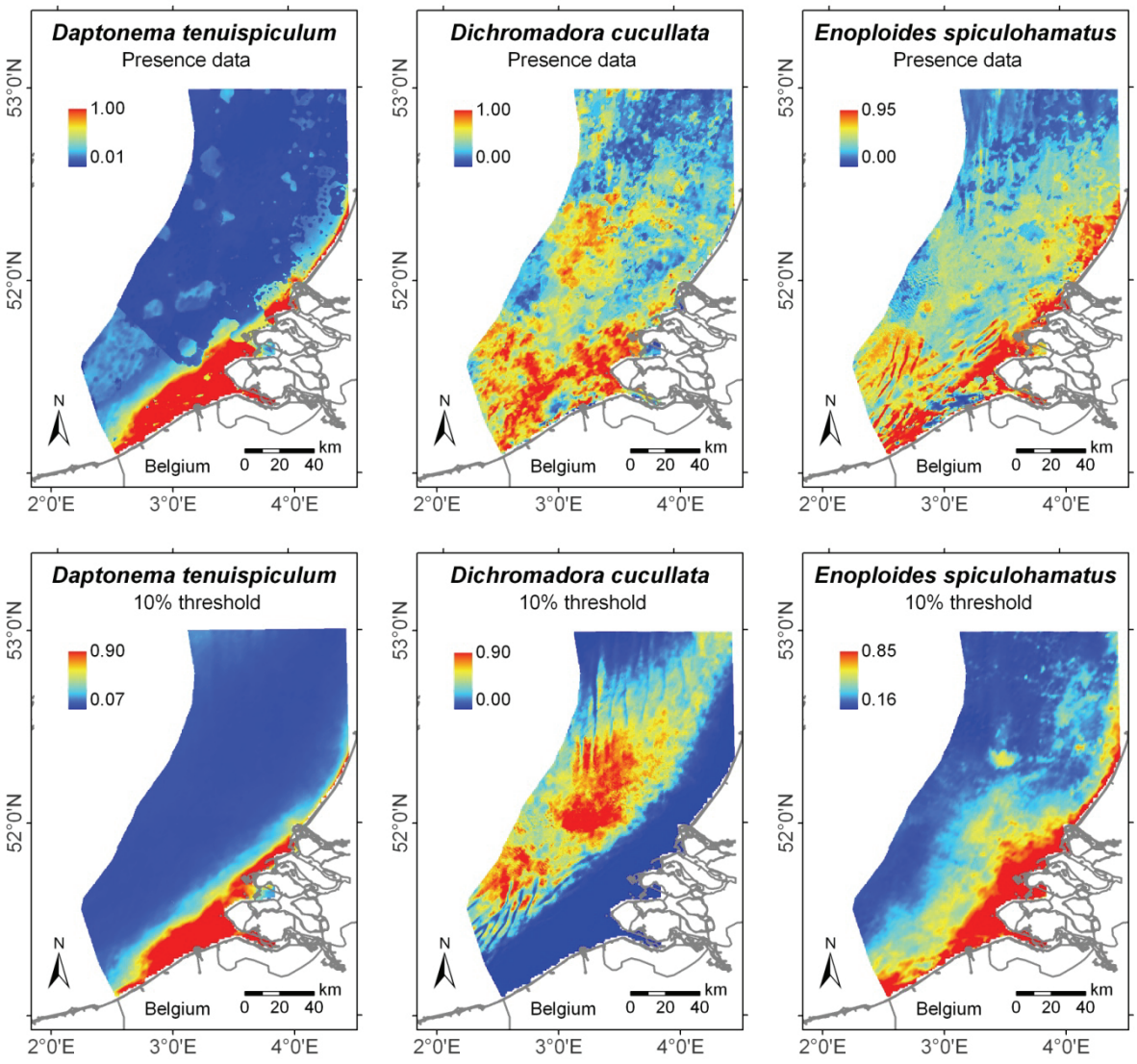


Fig. 6.5. Comparison of the two resulting models using all the presence data and the data incorporating relative abundances for Daptonema tenuispiculum*,* Dichromadora cucullata *and* Enoploides spiculohamatus*.*

Dichromadora cucullata is an epigrowth feeder (2A) (Wieser, 1953; Heip *et al.*, 1983; Vincx, 1989a) and is a common species (Vincx *et al.*, 1990). However it can be found in higher relative abundances in more offshore stations where sediments consist of clean medium sand and little gravel (Vincx, 1989a) and in coarse sand (Lorenzen, 1974). The original model

(Fig. 6.5) confirms the statement that *D. cucullata* is a common species. Limiting the samples to stations where high relative abundances (>10%) were observed restricts the suitable habitats to stations offshore, the region where sediments with higher median grain size prevail.

Enoploides spiculohamatus is a predator 2B (Wieser, 1953; Vincx, 1989a) and a common species (Vincx *et al.*, 1990) which is frequently found in fine medium sand with a low amount of silt (<5%) (Vincx, 1989a) and low chl *a* content in the water column (Vanreusel, 1990), or in coarse sand (Vanaverbeke *et al.*, 2002). It is mainly found along the Dutch coast and the western part of the Belgian coast (Vincx *et al.*, 1990). Both maps display a wide geographical range for this species. Restricting the observations to a threshold of at least 10%, results in a model that shows higher relative abundances along the coast. The original model with presence data clearly leaves out the part of the Belgian coast where a high amount of silt and clay is found (Eastern part of the Belgian coast), which is consistent with the literature sources, while the adjusted model does not exclude this region (Fig. 6.5). This is clearly an artefact of the adjusted model. Moreover, the species seems to occur in regions with high chl *a* content in the water column (Table 6.4), which seems to be in contradiction with the observations of Vanreusel (1990).

Onyx perfectus is a predator (2B) and is found in high relative abundances on the crests of the sand banks (Vincx, 1989a; Vanaverbeke *et al.*, 2007; Vanaverbeke and Vincx, 2008) and stations characterised by medium sand almost without silt (Vincx, 1989a; Vanaverbeke *et al.*, 2002). It is a very rare species in sediments containing more than 5% silt (Vanreusel, 1991). The amount of organic carbon (Vincx, 1989a) and chl *a* (Vanreusel, 1990) can be high. It is also found in high relative abundances in the gullies between the sandbanks of the Belgian Continental Shelf. These gullies are characterised by coarse sediments and high gravel content (Vincx, 1989a). The original model (Fig. 6.6) shows a broad geographical range of the species. The data with the 5% threshold results in a model where the suitability of the habitat is restricted to the sand banks and the gullies in between (Fig. 6.6). And Table 6.4 confirms the positive relation with increased chl *a*, which is consistent with the literature sources.

Sabatieria celtica is a non-selective deposit feeding nematode species (1B) (Wieser, 1953; Vincx, 1989a). It prefers fine to medium sand with a low amount of silt (<5%) (Vincx, 1989a; Soetaert *et al.*, 1995) but can also be found in lower densities in both silty environments and coarse sand (Lorenzen, 1974; Vanreusel, 1990; Vanreusel, 1991). It is mainly found at the Dutch coast and the western part of Belgian coast (Vincx *et al.*, 1990). This is the only model, where the best correlation is found when the relative abundances of the species are introduced as different samples (RA). The modified model (Fig. 6.6) restricts the suitable habitats to the coastal zones and the highest relative abundances are indeed found near the Dutch coast and the Western part of the Belgian coast while for the original model this differentiation is not clear (Fig. 6.6).

Sabatieria punctata is a non-selective deposit feeder (1B) (Wieser, 1953; Vincx and Heip, 1987) which is often found in stressed environments: in dredged material disposal sites (Boyd *et al.*, 2000) and in unstable sediments due to fluctuating current velocities (Vanreusel, 1990). It seems to thrive in fine medium sand with a high amount of silt and organic carbon (Vincx, 1989a; Vanreusel, 1991; Soetaert *et al.*, 1995). It is mainly found at the mouth of the Western Scheldt (Vincx and Heip, 1987) and at the Belgian coast, except for the Western part (Vincx, 1989a). The model selects TSM (Table 6.4) as most important factor which contributes to the presence of silt-clay and organic carbon. Both *Sabatieria punctata* models delineate the coastal zone. There is not so much difference between both models (Fig. 6.6), only at the mouth of the Scheldt estuary there is a small increase in suitability for the modified model, based on data with 1% threshold.

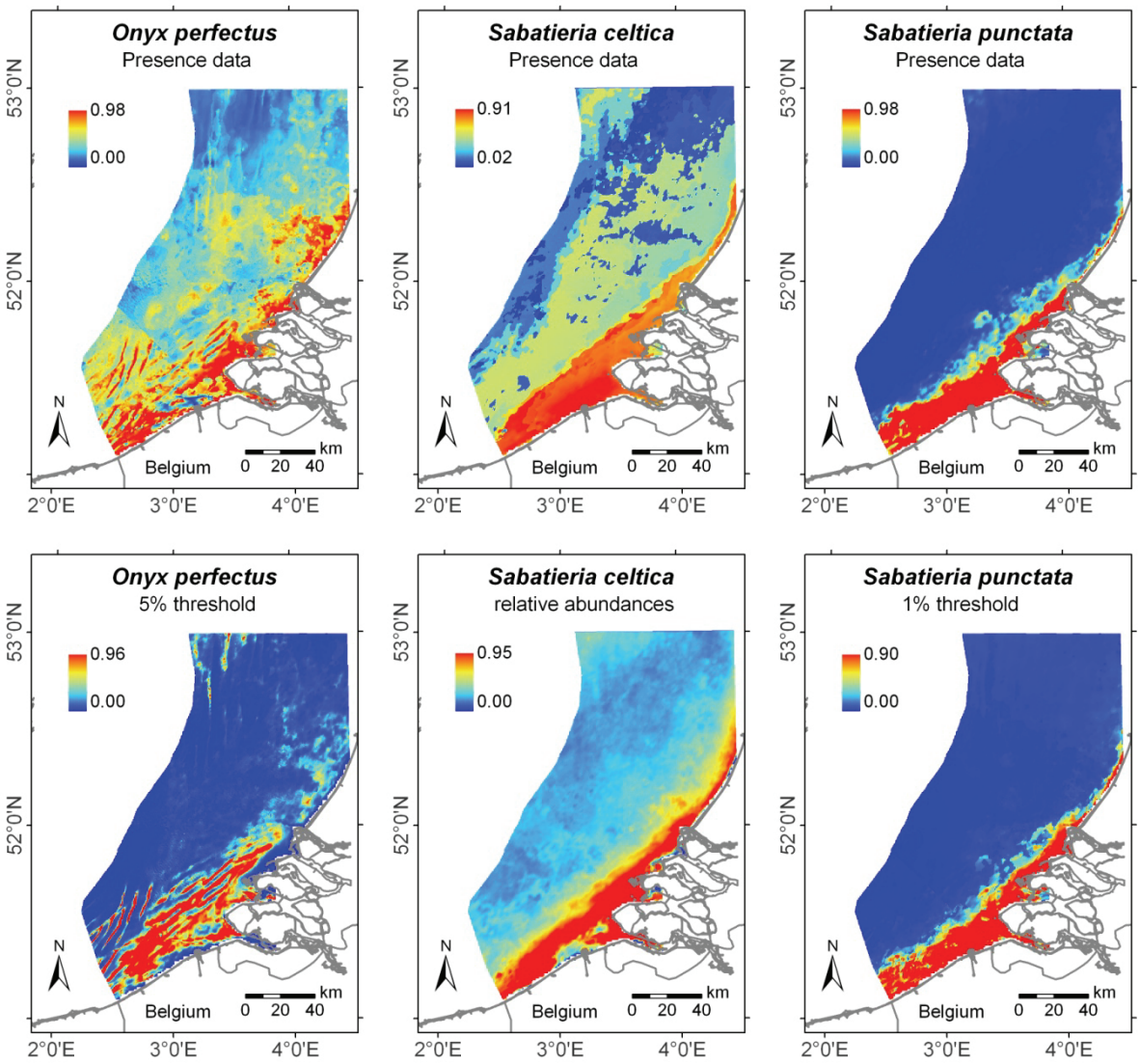


Fig. 6.6. Comparison of the two resulting models using all the presence data and the data incorporating relative abundances for Onyx perfectus*,* Sabatieria celtica *and* Sabatieria punctata*.*

According to literature *Daptonema tenuispiculum* and *Sabatieria punctata* often appear together (Vincx, 1989a; Vincx *et al.*, 1990; Vanaverbeke *et al.*, 2011). The models indeed indicate that the species have the same potential distribution. In addition, our models indicate that high relative abundances of *S. celtica* and *S. punctata* may coincide.

Five out of six nematode species appear in high relative abundances near the coast (Vincx, 1989a; Vanreusel, 1990). The environmental conditions enable species to reach high relative abundances in this region, and thus attributes to the effectiveness of this methodology for these species.

Niche concept

Generalist species tolerate a wide range of environmental conditions, and generally have a diverse diet and a good tolerance for disturbances. All this is important for defining the species' niche. Habitat suitability modelling estimates the species ecological niche. A species' fundamental niche represents a set of all conditions necessary for the survival of a species (Hutchinson, 1957). It is however assumed that every point within the niche has the same probability of persistence of the species, and all points outside the niche have zero probability of survival. This is clearly an oversimplification of reality. There will be optimal and suboptimal conditions in the niche (Hutchinson, 1957). By applying thresholds on the relative abundance, we filter out suboptimal conditions and try to define the fundamental niche for the survival of high relative abundances and dominance of a species. Species appearing across the whole region are generally hard to model since they show no real habitat preference. The relationship between the presence of the species and the environmental variables is therefore not always straightforward. In spite of this, some species may thrive or may be better competitors in certain restricted habitats, but not in others.

In this paper the data was converted to relative abundances, since the total density was not known for a considerable amount of data. In many other cases, absolute densities are known and most probably, this methodology can equally be applied to absolute densities.

CONCLUSIONS

In some cases knowing the potential density or relative abundance of a species in a region may be more important than knowing the suitability of the habitat. In this case it is reasonable to modify the data furnished to the habitat suitability modelling technique in such a way that habitats with high densities or relative abundances are preferentially predicted. The introduction of thresholds seems to be a reliable way to introduce this information into the model. Relating the model to existing knowledge of the species can help in identifying the most reliable model. Thus depending on the purpose of the model, we suggest different approaches in habitat suitability modelling: if the model concerns a rare species, knowing the potential niche may be the main focus of the research. If the species is

common and the species occurs in varying densities, applying thresholds may create opportunities to find the environments where the species can appear in high relative abundances.

ACKNOWLEDGEMENTS

This research is funded by the Fund for Scientific Research (FWO) of the Flemish government (FWO07/ASP/174). The authors wish to thank all the data providers! The environmental data was gathered from different institutes: ESA and MUMM/RBINS are acknowledged for providing and processing MERIS data (chlorophyll and TSM data, www.mumm.ac.be/BELCOLOUR), the Renard Centre of Marine Geology (RCMG, www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management for the oceanographic and sedimentological data. The study was conducted within the framework of the Ghent University BBSea Project (GOA 01600705) and the EU Network of Excellence MarBEF (GOCE-CT-2003-505446).

CHAPTER 7

APPLICATION TO MACROBENTHIC SPECIES

CHAPTER 7

APPLICATION TO MACROBENTHIC SPECIES

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

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

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

LANICE CONCHILEGA (PALLAS, 1766) AGGREGATIONS

Introduction

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

Lanice conchilega or sand mason is a polychaete, which builds linear tubes consisting of coarse sand grains cemented with mucus (Jones and Jago, 1993). The tube can reach a diameter of 5 mm and a length of 65 cm (Ziegelmeier, 1952). The tube is located mainly in

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

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

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

Material and methods

Research area

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

Lanice conchilega data

The *L. conchilega* data was retrieved from the MacroDat database (Degraer *et al.*, 2003a). This database is a compilation of macrobenthos data of the BPNS and beaches from 1163 stations (Fig. 7.1) taken within the time frame 1971-2008. From this database, the stations and densities with *L. conchilega* records were extracted, resulting in a dataset consisting of 231 stations where densities between 3 and 13 000 individuals per square meter were

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

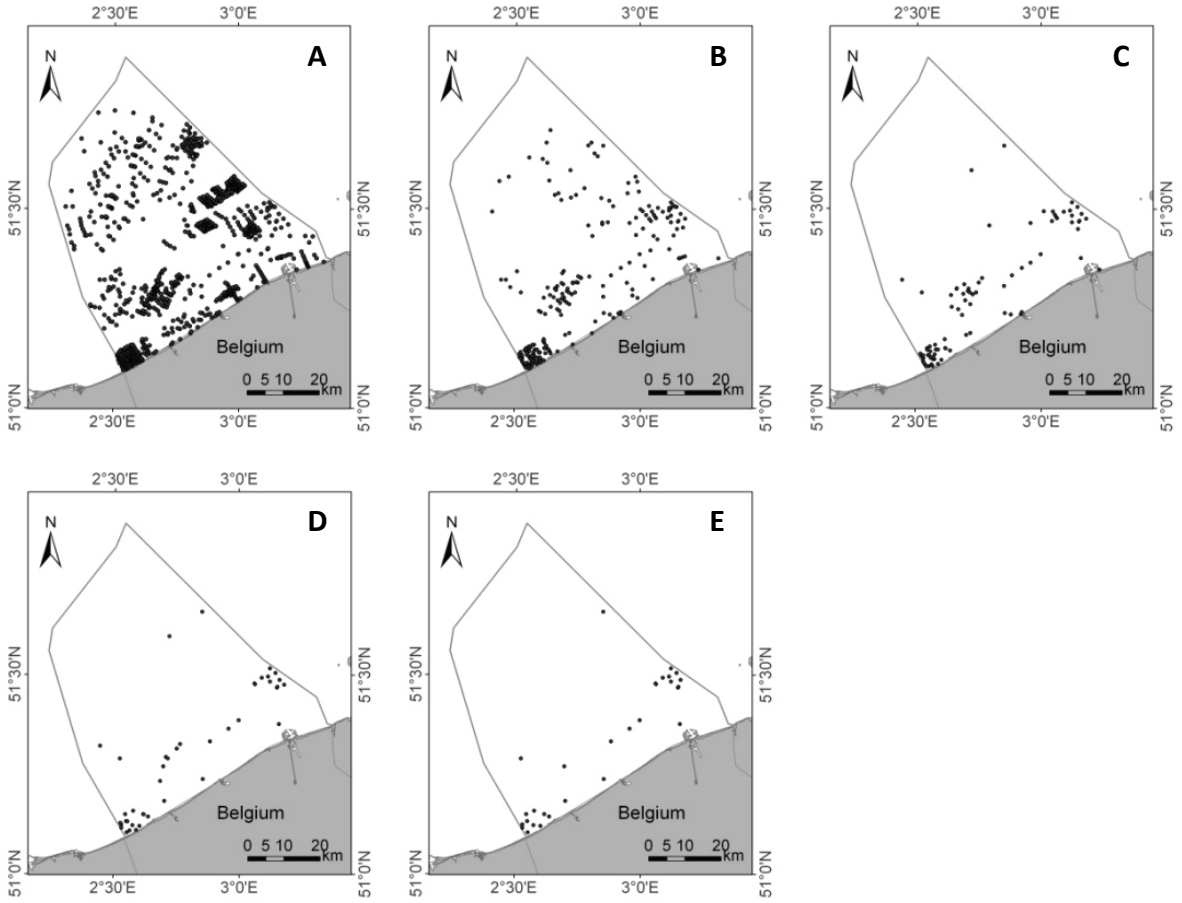


Fig. 7.1. Sampling stations (•) from the MacroDat database (A), stations where L. conchilega *has been observed (B),* L. conchilega *stations with at least 100 ind.m^{-2} (C),* L. conchilega *stations with at least 500 ind.m^{-2} (D) and* L. conchilega *stations with at least 1000 ind.m^{-2} (E).*

Environmental data

Fifteen environmental variables related to granulometry, the topography of the area and current properties were selected (Table 7.1). Chlorophyll *a* and total suspended matter were excluded from the analysis because these maps have no data values near the coastline, which would therefore result in maps without any prediction near the coast. Only those

current properties concerning bottom currents and bottom shear stress were considered, since these variables may be of direct influence to *L. conchilega*.

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

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

Modelling procedure

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

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

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

Results and discussion

Test for preferential sampling

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

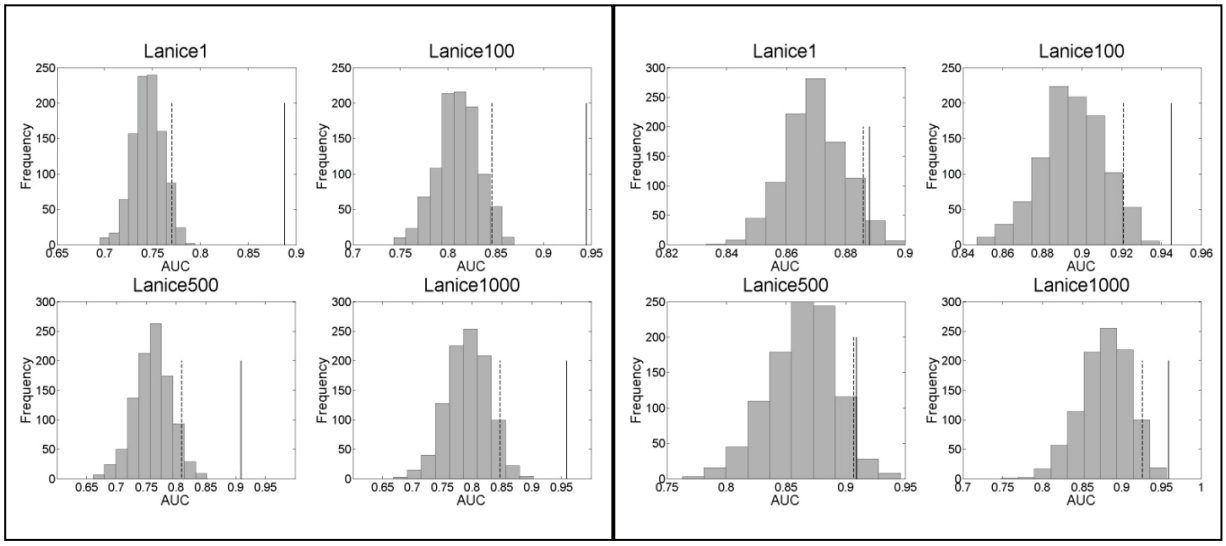


Fig. 7.2. Histograms of the randomisations of the four L. conchilega *densities: randomisations based on the total sampling area (left) and randomisations based on the actual sampling stations (right). The 95% quantile value (dotted line) and the AUC of the* Lanice *model (full line) are indicated.*

Contribution of the environmental variables

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

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

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

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

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

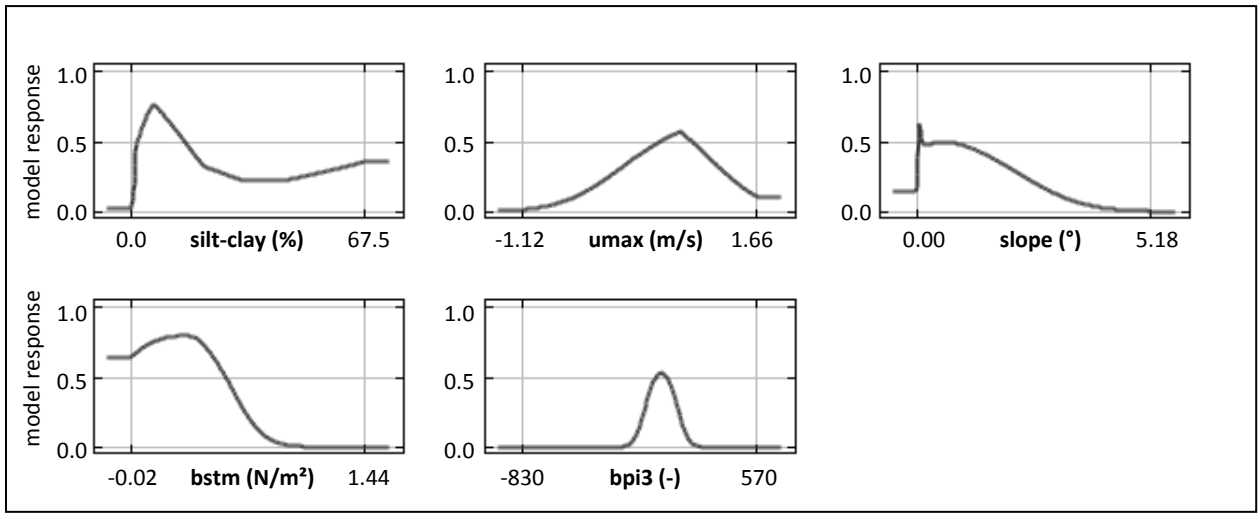


Fig. 7.3. Relation between the environmental variable and the logistic output of the Lanice conchilega models with densities of at least 500 individuals.m^{-2}. Each curve represents a model created using only the corresponding variable. See Table 7.1 for explanation of the abbreviations. The minimum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

Mapping of HSM of Lanice conchilega

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

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

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

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

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

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

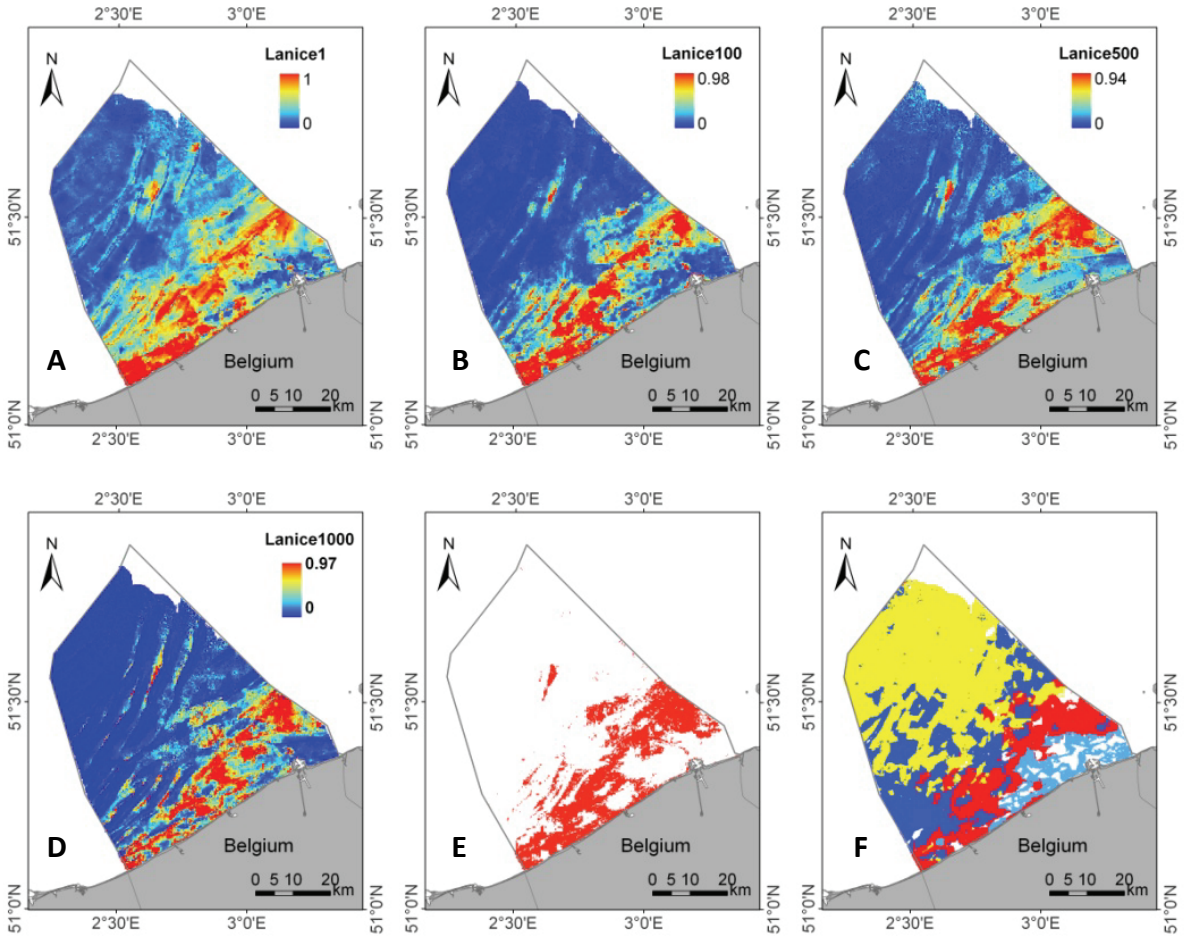


Fig. 7.4. Habitat suitability Models for different densities of Lanice conchilega*: presence (A), 100 ind.m^{-2} (B), 500 ind.m^{-2} (C), 1000 ind.m^{-2} (D). Model of Lanice500 with 'maximum training sensitivity plus specificity' threshold (E) and delineation of the four main macrobenthic communities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

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

The *Abra alba* community has previously been described as being of exceptional ecological importance because of its high macrobenthic abundance (6432 ind.m^{-2}) and diversity (30 species per sample of 0.12 m²). This community tends to hold some unique species for the BPNS (Van Hoey *et al.*, 2004) and the bivalves present in the community may serve as food source to sea ducks (Degraer *et al.*, 1999). Moreover, *L. conchilega* has the capacity to double the biodiversity of the *A. alba* community (Van Hoey, 2006). Table 7.3 and Fig. 7.4 clearly indicate that aggregations of *Lanice* pattern overlap with the distribution of the *Abra alba* community (Fig. 7.4 and Table 7.3). For the first threshold, about 51% of the area where the Lanice500 model has a high probability is situated within the *Abra alba* region, but on the other hand the *Abra alba* community lies for about 91 % in the Lanice500 region (Table 7.3). When the threshold increases, a smaller area, but with a higher probability of Lanice500 is predicted. Still 78% of the Lanice500 model is found within the *Abra alba* community and 68% of the Lanice500 surface lies within the distribution area of the *Abra alba* community. The clear overlap between the two areas can thus only promote the relevance of *Lanice conchilega* for nature conservation.

ENSIS DIRECTUS (CONRAD 1843)

Introduction

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

The species tends to occur in high densities (i.e. bivalve banks) and densities of 1000-2000 ind.m^{-2} are not uncommon (Armonies and Reise, 1999; Tulp *et al.*, 2010). These *Ensis* banks have a patchy distribution, but patches are not permanent. The European *E. directus*

populations show conspicuous events of mass mortality, mainly in late winter or early spring (Armonies and Reise, 1999). *Ensis directus* lives deep in the sediment and when in danger, it can retract fast into the sediment with its powerful foot down to a depth of 50 cm (Tulp *et al.*, 2010). It prefers muddy, fine sand with small amounts of silt (Beukema and Dekker, 1995; Kennish *et al.*, 2004) and is found in the intertidal and subtidal zones (Mühlenhardt-Siegel *et al.*, 1983; Swennen *et al.*, 1985). These sublittoral muddy fine sand sediments are also known to be the habitat of the diverse *Abra alba* community (Van Hoey *et al.*, 2004; Degraer *et al.*, 2008). Therefore, knowing the potential habitat of *E. directus* may indicate if a potential effect of *E. directus* on the *Abra alba* community is to be expected.

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

Materials and methods

Research area

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

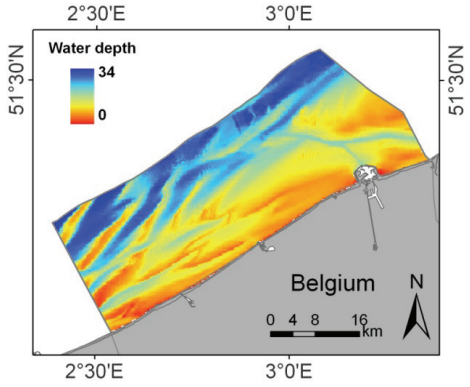


Fig. 7.5. Bathymetrical data of the 12-miles zone of the BPNS (Source: RCMG).

Ensis directus *data*

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

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

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

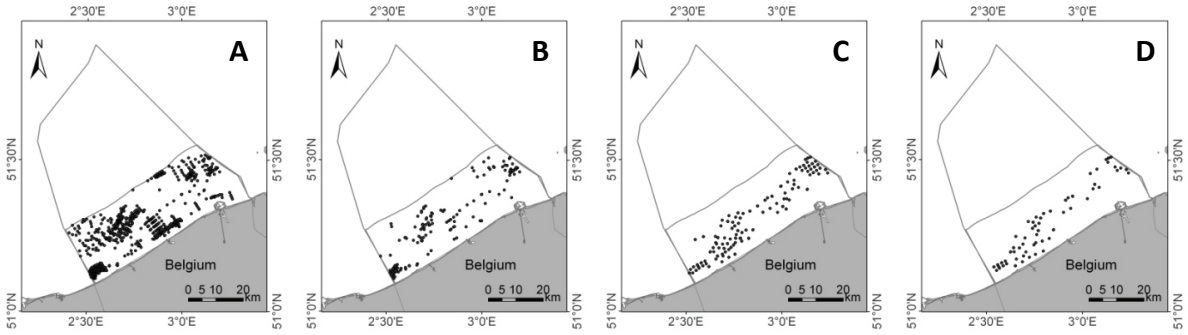


Fig. 7.6. Sampling stations (•) from the MacroDat database (A) within the 12-miles zone of the BPNS and stations where Ensis has been observed (B). Sampling stations from the MUMM database (C) and stations from the MUMM database where 1-year old specimens were found (D).

Environmental data

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

Modelling procedure

Habitat suitability modelling (HSM)

Ensis directus is an invasive species, which means that the species has not yet reached equilibrium in the area, and the absence of the species does thus not necessarily imply that

the habitat is unsuitable for the species. Therefore, we used a presence-only modelling algorithm, namely Maxent.

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

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

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

The presence of preferential sampling was checked by applying a randomisation test: 499 random models were created with locations sampled from the actually sampled stations. When preferential sampling is not evident from the data, the model is further refined by

applying a feature and variable selection. Maxent applies default different features: linear, quadratic, product, threshold and hinge features. Complex features may enhance overfitting, therefore a feature selection may result in a more realistic relation between the environmental variables and the output of the HSM. The variable and feature selection was done by a five-fold cross-validation.

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

Density map

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

Results and discussion

Habitat suitability modelling

Test for preferential sampling

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

sampling. Therefore, we will only focus on the HSM model derived from the data from the MacroDat database.

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

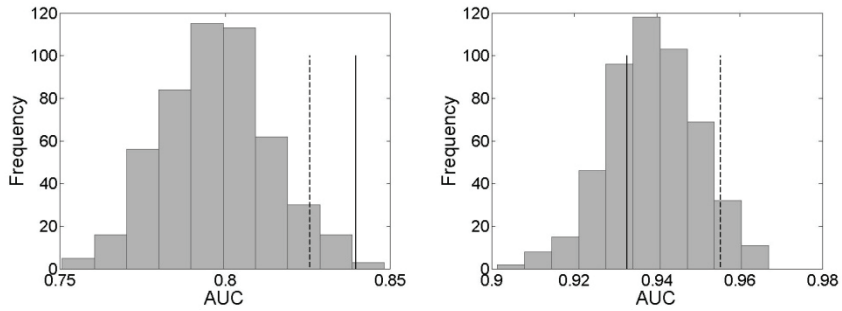


Fig. 7.7. Histograms of the randomisations of the MacroDat (left) and the MUMM database (right). The randomisations are based on the actual sampling stations. The 95% quantile value (dotted line) and the AUC of the Ensis *model (full line) are indicated.*

Contribution of the environmental variables

Both models, Ensis1km and Ensis5km, select bottom shear stress, water depth and sand fraction as the most important factors. For the Ensis1km model the minimum and maximum chlorophyll content are selected as well. Previous research pointed out that the American Jack knife clam is an opportunistic species, with little requirements regarding its environment. It prefers wave- and current-swept clean sands (Beukema and Dekker, 1995) with small amounts of silt (Kennish *et al.*, 2004), but it can also be found in muddy or coarse sediments (Armonies and Reise, 1999) and can thus be independent of sediment characteristics (Dauvin *et al.*, 2007). *E. directus* however has a limited tolerance to hypoxia and will thus avoid reduced sediments (Schiedek and Zebe, 1987). The positive relation between the sand fraction and *E. directus* is found in both HSM models (Fig. 7.8). Silt-clay was not selected as a variable. The preference for moving sands (Kenchington *et al.*, 1998)

and strong currents are not advocated by the model; the model indicates that the maximum bottom shear stress should be limited to about 4 $N.m^{-2}$ while a shear stress above 5 $N.m^{-2}$ corresponds to the threshold of sand transport (Mangelsdorf *et al.*, 1990). Likely areas for colonisation are subtidal and intertidal areas (ICES, 2005; Ovcharenko and Gollasch, 2009). The models do not favour near shore areas, which is probably because there is no environmental data available in the intertidal area. Intermediate water depths between 12 and 23 m are optimal according to the models. This is partially in agreement with observations by Armonies and Reise (1999), who found that sublittoral populations prefer water depths of 18 m and more. The influence of chlorophyll *a* on the Ensis1km seems contradictory: the species is favoured by low minimum chlorophyll *a* values and high maximum chlorophyll *a* values. This could mean that the species is preferably found in areas with annually strongly fluctuating chlorophyll *a* values, however, there are no literature sources supporting this.

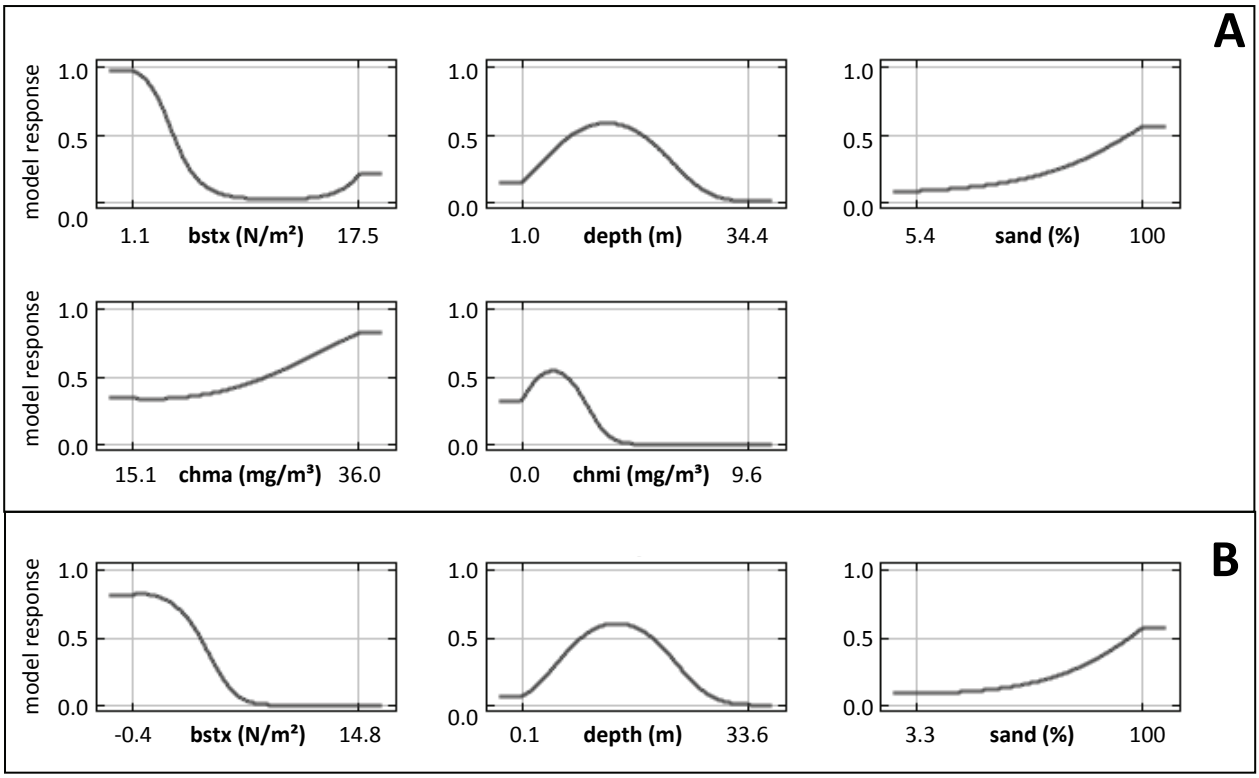


Fig. 7.8. Relation between the environmental variables and logistic output of the Ensis1km (A) and Ensis5km model (B). Each of the curves represents a model created using only the corresponding variable. See Table 7.4 for explanation of the abbreviations. The minimum and maximum values of the environmental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

The selected variables do not concord very well with findings in literature. This can be explained by two reasons: First, a lot of the data in literature is collected in intertidal areas. In the present study, information on the intertidal is lacking. Secondly, all observations were

taken into account, but *E. directus* is an opportunistic species which may occur in many different habitats. Its presence is therefore only slightly regulated by the environmental variables. However, it is possible that high densities of *E. directus* can only thrive in specific conditions. Therefore, by analogy to Chapter 6 and the HSM developed for *Lanice conchilega*, creating HSM for the species based on density thresholds could reveal the environmental variables which relate to high densities of the species.

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

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

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

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

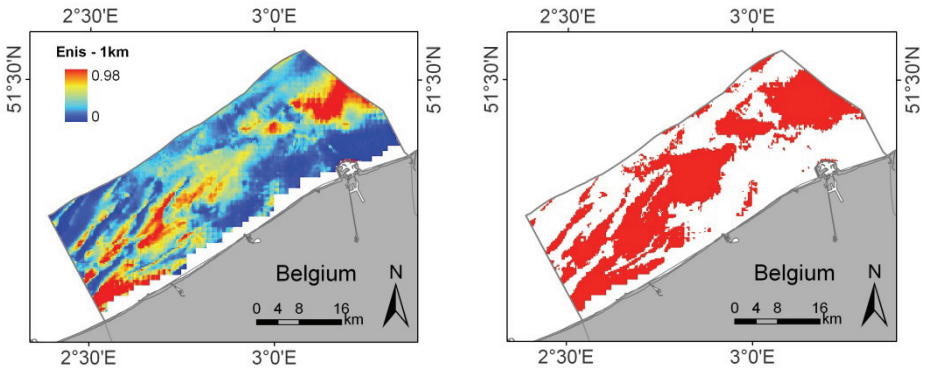


Fig. 7.9. Resulting maps of the Ensis1km habitat suitability model and the thresholded model ('10 percentile training presence' threshold) .

Density map

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

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

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

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

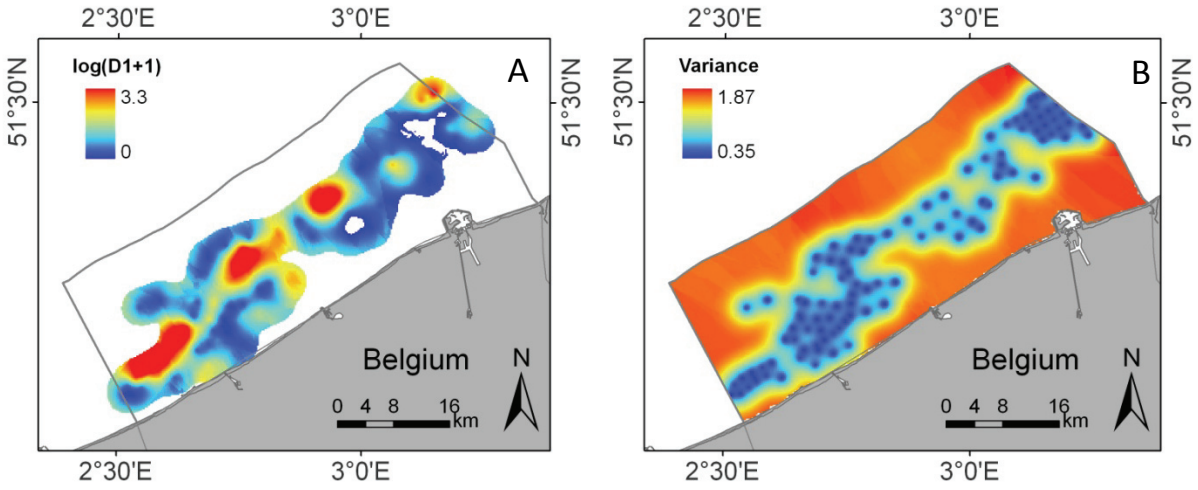


Fig. 7.10. Map of log(D1+1), with D1 the density of the 1-year old cohort (ind.m^{-2}) (A). This map is restricted to a variance smaller than 1.5. On the right side the variance map is shown (B).

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

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

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

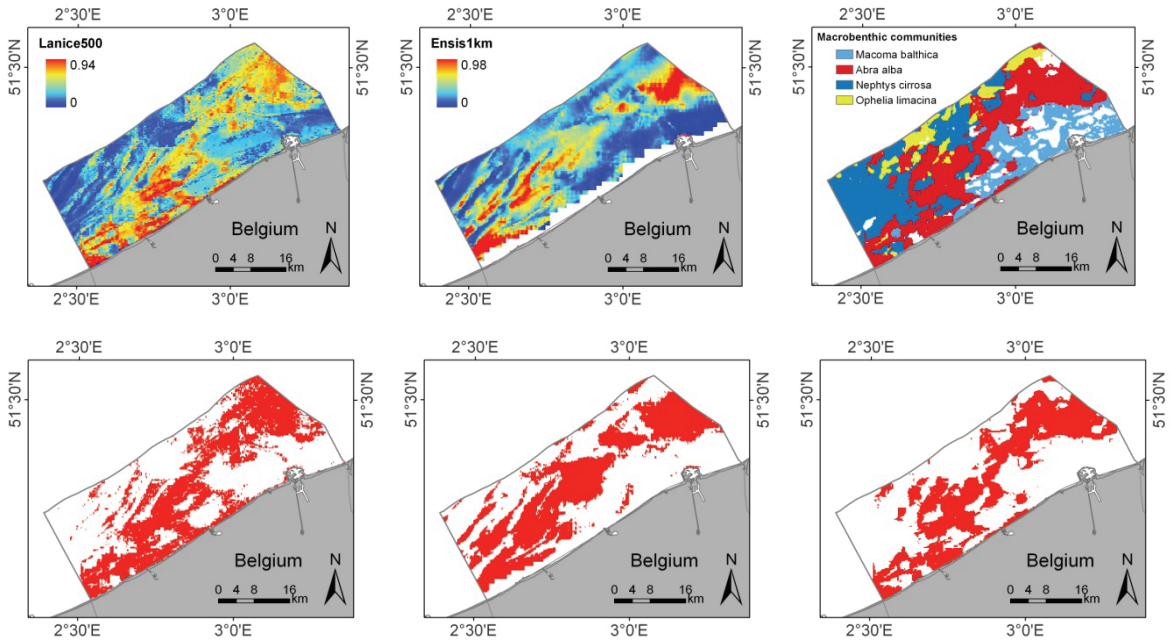


Fig. 7.11. Habitat suitability models of Lanice conchilega *(500 ind.m^{-2}),* Ensis directus *(1 km) and the four macrobenthic communities. Below, the reduction of these maps to binary maps representing* Lanice conchilega *(500 ind.m^{-2}, logistic threshold 0.358),* Ensis directus *(1 km, logistic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

CONCLUSIONS

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

ACKNOWLEDGMENTS

This research is funded by the Fund for Scientific Research(FWO) of the Flemish Government (FWO07/ASP/174). The authors wish to thank all the data providers! The environmental data was gathered from different institutes: ESA and MUMM/RBINS are acknowledged for

providing and processing MERIS data (chlorophyll *a* and TSM data, http://www.mumm.ac.be/BELCOLOUR), the Renard Centre of Marine Geology (RCMG, http://www.rcmg.ugent.be) of Ghent University and the Hydrographic Service of the Royal Netherlands Navy and the Directorate-General of Public Works and Water Management of the Dutch Ministry of Transport, Public Works and Water Management for the oceanographic and sedimentological data. The *Lanice conchilega* modelling was conducted within the Westbanks Project which is supported by the Belgian Science Policy (BELSPO): SSD Science for sustainable Development General Coordination and for the project 'Studie betreffende het opstellen van een lijst van potentiële Habitatrichtlijngebieden in het Belgische deel van de Noordzee' under the authority of Federale Overheidsdienst Volksgezondheid, Veiligheid van de Voedselketen en Leefmilieu, Directoraat-generaal Leefmilieu. The *Ensis directu*s modelling was performed as part of the Ensis project, financed by the Federal Science Policy.

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 ($\pm$ 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.

CITED LITERATURE

Abebe E., Andrássy I., Traunspurger, W. (eds.), 2006. Freshwater nematodes: ecology and taxonomy. CABI Publishing, Wallingford, UK, 752 pp.

Abrams, P.A., 1995. Monotonic or unimodal diversity-productivity gradients: what does competition theory predict? Ecology 76, 2019-2027.

Abugov, R. 1982. Species diversity and the phasing of disturbance. Ecology 63, 289-293.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC19, 716-723.

Alongi, D.M., Tietjen, J.H., 1980. Population growth and trophic interactions among free-living marine nematodes. Marine Benthic Dynamics, 151-166.

Amari, S., Murata, N., Muller, K.-R., Finke, M., Yang, H.H., 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactions on Neural Networks 8(5), 985-996.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33, 1677-1688.

Arlt, G., 1973. Vertical and horizontal microdistribution of the meiofauna in Greifswalder Bodden. Oikos, 105-111.

Armenteros, M., Perez-Garcia, J.A., Ruiz-Abierno, A., Diaz-Asencio, L., Helguera, Y., Vincx, M., Decraemer, W., 2010. Effects of organic enrichment on nematode assemblages in a microcosm experiment. Marine Environmental Research 70, 374-382.

Armonies, W., Reise, K., 1999. On the population development of the introduced razor clam *Ensis americanus* near the island of Sylt (North Sea). Helgolander Meeresuntersuchungen 52, 291-300.

Armonies, W., Reise, K., 2000. Faunal diversity across a sandy shore. Marine Ecology Progress Series 196, 49-57.

Austen, M.C., Widdicombe, S., 2006. Comparison of the response of meio- and macrobenthos to disturbance and organic enrichment. Journal of Experimental Marine Biology and Ecology 330, 96-104.

Austen, M.C., Widdicombe, S., Villano-Pitacco, N., 1998. Effects of biological disturbance on diversity and structure of meiobenthic nematode communities. Marine Ecology Progress Series 174, 233-246.

Baeyens, W., Chou, L., Frankignoulle, M., Laane, R., 2007. Project EV/20. Biogeochemical cycling of carbon, nitrogen and phosphorus in the North Sea. Final report. Published by The Belgian Science Policy, Belgium, 75 pp., available online: http://ns.belspo.be/belspo/home/publ/pub_ostc/EV/rappEV20_en.pdf

Bazzaz, F.A., 1975. Plant species diversity in old-field successional ecosystems in Southern Illinois. Ecology 56, 485-488.

Beale, M.H., Hagan, M.T., Demuth, H.B., 2010. Neural Network Toolbox. User's Guide by The MathWorks, Inc. Revised for Version 7.0., Natick, MA, USA, 849 pp., online only: http://www.mathworks.com/help/pdf_doc/nnet/nnet.pdf

Bell, G., 2000. The distribution of abundance in neutral communities. The American Naturalist 155, 606-617.

Bell, G., Lechowicz, M.J., Waterway, M.J., 2006. The comparative evidence relating to functional and neutral interpretations of biological communities. Ecology 87, 1378-1386.

Benito, B.M., Martinez-Ortega, M.M., Munoz, L.M., Lorite, J., Penas, J., 2009. Assessing extinction-risk of endangered plants using species distribution models: a case study of habitat depletion caused by the spread of greenhouses. Biodiversity and Conservation 18, 2509-2520.

Berry, P.M., Dawson, T.P., Harrison, P.A., Pearson, R.G., 2002. Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. Global Ecology and Biogeography 11, 453-462.

Beukema, J.J., 1974. The efficiency of the Van Veen grab compared with the Reineck box sampler ICES Journal of Marine Science: Journal du Conseil International pour l'Exploration de la Mer 35, 319-327.

Beukema, J.J., Dekker, R., 1995. Dynamics and growth of a recent invader into European coastal waters: The American razor clam, *Ensis directus*. Journal of the Marine Biological Association of the United Kingdom 75, 351-362.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK, 482 pp.

Blanchard, G.F., 1990. Overlapping microscale dispersion patterns of meiofauna and microphytobenthos. Marine Ecology Progress Series 68, 101-111.

Blome, D., Schleier, U., von Bernem, K.H., 1999. Analysis of the small-scale spatial patterns of free-living marine nematodes from tidal flats in the East Frisian Wadden Sea. Marine Biology 133, 717-726.

Bongers, T., Ferris, H., 1999. Nematode community structure as a bioindicator in environmental monitoring. Trends in Ecology & Evolution 14, 224-228.

Bonn, A., Schröder, B., 2001. Habitat models and their transfer for single and multi species groups: a case study of carabids in an alluvial forest. Ecography 24, 483-496.

Boyd, S.E., Rees, H.L., Richardson, C.A., 2000. Nematodes as sensitive indicators of change at dredged material disposal sites. Estuarine, Coastal and Shelf Science 51, 805-819.

Braeckman, U., Provoost, P., Gribsholt, B., Van Gansbeke, D., Middelburg, J.J., Soetaert, K., Vincx, M., Vanaverbeke, J., 2010. Role of macrofauna functional traits and density in biogeochemical fluxes and bioturbation. Marine Ecology Progress Series 399, 173-186.

Braeckman, U., Van Colen, C., Soetaert, K., Vincx, M., Vanaverbeke, J., 2011. Contrasting macrobenthic activities differentially affect nematode density and diversity in a shallow subtidal marine sediment. Marine Ecology Progress Series 422, 179-191.

Brown, K.A., Spector, S., Wu, W., 2008. Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. Journal of Applied Ecology 45, 1639-1648.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Morcillo, M.H., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vie, J.C., Watson, R., 2010. Global biodiversity: indicators of recent declines. Science 328, 1164-1168.

Caldeira, K., Wickett, M.E., 2003. Anthropogenic carbon and ocean pH. Nature 425, 365-365.

Callaway, R., 2006. Tube worms promote community change. Marine Ecology Progress Series 308, 49-60.

Callaway, R., Desroy, N., Dubois, S.F., Fournier, J., Frost, M., Godet, L., Hendrick, V.J., Rabaut, M., 2010. Ephemeral Bio-engineers or Reef-building Polychaetes: How Stable are Aggregations of the Tube Worm *Lanice conchilega* (Pallas, 1766)? Integrative and Comparative Biology 50, 237-250.

Cardinale, B.J., Bennett, D.M., Nelson, C.E., Gross, K., 2009. Does productivity drive diversity or vice versa? A test of the multivariate productivity-diversity hypothesis in streams. Ecology 90, 1227-1241.

Carnaval, A.C., Moritz, C., 2008. Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. Journal of Biogeography 35, 1187-1201.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2, 667-680.

Carranza, A., Arim, M., Scarabino, F., Defeo, O., 2010. Coexistence patterns of benthic gastropods on the Uruguayan shelf. Oikos 119, 1312-1318.

Carroll, R.J., Ruppert, D., 1988. Transformation and weighing in regression. Chapman and Hall, New York, USA, 264 pp.

CBD, 2006. Global Biodiversity Outlook 2. Secretariat of the Convention on Biological Diversity Montreal, Canada, 81 pp., available online: www.biodiv.org/GBO2

Chandler, G.T., Fleeger, J.W., 1987. Facilitative and inhibitory interactions among estuarine meiobenthic harpacticoid copepods. Ecology 68, 1906-1919.

Chao, A., 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11, 265-270.

Chao, A., 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783-791.

Chao, A., Shen, T.-J., 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10, 429-443.

Chase, J.M., Leibold, M.A., 2002. Spatial scale dictates the productivity-biodiversity relationship. Nature 416, 427-430.

Cheng, B., Titterington, D.M., 1994. Neural Networks: A Review from a statistical Perspective. Statistical Science 9, 2-54.

Chiarucci, A., Enright, N.J., Perry, G.L.W., Miller, B.P., Lamont, B.B., 2003. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions 9, 283-295.

Clarke, K.R., Warwick, R.M., 1998. A taxonomic distinctness index and its statistical properties. Journal of Applied Ecology 35 (4), 523-531.

Clarke, K.R., Warwick, R.M., 2001. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation, 2nd edition, Primer-E Ltd., Plymouth, UK, 172 pp.

Collins, S.L., Glenn, S.M., 1997. Intermediate disturbance and its relationship to within- and between-patch dynamics. New Zealand Journal of Ecology 21, 103-110.

Connell, J.H., 1978. Diversity in tropical rain forest and coral reefs. Science 199, 1302-1310.

Connell, J.H., 1980. Diversity and the coevolution of competitors, or the ghost of competition past. Oikos 35, 131-138.

Connell, J.H., Orias, E., 1964. The ecological regulation of species diversity. The American Naturalist 98, 399-414.

Connor, E.F., Simberloff, D., 1979. The assembly of species communities: Chance or competition? Ecology 60, 1132-1140.

Cordellier, M., Pfenninger, M., 2009. Inferring the past to predict the future: climate modelling predictions and phylogeography for the freshwater gastropod *Radix balthica* (Pulmonata Basommatophora). Molecular Ecology 18, 534-544.

Cressie, N.A.C., 1993. Statistics for spatial data. John Wiley & Sons, New York, USA, 900 pp.

Cunningham, H.R., Rissler, L.J., Apodaca, J.J., 2009. Competition at the range boundary in the slimy salamander: using reciprocal transplants for studies on the role of biotic interactions in spatial distributions. Journal of Animal Ecology 78, 52-62.

Danovaro, R., Gambi, C., Dell'Anno, A., Corinaidesi, C., Fraschetti, S., Vanreusel, A., Vincx, M., Gooday, A.J., 2008. Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. Current Biology 18, 1-8.

Darwin, C.R., 1876. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 6th edition with additions and corrections, John Murray, London, UK.

Dauvin, J.C., Ruellet, T., Thiebaut, E., Gentil, F., Desroy, N., Janson, A.L., Duhamel, S., Jourde, J., Simon, S., 2007. The presence of *Melinna palmata* (Annelida : Polychaeta) and *Ensis directus* (Mollusca : Bivalvia) related to sedimentary changes in the Bay of Seine (English Channel, France). Cahiers De Biologie Marine 48, 391-401.

Dayton, P.K., 1971. Competition, Disturbance, and Community Organization: The Provision and Subsequent. Utilization of Space in a Rocky Intertidal Community. Ecological Monographs 41, 351-389.

Dayton, P.K., Hessler, R.R., 1972. Role of biological disturbance in maintaining diversity in the deep sea. Deep-Sea Research 19, 199-208.

De Cauwer, V., Ruddick, K., Park, Y., Nechad, B., Kyramarios, M., 2004. Optical remote sensing in support of eutrophication monitoring in the southern North Sea. EARSeL eProceedings 3, 208-221.

De Ley, P.A., 2006. Quick tour of nematode diversity and the backbone of nematode phylogeny. WormBook (ed.) The *C. elegans* Research Community, p. 1-8, available online: http://www.wormbook.org.

De Mesel, I., Derycke, S., Swings, J., Vincx, M., Moens, T., 2006a. Role of nematodes in decomposition processes: does within-trophic group diversity matter? Marine Ecology Progress Series 321, 157-166.

De Mesel, I., Lee, H.J., Vanhove, S., Vincx, M., Vanreusel, A., 2006b. Species diversity and distribution within the deep-sea nematode genus *Acantholaimus* on the continental shelf and slope in Antarctica. Polar Biology 29, 860-871.

Debenham, N.J., Lambshead, P.J.D., Ferrero, T.J., Smith, C.R., 2004. The impact of whale falls on nematode abundance in the deep sea. Deep-Sea Research Part I-Oceanographic Research Papers 51, 701-706.

Dedecker, A.P., Goethals, P.L.M., Gabriels,W., De Pauw, N., 2004. Optimization of artificial neural network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecological Modelling 174, 161-173.

Degraer, S., Vincx, M., Meire, P., Offringa, H., 1999. The macrozoobenthos of an important wintering area of the common scoter (*Melanitta nigra*). Journal of the Marine Biological Association of the United Kingdom 79, 243-251.

Degraer, S., Van Hoey, G., Willems, W., Speybroeck, J., Vincx, M., 2003a. MacroDat Belgium. Macrobenthic data from the Belgian part of the North Sea from 1976 onwards. Ghent University, Biology Department, Marine Biology Section, Belgium.

Degraer, S., Van Lancker, V., Moerkerke, G., Van Hoey, G., Vanstaen, K. Vincx, M., Henriet, J.-P., 2003b. Evaluation of the ecological value of the foreshore: habitat-model and macrobenthic side-scan sonar interpretation: extension along the Belgian Coastal Zone. Final report. Ministry of the Flemish Community, Environment and Infrastructure Department, Waterways and Marine Affairs Administration, Coastal Waterways, Belgium, 63 pp.

Degraer, S., Verfaillie, E., Willems, W., Adriaens, E., Vincx, M., Van Lancker, V., 2008. Habitat suitability modelling as a mapping tool for macrobenthic communities: An example from the Belgian part of the North Sea. Continental Shelf Research 28, 369-379.

Degraer, S., Braeckman, U., Haelters, J., Hostens, K., Jacques, T.G., Kerckhof, F., Merckx, B., Rabaut, M., Stienen, E.W.M., Van Hoey, G., Van Lancker, V.R.M., Vincx, M., 2009. Studie betreffende het opstellen van een lijst met potentiële Habitatrichtlijngebieden in het Belgische deel van de Noordzee. Eindrapport. Federale Overheidsdienst Volksgezondheid, Veiligheid van de Voedselketen en Leefmilieu: Brussels, Belgium, 93 pp.

Demuth, H.B., Beale, M.H., 1998. Neural Network Toolbox for Use with MATLAB. User's Guide. Version 3.0., Natick, MA, USA, 849 pp.

Derous, S., 2008. Marine biological valuation as a decision support tool for marine management. PhD Thesis, Ghent University, Faculty of Sciences, Marine Biology Section, Ghent, Belgium, 298 pp.

Derous, S., Agardy, M.T., Hillewaert, H., Hostens, K., Jamieson, G., Lieberknecht, L., Mees, J., Moulaert, I., Olenin, S., Paelinckx, D., Rabaut, M., Rachor, E., Roff, J.C., Stienen, E., van der Wal, J.T., Van Lancker, V.R.M., Verfaillie, E., Vincx, M., Weslawski, J.M., Degraer, S., Agardy, T., Paelinckx, D., Roff, J., Stienen, E.W.M., 2007. A concept for biological valuation in the marine environment. Oceanologia 49, 99-128.

Deutsch, A., 1978. Gut ultrastructure and digestive physiology of two marine nematodes, *Chromadorina germanica* (Bütschli, 1874) and *Diplolaimella* sp. Biological Bulletin 155, 317-335.

Deutsch, C.V., Journel, A.G., 1992. GSLIB: geostatistical software library and user's guide. Oxford University Press, New York, USA, 369 pp.

Diamond, J.M., 1975. Assembly of Species Communities. In: Cody, M.L., Diamond, J.M. (eds.),Ecology and Evolution of Communities, Belknap, Harvard, UK, p. 342-444.

Diamond, J.M., Gilpin, M.E., 1982. Examination of the 'null' model of Connor and Simberloff for species co-occurrences on islands. Oecologia 52, 64-74.

Diniz-Filho, J.A.F., Bini, L.M., Hawkins, B.A., 2003. Spatial autocorrelation and red herrings in geographical ecology. Global Ecology & Biogeography 12, 53-64.

Dobzhansky, T., 1950. Mendelian populations and their evolution. The American Naturalist 84, 401-418.

Dormann, C.F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography 16, 129-138.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609-628.

dos Santos, G.A.P., Derycke, S., Genevois, V.G.F., Coelho, L., Correia, M.T.S., Moens, T., 2009. Interactions among bacterial-feeding nematode species at different levels of food availability. Marine Biology 156, 629-640.

Drew, G.A., 1907. The habits and movements of the razor-shell clam, *Ensis directus*, Conrad. Biological Bulletin 12, 127-140.

Druon, J.N., Schrimpf, W., Dobricic, S., Stips, A., 2004. Comparative assessment of large-scale marine eutrophication: North Sea area and Adriatic Sea as case studies. Marine Ecology Progress Series 272, 1-23.

Dudík, M., Phillips, S.J., Schapire, R.E., 2004. Performance guarantees for regularized maximum entropy density estimation. In: Proceedings of the 17th Annual Conference on Computational Learning Theory, ACM Press, New York, USA, p. 655-662.

EC (European Commission), 2006. Annexes to the communication from the Commission - Halting the loss of biodiversity by 2010 - and beyond - Sustaining ecosystem services for human well-being. EC, Brussels, Belgium, 14 pp., online available: http://ec.europa.eu/environment/nature/biodiversity/comm2006/pdf/sec_2006_621.pdf

EC (European Commission), 2009. Green Paper: Reform of the Common Fisheries Policy. EC, Com 163, Brussels, Belgium, 27 pp., online available: http://ec.europa.eu/fisheries/reform.

Echarri, F., Tambussi, C., Hospitaleche, C.A., 2009. Predicting the distribution of the crested tinamous, Eudromia spp. (Aves, Tinamiformes). Journal of Ornithology 150, 75-84.

Eckman, J.E. 1983. Hydrodynamic processes affecting benthic recruitment. Limnology and Oceanography, 28, 241-257.

Edwards, M., Richardson, A.J., 2004. Impact of climate change on marine pelagic phenology and trophic mismatch. Nature 430, 881-884.

Eleveld, M.A., Pasterkamp, R., van der Woerd, H.J., 2004. A survey of total suspended matter in the southern North Sea based on the 2001 SeaWiFS data. EARSeL eProceedings 3, 166-178.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43-57.

Etter, R.J., Mullineaux, L., 2001. Deep-Sea communities. In: Bertness, M.D., Gaines, S., Hay, M., (eds.) Marine Community Ecology, p. 367-393. Sunderland, Sinauer Associates, Sunderland, MA, USA, 550 pp.

Evangelista, P.H., Kumar, S., Stohlgren, T.J., Jarnevich, C.S., Crall, A.W., Norman, J.B., Barnett, D.T., 2008. Modelling invasion for a habitat generalist and a specialist plant species. Diversity and Distributions 14, 808-817.

FAO, 2010. The State of World Fisheries and Aquaculture. FAO Fisheries and Aquaculture Department. Food and Agriculture Organization of the United Nations, Rome, Italy, 207 pp., available online: http://www.fao.org/docrep/013/i1820e/i1820e.pdf

Fayle, T.M., Manica, A., 2010. Reducing over-reporting of deterministic co-occurrence patterns in biotic communities. Ecological Modelling 221, 2237-2242.

Fayle, T.M., Manica, A., 2011. Bias in null model analyses of species co-occurrence: A response to Gotelli and Ulrich. Ecological Modelling 222, 1340–1341.

Feilhauer, H., Schmidtlein, S., 2009. Mapping continuous fields of forest alpha and beta diversity. Applied Vegetation Science 12, 429-439.

Fernandes, F.A.N., Lona, L.M.F., 2005. Neural network applications in polymerization processes. Brazilian Journal of Chemical Engineering 22, 401-418.

Ferrier, S., Watson, G., 1996. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. NSW National Parks and Wildlife Service, Canberra, Australia, 184 pp., available online: http://www.deh.gov.au/biodiversity/publications/technical/surrogates

Fettweis, M., Van den Eynde, D., 2003. The mud deposits and the high turbidity in the Belgian-Dutch coastal zone, southern bight of the North Sea. Continental Shelf Research 23, 669-691.

Ficetola, G.F., Thuiller, W., Padoa-Schioppa, E., 2009. From introduction to the establishment of alien species: bioclimatic differences between presence and reproduction localities in the slider turtle. Diversity and Distributions 15, 108-116.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38-49.

Findlay, S.E.G., 1981. Small-scale spatial distribution of meiofauna on a mud- and sandflat. Estuarine Coastal and Shelf Science 12, 471-484.

Fischer, A.G., 1960. Latitudinal Variations in Organic Diversity. Evolution 14, 64-81.

Forster, S., Graf, G., 1995. Impact of irrigation on oxygen flux into the sediment: intermittent pumping by *Callianassa subterranea* and 'piston-pumping' by *Lanice conchilega*. Marine Biology 123, 335-346.

Fox, B. J., Brown, J., 1993. Assembly rules for functional groups in North American desert rodent communities. Oikos 67, 358-370.

Fox, B.J., Fox, M.D., 2000. Factors determining mammal species richness on habitat islands and isolates: habitat diversity, disturbance, species interactions and guild assembly rules. Global Ecology & Biogeography 9, 19-37.

Franco, M.A., Soetaert, K., Van Oevelen, D., Van Gansbeke, D., Costa, M.J., Vincx, M., Vanaverbeke, J., 2008. Density, vertical distribution and trophic responses of metazoan meiobenthos to phytoplankton deposition in contrasting sediment types. Marine Ecology Progress Series 358, 51-62.

Franco, M.D., Vanaverbeke, J., Van Oevelen, D., Soetaert, K., Costa, M.J., Vincx, M., Moens, T., 2010. Respiration partitioning in contrasting subtidal sediments: seasonality and response to a spring phytoplankton deposition. Marine Ecology-an Evolutionary Perspective 31, 276-290.

Fraser, H.M., Greenstreet, S.P.R., Fryer, R.J., Piet, G.J., 2008. Mapping spatial variation in demersal fish species diversity and composition in the North Sea: accounting for species and size-related catchability in survey trawls. Ices Journal of Marine Science 65, 531-538.

Froese, R., Branch, T.A., Proelß, A., Quaas, M. , Sainsbury, K., Zimmermann, C., 2010. Generic harvest control rules for European fisheries. Fish and Fisheries, 12, early view, available online: http://dx.doi.org/10.1111/j.1467-2979.2010.00387.x

Gage, J.D., 1996. Why are there so many species in deep-sea sediments? Journal of Experimental Marine Biology and Ecology 200, 257-286.

Gaston, K.J., Fuller, R.A., 2008. Commonness, population depletion and conservation biology. Trends in Ecology & Evolution 23, 14-19.

Gastón, A., García-Viñas, J.I., 2011. Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. Ecological modelling 222, 2037-2041.

Geetanjali, Malhotra, S.K., Malhotra, A., Ansari, Z., Chatterji, A., 2002. Role of nematodes as bioindicators in marine and freshwater habitats. Current Science 82, 505-507.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural networkmodels. Ecological Modelling 160, 249-264.

Gevrey, M., Park, Y.-S., Oberdorff, T., Lek, S., 2005. Predicting fish assemblages in France and evaluating the influence of their environmental variables. In: Lek, S., Scardi, M., Verdonshot, P.F.M., Descy, J.-P., Park, Y.-S. (eds.), Modelling Community Structure in Freshwater Ecosystems, p. 54-63. Springer-Verlag, Berlin, 530 pp.

Ghertsos, K., Luczak, C., Dewarumez, J.M., Dauvin, J.C., 2000. Influence of spatial scales of observation on temporal change in diversity and trophic structure of fine-sand communities from the English Channel and the southern North Sea. Ices Journal of Marine Science 57, 1481-1487.

Gheskiere, T., Hoste, E., Vanaverbeke, J., Vincx, M., Degraer, S., 2004. Horizontal zonation patterns and feeding structure of marine nematode assemblages on a macrotidal, ultra-dissipative sandy beach (De Panne, Belgium). Journal of Sea Research 52, 211-226.

Giere, O., 2009. Meiobenthology. The microscopic motile fauna of aquatic sediments. 2nd edition, Springer-Verlag, Berlin, Germany, 527 pp.

Gingold, R., Ibarra-Obando, S.E., Rocha-Olivares, A., 2010a. Spatial aggregation patterns of free-living marine nematodes in contrasting sandy beach micro-habitats. Journal of the Marine Biological Association of the United Kingdom, 1-8.

Gingold, R., Mundo-Ocampo, M., Holovachov, O., Rocha-Olivares, A., 2010b. The role of habitat heterogeneity in structuring the community of intertidal free-living marine nematodes. Marine Biology 157, 1741-1753.

Glasser, S.P. (ed.), 2008. Essentials of Clinical Research. Springer, The Netherlands, 360 pp.

Glowka, L., Burhenne-Guilmin, F., Synge, H., McNeely, J.A., Gündling, L., 1994. A Guide to the Convention on Biological Diversity. Environmental Policy and Law Paper 30, IUCN-The World Conservation Union, Gland & Cambridge, UK, 161 pp.

Godet, L., Toupoint, N., Olivier, F., Fournier, J., Retiere, C., 2008. Considering the functional value of common marine species as a conservation stake: The case of sandmason worm *Lanice conchilega* (Pallas 1766) (Annelida, Polychaeta) beds. Ambio 37, 347-355.

Goethals, P.L.M., 2005. Data Driven Development of Predictive Ecological Models for Benthic Macroinvertebrates in Rivers. PhD Thesis, Ghent University, Faculty of of Agricultural and Applied Biological Sciences, Ghent, Belgium, 376 pp.

Gollasch, S., Minchin, D., Rosenthal, H., Voigt, M. (eds.), 1999. Exotics across the ocean. Case histories on introduced species: their general biology, distribution, range expansion and impact: prepared by Members of the European Union Concerted Action on testing monitoring systems for risk assessment of harmful introductions by ships to

European waters (MAS-CT-97-0111). Department of Fishery Biology, Institute for Marine Science, University of Kiel, Germany, 73 pp.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York, USA, 483 pp.

Gotelli, N.J., 2000. Null model analysis of species co-occurrence patterns. Ecology 81, 2606-2621.

Gotelli, N.J., 2001. Research frontiers in null model analysis. Global Ecology and Biogeography 10, 337-343.

Gotelli, N.J., Graves, G.R., 1996. Null Models in Ecology. Smithsonian Institution Press, Washington, DC, USA, 368 pp.

Gotelli, N.J., McCabe, D.J., 2002. Species co-occurrence: a meta-analysis of J.M. Diamond's assembly rules model. Ecology 83, 2091-2096.

Gotelli, N.J., Rohde, K., 2002. Co-occurrence of ectoparasites of marine fishes: a null model analysis. Ecology Letters 5, 86-94.

Gotelli, N.J., Entsminger, G.L., 2003. Swap algorithms in null model analysis. Ecology 84, 532-535.

Gotelli, N.J., Entsminger, G.L., 2005. EcoSim: Null models software for ecology. Version 7. Acquired Intelligence Inc. & Kesey-Bear. Jericho, VT 05465. Online available: http://garyentsminger.com/ecosim.htm.

Gotelli, N.J., Ulrich, W., 2010. Over-reporting bias in null model analysis: A response to Fayle and Manica. Ecological Modelling 222, 1337-1339.

Graf, G., 1992. Benthic-pelagic coupling: a benthic view. Oceanography and Marine Biology 30, 149-190.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19, 497-503.

Grassle, J.F., Sanders, H.L., 1973. Life histories and the role of disturbance. Deep-Sea Research 20, 643-659.

Grassle, J.F., Morse-Porteus, L.S., 1987. Macrofaunal colonisation of disturbed deep-sea environments and the structure of deep-sea benthic communities. Deep-Sea Research 34, 1911–1950.

Gray, J.S., Wu, R.S.S., Or, Y.Y., 2002. Effects of hypoxia and organic enrichment on the coastal marine environment. Marine Ecology Progress Series 238, 249-279.

Gross, K., Cardinale, B.J., 2007. Does species richness drive community production or vice versa? Reconciling historical and contemporary paradigms in competitive communities. The American Naturalist 170, 207-220.

Guilini, K., Soltwedel, T., van Oevelen, D., Vanreusel, A., 2011. Deep-Sea Nematodes Actively Colonise Sediments, Irrespective of the Presence of a Pulse of Organic Matter: Results from an In-Situ Experiment. Plos One 6, 12.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147-186.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993-1009.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., Peterson, A.T., 2007. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? Ecological Monographs 77, 615-630.

Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., Fujita, R., Heinemann, D., Lenihan, H.S., Madin, E.M.P., Perry, M.T., Selig, E.R., Spalding, M., Steneck, R., Watson, R., 2008. A global map of human impact on marine ecosystems. Science 319, 948-952.

Hammond, P.M., 1992. Species inventory. In: Groombridge, B. (ed.), Global diversity, status of the earth's living resources, p. 17-39, Chapman & Hall, London, UK, 585 pp.

Hampe, A., 2004. Bioclimate envelope models: what they detect and what they hide. Global Ecology and Biogeography 13, 469-471.

Hartmann-Schröder, G., 1996. Annelida, Borstenwürmer, Polychaeta [Annelida, bristleworms, Polychaeta]. In: The fauna of Germany and adjacent seas with their characteristics and ecology, p. 58. 2nd revised edition, Gustav Fischer, Jena, Germany, 648 pp.

Heino, J., 2009. Species co-occurrence, nestedness and guild-environment relationships in stream macroinvertebrates. Freshwater Biology 54, 1947-1959.

Heip, C., 1975. On the significance of aggregation in some benthic marine invertebrates. Proceedings of Ninth European Marine Biology Symposium, 527-538.

Heip, C., Engels, P., 1977. Spatial segregation in copepod species from a brackish water habitat. Journal of Experimental Marine Biology and Ecology 26, 77-96.

Heip, C., Herman, R., Vincx, M., 1983. Subtidal meiofauna of the North Sea: A review. Biologisch Jaarboek Dodonaea 51, 116-170.

Heip, C., Vincx, M., Vranken, G., 1985. The ecology of marine nematodes. Oceanography and Marine Biology 23, 399-489.

Hengl, T., 2007. A practical guide to geostatistical mapping of environmental variables. Office for Official Publications of the European Communities, Luxembourg, 143 pp., available online: http://eusoils.jrc.ec.europa.eu/esdb_archive/eusoils_docs/other/eur22904en.pdf

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75-93.

Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: From equations to case studies. Computers & Geosciences 33, 1301-1315.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773-785.

Hernandez, P.A., Franke, I., Herzog, S.K., Pacheco, V., Paniagua, L., Quintana, H.L., Soto, A., Swenson, J.J., Tovar, C., Valqui, T.H., Vargas, J., Young, B.E., 2008. Predicting species distributions in poorly-studied landscapes. Biodiversity and Conservation 17, 1353-1366.

Hernández-Stefanoni, J.L., Dupuy, J.M., 2007. Mapping species density of trees, shrubs and vines in a tropical forest, using field measurements, satellite multiespectral imagery and spatial interpolation. Biodiversity and Conservation 16, 3817-3833.

Hernández-Stefanoni, J.L., Ponce-Hernandez, R., 2004. Mapping the spatial distribution of plant diversity indices in a tropical forest using multi-spectral satellite image classification and field measurements. Biodiversity and Conservation 13, 2599-2621.

Heuers, J., Jaklin, S., Zühlke, R., Dittman, S., Günther, C.-P., Hildenbrandt, H., Grimm, V., 1998. A model on the distribution and abundance of the tube building polychaete *Lanice conchilega* (Pallas, 1766) in the intertidal of the Wadden Sea. Verhandlungen der Gesellschaft für Ökologie 28, 207-215.

Hijmans, R.J., Graham, C.H., 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12, 2272-2281.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83, 2027-2036.

Hoegh-Guldberg, O., Bruno, J.F., 2010. The impact of climate change on the World's marine ecosystems. Science 328, 1523-1528.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561-7578.

Hogue, E.W., Miller, C.B., 1981. Effects of sediment microtopography on small-scale spatial distributions of meiobenthic nematodes. Journal of Experimental Marine Biology and Ecology 53, 181-191.

Holt, A.C., Salkeld, D.J., Fritz, C.L., Tucker, J.R., Gong, P., 2009. Spatial analysis of plague in California: niche modelling predictions of the current distribution and potential response to climate change. International Journal of Health Geographics 8, 38.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359-366.

Hortal, J., Borges, P.A.V., Gaspar, C., 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75, 274-287.

Houziaux, J.-S., Kerckhof, F., Merckx, B., Vincx, M., Courtens, W., Stienen, E., Van Lancker, V., Craeymeersch, J., Van Hoey, G., Hostens, K., Degraer, S., 2010. Invasion of the southern bight of the North Sea by *Ensis directus*: ecological consequences and fishery perspectives C.M. International Council for the Exploration of the Sea K22, 10.

Hubbell, S.P., 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, USA, 375 pp.

Hugot, J.P., Baujard, P., Morand, S., 2001. Biodiversity in helminths and nematodes as a field of study: an overview. Nematology 3, 199-208.

Hurlbert, S.H., 1971. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52, 577-586.

Huston, M., 1979. A general hypothesis of species diversity. The American Naturalist 113, 81-101.

Hutchinson, G.E., 1957. Population studies - Animal ecology and demography - Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology 22, 415-427.

ICES, 2005. Report of the Working Group on Introductions and Transfers of Marine Organisms (WGITMO). By correspondence, ICES CM 2005/ACME:05, Copenhagen, Denmark, 173 pp., available online: http://www.ices.dk/reports/acme/2005/wgitmo05.pdf

Ingels, J., Tchesunov, A.V., Vanreusel, A., 2011. Meiofauna in the Gollum Channels and the Whittard Canyon, Celtic Margin-How Local Environmental Conditions Shape Nematode Structure and Function. Plos One 6, 15.

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques: Regression, Classication and Manifold Learning. Springer-Verlag, New York, USA, 732 pp.

Jackson, J.B.C., 1992. Pleistocene perspectives on coral reef community structure. American Zoologist 32, 719-731.

Joint, I.R., Gee, J.M., Warwick, R.M., 1982. Determination of fine-scale vertical distribution of microbes and meiofauna in an intertidal sediment. Marine Biology 72, 157-164.

Jones, C.G., Lawton, J.H., Shachak, M., 1994. Organisms as ecosystem engineers. Oikos, 69, 373-386.

Jones, S.E., Jago, C.F., 1993. In situ assessment of modification of sediment properties by burrowing invertebrates. Marine Biology 115, 133-142.

Journel, A.G., Huijbregts, C.J., 1978. Mining geostatistics. Academic Press Inc, London, UK, 600 pp.

Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14, 401-413.

Kamilar, J.M., Ledogar, J.A., 2011. Species Co-Occurrence Patterns and Dietary Resource Competition in Primates. American Journal of Physical Anthropology 144, 131-139.

Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., German, E., 2000. Case studies on the use of neural networks in eutrophication modelling. Ecological Modeling 134, 145-152.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. 1st edition, Wiley, New York, USA, 368 pp.

Kautsky, N., Evans, S., 1987. Role of biodeposition by *Mytilus edulis* in the circulation of matter and nutrients in a Baltic coastal ecosystem. Marine Ecology Progress Series 38, 201-212.

Kenchington, E., Duggan, R., Riddell, T., 1998. Early life history characteristics of the razor clam (*Ensis directus*) and the moonsnails (*Euspira* spp.) with applications to fisheries and aquaculture Canadian technical report of fisheries and aquatic sciences 2223, 1-32.

Kennedy, A.D., Jacoby, C.A., 1999. Biological indicators of marine environmental health: meiofauna - a neglected benthic component? Environmental Monitoring and Assessment 54, 47-68.

Kennish, M.J., Haag, S.M., Sakowicz, G.P., Durand, J.B., 2004. Benthic macrofaunal community structure along a well-defined salinity gradient in the Mullica River Great Bay estuary. Journal of Coastal Research, 209-226.

Kerckhof, F., Dumoulin, E., 1987. Eerste vondsten van de Amerikaanse zwaardschede *Ensis directus* (Conrad, 1843) langs de Belgische kust. De Strandvlo 7, 51-52.

Kerckhof, F., Haelters, J., Gollasch, S., 2007. Alien species in the marine and brackish ecosystem: the situation in Belgian waters. Aquatic Invasions 2, 243-257.

Kitanidis, P.K., 1993. Generalized Covariance Functions in Estimation. Mathematical Geology 25, 525-540.

Klopfer, P.H., 1959. Environmental determinants of faunal diversity. The American Naturalist 93, 337-342.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI) 2, 1137-1143.

Kondoh, M., 2001. Unifying the relationships of species richness to productivity and disturbance. Proceedings of the Royal Society of London Series B-Biological Sciences 268, 269-271.

Kotta, J., Orav-Kotta, H., Vuorinen, P., 2005. Field measurements on the variability in biodeposition and estimates of grazing pressure of suspension-feeding bivalves in the northern Baltic Sea. Comparative Roles of Suspension-Feeders in Ecosystems 47, 11-29.

Kristensen, E., 2008. Mangrove crabs as ecosystem engineers; with emphasis on sediment processes. Journal of Sea Research 59, 30-43.

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Diversity and Distributions 13, 66-69.

Lambshead, P.J.D., Boucher, G., 2003. Marine nematode deep-sea biodiversity-hyperdiverse or hype? Journal of Biogeography 30, 475-485.

Lambshead, P.J.D., Elce, B.J., Thistle, D., Eckman, J.E., Barnett, P.R.O., 1994. A comparison of the biodiversity of deep-sea marine nematodes from three stations in the Rockall Trough, Northeast Atlantic, and one station in the San Diego Trough, Northeast Pacific. Biodiversity Letters 2, 95–107.

Lambshead, P.J.D., Tietjen, J., Ferrero, T., Jensen, P., 2000. Latitudinal diversity gradients in the deep-sea with special reference to North Atlantic nematodes. Marine Ecology Progress Series 194, 159-167.

Lambshead, P.J.D., Tietjen, J., Glover, A., Ferrero, T., Thistle, D., Gooday, A., 2001. The impact of largescale natural physical disturbance on the diversity of deep-sea North Atlantic nematodes. Marine Ecology Progress Series 214, 121-126.

Lanckneus, J., Van Lancker, V.R.M., Moerkerke, G., Van den Eynde, D., Fettweis, M., De Batist, M., Jacobs, P., 2002. Onderzoek van natuurlijke zandtransporten op het Belgisch continentaal plat. BUDGET (Beneficial usage of data and geo-environmental techniques): samenvatting van het onderzoek. Eerste plan voor wetenschappelijke ondersteuning van een beleid gericht op duurzame ontwikkeling (PODO I) Programma "Duurzaam beheer van de Noordzee", Brussels, Belgium, 9 pp.

Lee, A.J., 1980. North Sea: physical oceanography. In: Banner, F.T., Collins, M.B., Massie, K.S. (eds.), The North-West European Shelf Sea: The Seabed and the Sea in Motion. II. Physical and Chemical Oceanography, and Physical Resources, p. 467-493, Elsevier, Amsterdam, Netherlands, 616 pp.

Lee, J.J., Tietjen, J.H., Mastropaolo, C., Rubin, H., 1977. Food quality and the heterogeneous spatial distribution of meiofauna. Helgolander Wissenschaftliche Meeresuntersuchungen 30, 272-282.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659-1673.

Legendre, P., Legendre, L., 1998. Numerical ecology. 2nd English edition, Elsevier Science BV, Amsterdam, Netherlands, 853 pp.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M., Gonzalez, A., 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7, 601-613.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120, 65-73.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Marine Freshwater Research 46, 1229-1236.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996a. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90, 39-52.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996b. Role of some environmental variables in trout abundance models using neural networks. Aquatic Living Resources 9, 23-29.

Levin, L.A., Demaster, D.J., McCann, L.D., Thomas, C.L., 1986. Effects of giant protozoans (class: Xenophyophorea) on deep-seamount benthos. Marine Ecology Progress Series 29, 99-104.

Levin, L.A., Etter, R.J., Rex, M.A., Gooday, A.J., Smith, C.R., Pineda, J., Stuart, C.T., Hessler, R.R., Pawson, D., 2001. Environmental influences on regional deep-sea species diversity. Annual Review of Ecology and Systematics 32, 51-93.

Li, J., Vincx, M., Herman, P.M.J., Heip, C., 1997. Monitoring meiobenthos using cm-, m- and km-Scales in the Southern Bight of the North Sea. Marine Environmental Research 43, 265-278.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255-268.

Liu C., Pam M., Dawson T.P., Pearson R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385-393.

Lobo, J.M., Jimenez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145-151.

Loiselle, B.A., Jorgensen, P.M., Consiglio, T., Jimenez, I., Blake, J.G., Lohmann, L.G., Montiel, O.M., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35, 105-116.

Long, D., 2006. BGS detailed explanation of seabed sediment modified folk classification, European Commission, Brussels, Belgium, 7 pp. Online available: http://ec.europa.eu/maritimeaffairs/emodnet/documents/standards/mesh_geology.pdf

Lorenzen, S., 1974. Die Nematodenfauna der sublitoralen Region der Deutschen Bucht, insbesondere im Titan-Abwassergebiet bei Helgoland. Veroffentlichungen des Instituts fur Meeresforschung in Bremerhaven, 14, 305-327.

Lorenzen, S., 2000. The role of the biogenetic convergence rule in polarizing transformation series - Arguments from nematology, chaos science, and phylogenetic systematics. Annales Zoologici 50, 267-275.

MacArthur, R.H., MacArthur, J.W., 1961. On bird species diversity. Ecology 42, 594- 598.

MacArthur, R.H., Wilson, E.O., 1967. The Theory of Island Biogeography. Princeton University Press, Princeton, USA, 203 pp.

Mackey, R.L., Currie, D.J., 2001. The diversity-disturbance relationship: Is it generally strong and peaked? Ecology 82, 3479-3492.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software v15, 101-124.

Malakoff, D., 2004. Marine science - New tools reveal treasures at ocean hot spots. Science 304, 1104-1105.

Mangelsdorf, J., Scheurmann, K., Weiss, F.H., 1990. River morphology. A guide for geoscientists and engineers. Springer-Verlag, Berlin, Germany, 243 pp.

Masters, T., 1993. Practical Neural Network Recipes in C++. Academic Press, San Diego, CA, USA, 493 pp.

Matheron, G., 1963. Principles of Geostatistics. Economic Geology, 58, 1246-1266.

May, R.M., 1988. How many species are there on the Earth? Science 241, 1441-1449.

Mello, L.G.S., Rose, G.A., 2005. Using geostatistics to quantify seasonal distribution and aggregation patterns of fishes: an example of Atlantic cod (*Gadus morhua*). Canadian Journal of Fisheries and Aquatic Sciences 62, 659-670.

Menge, B.A., Sutherland, J.P., 1976. Species diversity gradients: synthesis of the roles of predation, competition, and temporal heterogeneity. The American Naturalist 110 (973), 351-369.

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

Merckx, B., Van Meirvenne, M., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2010. Mapping nematode diversity in the Southern Bight of the North Sea. Marine Ecology Progress Series 406, 135-145.

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

Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I., Gough, L., 2001. What is the observed relationship between species richness and productivity? Ecology 82, 2381-2396.

Moens, T., Vincx, M., 1997. Observations on the feeding ecology of estuarine nematodes. Journal of the Marine Biological Association of the United Kingdom 77, 211-227.

Moens, T., dos Santos, G.A.P., 2010. Horizontal and vertical interactions and the structure and functioning of marine nematode assemblages. In: Gheerardyn, H. *et al.* (eds.), Book of Abstracts, Fourteenth International Meiofauna Conference (FouthIMCo),

Aula Academica, Ghent, 11-16 July 2010, p. 9., VLIZ Special Publication 44, Ostend, Belgium, 237 pp.

Moens, T., Verbeeck, L., de Maeyer, A., Swings, J., Vincx, M., 1999. Selective attraction of marine bacterivorous nematodes to their bacterial food. Marine Ecology-Progress Series 176, 165-178.

Moens, T., Herman, P., Verbeeck, L., Steyaert, M., Vincx, M., 2000. Predation rates and prey selectivity in two predacious estuarine nematode species. Marine Ecology-Progress Series 205, 185-193.

Moens, T., Yeates, G.W., Ley, P., 2004. Use of carbon and energy sources by nematodes. In: Cook, R.C., Hunt, D.J. (eds.), Proceeding of the Fourth International Congress of Nematology, June 2002, Tenerife Spain, p. 529-545. Brill, Leiden, Netherlands, 866 pp.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17-23.

Mouillot, D., Dumay, O., Tomasini, J.A., 2007. Limiting similarity, niche filtering and functional diversity in coastal lagoon fish communities. Estuarine Coastal and Shelf Science 71, 443-456.

Mühlenhardt-Siegel, U., Dörjes, J., von Cosel, R., 1983. Die amerikanische Schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. II. Populationsdynamik. Senckenbergiana Maritima 15, 93-110.

Murray-Smith, C., Brummitt, N.A., Oliveira-Filho, A.T., Bachman, S., Moat, J., Lughadha, E.M.N., Lucas, E.J., 2009. Plant diversity hotspots in the Atlantic Coastal forests of Brazil. Conservation Biology 23, 151-163.

Muylaert, K., Gonzales, R., Franck, M., Lionard, M., Van der Zee, C., Cattrijsse, A., Sabbe, K., Chou, L., Vyverman, W., 2006. Spatial variation in phytoplankton dynamics in the Belgian coastal zone of the North Sea studied by microscopy, HPLC-CHEMTAX and underway fluorescence recordings Journal of Sea Research 55, 253-265.

Ng, A.Y., Jordan, M.I., 2001. On discriminative versus generative classifiers: a comparison of logistic regression and naive Bayes. Advances in Neural Information Processing Systems 14, 605-610.

Nielsen, U.N., Osler, G.H.R., Campbell, C.D., Neilson, R., Burslem, D., van der Wal, R., 2010. The Enigma of Soil Animal Species Diversity Revisited: The Role of Small-Scale Heterogeneity. Plos One 5.

Nix, H.A., 1986. BIOCLIM - a Bioclimatic Analysis and Prediction System. Research report, CSIRO Division of Water and Land Resources 1983-1985, 59-60.

Odland, J., 1988. Spatial Autocorrelation. Sage Publications, Newbury Park, CA, USA, 87 pp.

Olden, J.D., Jackson, D.A., 2002. Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154, 135-150.

Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. The Quarterly Review in Biology 83, 171-193.

Olson, D.L., Delen., D., 2008. Advanced Data Mining Techniques. Springer, New York, USA, 180 pp.

Ortega-Huerta, M.A., Peterson, A.T., 2008. Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods. Revista Mexicana De Biodiversidad 79, 205-216.

Ovcharenko, S.O., Gollasch, S., 2009. *Ensis americanus* (Gould), American Jack knife clam (Solenidae, Mollusca). In: DAISIE, Handbook of alien species in Europe, p. 281. Springer, Dordrecht, Netherlands, 400 pp.

Pagliosa, P.R., 2005. Another diet of worms: the applicability of polychaete feeding guilds as a useful conceptual framework and biological variable. Marine Ecology-an Evolutionary Perspective 26, 246-254.

Paine, R.T., 1969. A Note on Trophic Complexity and Community Stability. The American Naturalist 103, 91-93.

Panatier, Y., 1996. Variowin: software for spatial data analysis in 2D, statistics and computing. Springer-Verlag, New York, USA, 91 pp.

Parisien, M.A., Moritz, M.A., 2009. Environmental controls on the distribution of wildfire at multiple spatial scales. Ecological Monographs 79, 127-154.

Park, Y., Van Mol, B., Ruddick, K., 2006. Validation of MERIS water products for Belgian coastal waters: 2002-2005. In: Danesy, D. (ed.), Proceedings of the 2nd working meeting on MERIS and AATSR calibration and geophysical validation (MAVT-2006), 20-24 March 2006, ESRIN, Frascati, Italy, p.1-7. ESA Special Publications 615, Noordwijk, Netherlands, available online: http://envisat.esa.int/workshops/mavt_2006/

Park, Y.-S., Céréghino, R., Compin, A., Lek, S., 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling 160, 265-280.

Park, Y.-S., Verdonschot, P.F.M., Lek, S., 2005. Reviewofmodelling techniques. In: Lek, S., Scardi, M., Verdonshot, P.F.M., Descy, J.-P., Park, Y.-S. (eds.), Modelling Community

Structure in Freshwater Ecosystems, p. 21-40, Springer-Verlag, Berlin, Germany, 530 pp.

Parolo, G., Rossi, G., Ferrarini, A., 2008. Toward improved species niche modelling: *Arnica montana* in the Alps as a case study. Journal of Applied Ecology 45, 1410-1418.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R., Torres, F., 1998. Fishing down marine food webs. Science 279, 860-863.

Pearson, R.G. 2007. Species' Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History, New York, USA, 50 pp., online only: http://biodiversityinformatics.amnh.org/files/SpeciesDistModelingSYN_1-16-08.pdf

Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102-117.

Perry, A.L., Low, P.J., Ellis, J.R., Reynolds, J.D., 2005. Climate change and distribution shifts in marine fishes. Science 308, 1912-1915.

Peterson, A.T., Papes, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550-560.

Phillips, S., 2010. A Brief Tutorial on Maxent. AT&T Research, New York, USA, 38 pp., online only: http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161-175.

Phillips, S.J., Dudík, M., Schapire, R.E., 2004. A maximum entropy approach to species distribution modeling. In: Brodley, C.E. (ed.), Proceedings of the 21st international conference on machine learning, p. 655-662, AMC Press, New York, USA, available online: http://www.informatik.uni-trier.de/~ley/db/conf/icml/icml2004.html

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231-259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181-197.

Pingree, R.D., Griffiths, D.K., 1979. Sand transport paths around the British Isles resulting from M2 and M4 tidal interactions. Journal of the Marine Biological Association of the United Kingdom 59, 497-513.

Pomeroy, R.S., Parks, J.E. and Watson, L.M., 2004. How is your MPA doing? A Guidebook of Natural and Social Indicators for Evaluating Marine Protected Area Management Effectiveness. IUCN/NOAA/WWF Guidebook. IUCN, Gland, Switzerland and Cambridge, UK, 216 pp.

Postma-Blaauw, M.B., de Vries, F.T., de Goede, R.G.M., Bloem, J., Faber, J.H., Brussaard, L., 2005. Within-trophic group interactions of bacterivorous nematode species and their effects on the bacterial community and nitrogen mineralization. Oecologia 142, 428-439.

Pritchard, J.R., Schlüter, D., 2001. Declining interspecific competition during character displacement: Summoning the ghost of competition past. Evolutionary Ecology Research 3, 209-220.

Purves, D.W., Turnbull, L.A., 2010. Different but equal: the implausible assumption at the heart of neutral theory. Journal of Animal Ecology 79, 1215-1225.

Rabaut, M., 2009. *Lanice conchilega*, fisheries and marine conservation: towards an ecosystem approach to marine management. PhD Thesis, Ghent University, Faculty of Sciences, Marine Biology Section, Ghent, Belgium, 354 pp.

Rabaut, M., Guilini, K., Van Hoey, G., Magda, V., Degraer, S., 2007. A bio-engineered soft-bottom environment: The impact of Lanice conchilega on the benthic species-specific densities and community structure. Estuarine Coastal and Shelf Science 75, 525-536.

Rabaut, M., Braeckman, U., Hendrickx, F., Vincx, M., Degraer, S., 2008. Experimental beam-trawling in Lanice conchilega reefs: Impact on the associated fauna. Fisheries Research 90, 209-216.

Rabaut, M., Vincx, M., Degraer, S., 2009. Do *Lanice conchilega* (sandmason) aggregations classify as reefs? Quantifying habitat modifying effects. Helgoland Marine Research 63, 37-46.

Rabaut, M., Van de Moortel, L., Vincx, M., Degraer, S., 2010. Biogenic reefs as structuring factor in Pleuronectes platessa (Plaice) nursery. Journal of Sea Research 64, 102-106.

Raes, N., ter Steege, H., 2007. A null-model for significance testing of presence-only species distribution models. Ecography 30, 727-736.

Reese, D.C., Brodeur, R.D., 2006. Identifying and characterizing biological hotspots in the northern California Current. Deep-Sea Research Part Ii-Topical Studies in Oceanography 53, 291-314.

Reidenauer, J.A., 1989. Sand-dollar *Melitta quinquiesperforata* (Leske) burrow trails: sites of harpacticoid disturbance and nematode attraction. Journal of Experimental Marine Biology and Ecology 130, 223-235.

Reise, K., 1981. High abundance of small zoobenthos around biogenic structures in tidal sediments of the Wadden Sea. Helgolander Meeresuntersuchungen 34, 413-425.

Rex, M.A., 1981. Community structure in the deep-sea benthos. Annual Review of Ecology and Systematics 12, 331-353.

Rex, M.A., Stuart, C.T., Hessler, R.R., Allen, J.A., Sanders, H.L., Wilson, G.D.F., 1993. Global-scale latitudinal patterns of species diversity in the deep-sea benthos. Nature 365, 636-639.

Ribichich, A.M., 2005. From null community to non-randomly structured actual plant assemblages: parsimony analysis of species co-occurrences. Ecography 28, 88-98.

Richerson, P., Armstrong, R. and Goldman, C.R., 1970. Contemporaneous disequilibrium, a new hypothesis to explain the 'paradox of plankton.' Proceedings of the National Academy of Sciences 67, 1710-1714.

Ricklefs, R.E., Schluter, D., 1993. Species diversity in ecological communities. Historical and geographical perspectives. The University of Chicago Press, Chicago and London, UK, 414 pp.

Riordan, E.C., Rundel, P.W., 2009. Modelling the distribution of a threatened habitat: the California sage scrub. Journal of Biogeography 36, 2176-2188.

Rios-Lara, V., Salas, S., Javier, B.P., Irene-Ayora, P., 2007. Distribution patterns of spiny lobster (*Panulirus argus*) at Alacranes reef, Yucatan: Spatial analysis and inference of preferential habitat. Fisheries Research 87, 35-45.

Rodriguez, J.P., Brotons, L., Bustamante, J., Seoane, J., 2007. The application of predictive modelling of species distribution to biodiversity conservation. Diversity and Distributions 13, 243-251.

Rohde, K., 1992. Latitudinal Gradients in Species Diversity: The Search for the Primary Cause. Oikos 65, 514-527.

Rosemann, M., 2006a. Potential Pitfalls of Process Modelling (Part A). Business Process Management Journal 12, 249-254.

Rosemann, M., 2006b. Potential Pitfalls of Process Modelling (Part B). Business Process Management Journal 12, 377-384.

Roura-Pascual, N., Brotons, L., Peterson, A.T., Thuiller, W., 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biological Invasions 11, 1017-1031.

Rzonzef, L., 1993. Effecten op hetmarien leefmilieu van de zand- en grindwinningen op het Belgisch Kontinentaal Plat. Annalen der mijnen van België 2, 1-49.

Samson, F.B., Knopf, F.L., 1982. In search of a diversity ethic for wildlife management. Transactions of the North American Wildlife and Natural Resources Conference 47, 421-431.

Sanders, H.L., 1968. Marine benthic diversity: a comparative study. The American Naturalist 102, 243-282.

Sanders, N.J., Gotelli, N.J., Wittman, S.E., Ratchford, J.S., Ellison, A.M., Jules, E.S., 2007. Assembly rules of ground-foraging ant assemblages are contingent on disturbance, habitat and spatial scale. Journal of Biogeography 34, 1632-1641.

Sawada, M., 1999. ROOKCASE: an excel 97/2000 visual basic (VB) add-in for exploring global and local spatial autocorrelation. Bulletin of the Ecological Society of America 80, 231-234.

Scardi, M., Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecological Modelling 120, 213-223.

Schiedek, D. and Zebe, E., 1987. Functional and environmental anaerobiosis in the razor clam *Ensis directus* (Mollusca: Bivalvia). Marine Biology 94, 31-37.

Schluter, D., 1984. A Variance Test for Detecting Species Associations, with Some Example Applications. Ecology 65, 998-1005.

Schmid, B., 2002. The species richness-productivity controversy. Trends in Ecology & Evolution 17, 113-114.

Schratzberger, M., Warwick, R.M., 1998. Effects of physical disturbance on nematode communities in sand and mud: a microcosm experiment. Marine Biology 130, 643-650.

Schratzberger, M., Gee, J.M., Rees, H.L., Boyd, S.E., Wall, C.M., 2000a. The structure and taxonomic composition of sublittoral meiofauna assemblages as an indicator of the status of marine environments. Journal of the Marine Biological Association of the United Kingdom 80, 969-980.

Schratzberger, M., Rees, H.L., Boyd, S.E., 2000b. Effects of simulated deposition of dredged material on structure of nematode assemblages - the role of burial. Marine Biology 136, 519-530.

Schratzberger, M., Dinmore, T.A., Jennings, S., 2002. Impacts of trawling on the diversity, biomass and structure of meiofauna assemblages. Marine Biology 140, 83-93.

Schratzberger, M., Lampadariou, N., Somerfield, P., Vandepitte, L., Vanden Berghe, E., 2009. The impact of seabed disturbance on nematode communities: linking field and laboratory observations. Marine Biology 156, 709-724.

Schrijvers, J., Okondo, J., Steyaert, M., Vincx, M., 1995. Influence of epibenthos on meiobenthos of the Ceriops tagal mangrove sediment at Gazi Bay, Kenya. Marine Ecology Progress Series 128, 247-259.

Schroeder, L.D., Sjoquist, D.L., Stephan, P.E., 1986. Understanding regression analysis: an introductory guide. Sage Publications, Beverly Hills, CA, USA, 95 pp.

Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31, 1555-1568.

Segurado, P.,Araújo, M.B.,Kunin,W.E., 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology 43, 433-444.

Semmens, B.X., Auster, P.J., Paddack, M.J., 2010. Using Ecological Null Models to Assess the Potential for Marine Protected Area Networks to Protect Biodiversity. Plos One 5.

Sergio, C., Figueira, R., Draper, D., Menezes, R., Sousa, A.J., 2007. Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. Biological Conservation 135, 341-351.

Sfenthourakis, S., Tzanatos, E., Giokas, S., 2006. Species co-occurrence: the case of congeneric species and a causal approach to patterns of species association. Global Ecology and Biogeography 15, 39-49.

Shahin, M.A., Maier, H.R., Jaksa, M.B., 2004. Data division for developing neural networks applied to geotechnical engineering. Journal of Computing in Civil Engineering 18, 105-114.

Shannon, C.E., 1948. A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379-423, 623-656.

Shephard, S., Brophy, D., Reid, D.G., 2010. Can bottom trawling indirectly diminish carrying capacity in a marine ecosystem? Marine Biology 157, 2375-2381.

Shi, J.J.S., 2000. Reducing prediction error by transforming input data for neural networks. Journal of Computing in Civil Engineering 14 (2), 109-116.

Simberloff, D.S., 1972. Properties of the rarefaction diversity measurement. The American Naturalist 106, 414-418.

Simberloff, D., Connor, E.F., 1981. Missing species combinations. The American Naturalist 118, 215-239.

Smith, M.D., Knapp, A.K., 2003. Dominant species maintain ecosystem function with non-random species loss. Ecology Letters 6, 509-517.

Smith, M.D., Wilcox, J.C., Kelly, T., Knapp, A.K., 2004. Dominance not richness determines invasibility of tallgrass prairie. Oikos 106, 253-262.

Soetaert, K., Heip, C., 1995. Nematode assemblages of deep-sea and shelf break sites in the North Atlantic and Mediterranean Sea. Marine Ecology Progress Series 125, 171-183.

Soetaert, K., Vincx, M., Wittoeck, J., Tulkens, M., Vangansbeke, D., 1994. Spatial Patterns of Westerschelde Meiobenthos. Estuarine Coastal and Shelf Science 39, 367-388.

Soetaert, K., Vincx, M., Wittoeck, J., Tulkens, M., 1995. Meiobenthic distribution and nematode community structure in 5 European estuaries. Hydrobiologia 311, 185-206.

Solan, M., Cardinale, B.J., Downing, A.L., Engelhardt, K.A.M., Ruesink, J.L., Srivastava, D.S., 2004. Extinction and ecosystem function in the marine benthos. Science 306, 1177-1180.

Somerfield, P.J., Dashfield, S.L., Warwick, R.M., 2007. Three-dimensional spatial structure: nematodes in a sandy tidal flat. Marine Ecology Progress Series 336, 177-186.

Somerfield, P.J., Arvanitidis, C., Faulwetter, S., Chatzigeorgiou, G., Vasileiadou, A., Amouroux, J.M., Anisimova, N., Cochrane, S.J., Craeymeersch, J., Dahle, S., Denisenko, S., Dounas, K., Duineveld, G., Gremare, A., Heip, C.H.R., Herrmann, M., Karakassis, I., Kedra, M., Kendall, M.A., Kingston, P., Kotwicki, L., Labrune, C., Laudien, J., Nevrova, H., Nicolaidou, A., Occhipinti-Ambrogi, A., Palerud, R., Petrov, A., Rachor, E., Revkov, N., Rumohr, H., Sarda, R., Janas, U., Vanden Berghe, E., Wlodarska-Kowalczuk, M., 2009. Assessing evidence for random assembly of marine benthic communities from regional species pools. Marine Ecology Progress Series 382, 279-286.

Stachura-Skierczynska, K., Tumiel, T., Skierczynski, M., 2009. Habitat prediction model for three-toed woodpecker and its implications for the conservation of biologically valuable forests. Forest Ecology and Management 258, 697-703.

Steyaert, M., 2003. Spatial and temporal scales of nematode communities in the North Sea and Westerschelde. Ghent University, Faculty of Sciences, Marine Biology Section, Ghent, Belgium, 114 pp.

Steyaert, M., Garner, N., van Gansbeke, D., Vincx, M., 1999. Nematode communities from the North Sea: environmental controls on species diversity and vertical distribution within the sediment. Journal of the Marine Biological Association of the United Kingdom 79, 253-264.

Steyaert, M., Vanaverbeke, J., Vanreusel, A., Barranguet, C., Lucas, C., Vincx, M., 2003. The importance of fine-scale, vertical profiles in characterising nematode community structure. Estuarine Coastal and Shelf Science 58, 353-366.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13, 143-158.

Stone, L., Roberts, A., 1990. The checkerboard score and species distributions. Oecologia 85, 74-79.

Stone, L., Roberts, A., 1992. Competitive exclusion, or species aggregation? An aid in deciding. Oecologia 91, 419-424.

Suarez-Seoane, S., de la Morena, E.L.G., Prieto, M.B.M., Osborne, P.E., de Juana, E., 2008. Maximum entropy niche-based modelling of seasonal changes in little bustard (*Tetrax tetrax*) distribution. Ecological Modelling 219, 17-29.

Sun, B., Fleeger, J.W., Carney, R.S., 1993. Sediment microtopography and the small-scale spatial distribution of meiofauna. Journal of Experimental Marine Biology and Ecology 167, 73-90.

Svensson, J.R., Lindegarth, M., Pavia, H., 2010. Physical and biological disturbances interact differently with productivity: effects on floral and faunal richness. Ecology 91, 3069-3080.

Swennen, C., Leopold, M.F., Stock, M., 1985. Notes on growth and behaviour of the American razor clam *Ensis directus* in the Wadden Sea and the predation on it by birds. Helgoländer Meeresuntersuchungen 39, 255-261.

Telford, R.J., Birks, H.J.B., 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quaternary Science Reviews 24, 2173-2179.

Tews, J., Brose, U., Grimm, V., Tielborger, K., Wichmann, M.C., Schwager, M., Jeltsch, F., 2004. Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. Journal of Biogeography 31, 79-92.

Thuiller, W., Richardson, D.M., Pysek, P., Midgley, G.F., Hughes, G.O., Rouget, M., 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biology 11, 2234-2250.

Tietjen, J., 1984. Distribution and species diversity of deep-sea nematodes in the Venezuela Basin. Deep-Sea Research 31, 119-132.

Tietjen, J., 1989. Ecology of deep-sea nematodes from the Puerto Rico Trench area and Hatteras Abyssal Plain. Deep-Sea Research 36, 1579-1594.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Vanden Berghe, E., Worm, B., 2010. Global patterns and predictors of marine biodiversity across taxa. Nature 466, 1098-U1107.

Tokeshi, M., 1986. Resource utilization, overlap and temporal community dynamics : a null model analysis of an epiphytic chironomid community. Journal of Animal Ecology 55, 491-506.

Tomašových, A., 2008. Evaluating neutrality and the escalation hypothesis in brachiopod communities from shallow, high-productivity habitats. Evolutionary Ecology Research 10, 667-698.

Toropova, C., Meliane, I., Laffoley, D., Matthews, E. and Spalding, M. (eds.), 2010. Global Ocean Protection: Present Status and Future Possibilities. Brest, France: Agence des aires marines protégées, Gland, Switzerland, Washington, DC and New York, USA: IUCN WCPA, Cambridge, UK : UNEP-WCMC, Arlington, USA: TNC, Tokyo, Japan: UNU, New York, USA: WCS., 96 pp., available online: http://data.iucn.org/dbtw-wpd/edocs/2010-053.pdf

Toupoint, N., Godet, L., Fournier, J., Retiere, C., Olivier, F., 2008. Does Manila clam cultivation affect habitats of the engineer species *Lanice conchilega* (Pallas, 1766)? Marine Pollution Bulletin 56, 1429-1438.

Tulp, I., Craeymeersch, J., Leopold, M., van Damme, C., Fey, F., Verdaat, H., 2010. The role of the invasive bivalve *Ensis directus* as food source for fish and birds in the Dutch coastal zone. Estuarine Coastal and Shelf Science 90, 116-128.

Ülgen, O., Gunal, A., Shore, J., 1996. Pitfalls of simulation modeling and how to avoid them by using a robust simulation methodology. Proceedings of the 1996 Winter Auto Simulations Symposium, Bountiful, Utah, USA, 21-31.

Ulrich, W., 2004. Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules. Oikos 107, 603-609.

Ulrich, W., Almeida, M., Gotelli, N.J., 2009. A consumer's guide to nestedness analysis. Oikos 118, 3-17.

Van Colen, C., De Backer, A., Meulepas, G., van der Wal, D., Vincx, M., Degraer, S., Ysebaert, T., 2010. Diversity, trait displacements and shifts in assemblage structure of tidal flat deposit feeders along a gradient of hydrodynamic stress. Marine Ecology Progress Series 406, 79-89.

van der Zee, C., Chou, L., 2005. Seasonal cycling of phosphorus in the southern bight of the North Sea. Biogeosciences 2, 27-42.

Van Hoey, G., 2006. Spatio-temporal variability within the macrobenthic *Abra alba* community, with emphasis on the structuring role of *Lanice conchilega*. PhD Thesis, Ghent University, Faculty of Sciences, Marine Biology Section, Ghent, Belgium, 187 pp.

Van Hoey, G., Degraer, S., Vincx, M., 2004. Macrobenthic community structure of soft-bottom sediments at the Belgian Continental Shelf. Estuarine, Coastal and Shelf Science 59, 599-613.

Van Hoey, G., Guilini, K., Rabaut, M., Vincx, M., Degraer, S., 2008. Ecological implications of the presence of the tube-building polychaete *Lanice conchilega* on soft-bottom benthic ecosystems. Marine Biology 154, 1009-1019.

Van Meirvenne, M., 2007. Geostatistics. Course at Department of Soil Management and Soil Care, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium, 170 pp.

Van Meirvenne, M., Meklit, T., Verstraete, S., De Boever, M., Tack, F., 2008. Could shelling in the First World War have increased copper concentrations in the soil around Ypres? European Journal of Soil Science 59, 372-379.

van Oevelen, D., Soetaert, K., Franco, M.A., Moodley, L., van Ijzerloo, L., Vincx, M., Vanaverbeke, J., 2009. Organic matter input and processing in two contrasting North Sea sediments: insights from stable isotope and biomass data. Marine Ecology Progress Series 380, 19-32.

van Teeffelen, A.J.A., Ovaskainen, O., 2007. Can the cause of aggregation be inferred from species distributions? Oikos 116, 4-16.

Vanaverbeke, J., Vincx, M., 2008. Short-term changes in nematode communities from an abandoned intense sand extraction site on the Kwintebank (Belgian Continental Shelf) two years post-cessation. Marine Environmental Research 66, 240-248.

Vanaverbeke, J., Gheskiere, T., Steyaert, M., Vincx, M., 2002. Nematode assemblages from subtidal sandbanks in the Southern Bight of the North Sea: effect of small sedimentological differences. Journal of Sea Research 48, 197-207.

Vanaverbeke, J., Soetaert, K., Vincx, M., 2004a. Changes in morphometric characteristics of nematode communities during a spring phytoplankton bloom deposition. Marine Ecology Progress Series 273, 139-146.

Vanaverbeke, J., Steyaert, M., Soetaert, K., Rousseau, V., Van Gansbeke, D., Parent, J.-Y., Vincx, M., 2004b. Changes in structural and functional diversity of nematode

communities during a spring phytoplankton bloom in the southern North Sea. Journal of Sea Research 52, 281-292.

Vanaverbeke, J., Deprez, T., Vincx, M., 2007. Changes site in nematode communities at the long-term sand extraction of the Kwintebank (Southern Bight of the North Sea). Marine Pollution Bulletin 54, 1351-1360.

Vanaverbeke, J., Braeckman, U., Claus, S., Courtens, W., De Hauwere, N., Degraer, S., Deneudt, K., Goffin, A., Mees, J., Merckx, B., Provoost, P., Rabaut, M., Soetaert, K., Stienen, E., Vincx, M., 2009. Long-term data from the Belgian Continental Shelf in the framework of science-based management of the coastal North Sea: Report of the WestBanks integrative workshop, October 2008. Belgian Science Policy, Brussels, Belgium, 23 pp.

Vanaverbeke, J., Merckx, B., Degraer, S., Vincx, M., 2011. Sediment-related distribution patterns of nematodes and macrofauna: Two sides of the benthic coin? Marine Environmental Research 71, 31-40.

Vandepitte, L., Vanaverbeke, J., Vanhoorne, B., Hernandez, F., Bezerra, T.N., Mees, J., Vanden Berghe, E., 2009. The MANUELA database: an integrated database on meiobenthos from European marine waters. Meiofauna Marina 17, 35-60.

Vanreusel, A., 1990. Ecology of the free-living marine nematodes from the Voordelta (Southern Bight of the North Sea). I. Species composition and structure of the nematode communities. Cahiers De Biologie Marine 31, 439-462.

Vanreusel, A., 1991. Ecology of the free-living marine nematodes in The Voordelta (Southern Bight of the North Sea). II. Habitat preferences of the dominant species. Nematologica 37, 343-359.

Vanreusel, A., Fonseca, G., Danovaro, R., da Silva, M.C., Esteves, A.M., Ferrero, T., Gad, G., Galtsova, V., Gambi, C., Genevois, V.D., Ingels, J., Ingole, B., Lampadariou, N., Merckx, B., Miljutin, D., Miljutina, M., Muthumbi, A., Netto, S., Portnova, D., Radziejewska, T., Raes, M., Tchesunov, A., Vanaverbeke, J., Van Gaever, S., Venekey, V., Bezerra, T.N., Flint, H., Copley, J., Pape, E., Zeppilli, D., Martinez, P.A., Galeron, J., 2010. The contribution of deep-sea macrohabitat heterogeneity to global nematode diversity. Marine Ecology-an Evolutionary Perspective 31, 6-20.

Verfaillie, E., Van Lancker, V., Van Meirvenne, M., 2006. Multivariate geostatistics for the predictive modelling of the surficial sand distribution in shelf seas. Continental Shelf Research 26, 2454-2468.

Villa, F., Tunesi, L., Agardy, T., 2002. Zoning marine protected areas through spatial multiple-criteria analysis: the case of the Asinara Island National Marine Reserve of Italy. Conservation Biology 16, 515-526.

Vincx, M., 1989a. Free-living marine nematodes from the southern bight of the North sea. Mededelingen van de Koninklijke Academie voor Wetenschappen, Letteren en Schone Kunsten van België Academia Analecta 51, 39-70.

Vincx, M., 1989b. Seasonal fluctuations and production of nematode communities in the Belgian coastal zone of the North Sea. Verhandelingen van het symposium 'Invertebraten van België', 57-66.

Vincx, M., 1990. Diversity of the nematode communities in the Southern Bight of the North Sea. Netherlands Journal of Sea Research 25, 181-188.

Vincx, M., Heip, C., 1987. The use of meiobenthos in pollution monitoring studies. A review. ICES Techniques in Marine Environmental Sciences 16, 50-67.

Vincx, M., Meire, P., Heip, C., 1990. The distribution of nematode communities in the Southern Bight of the North-Sea. Cahiers De Biologie Marine 31, 107-129.

von Cosel, R., 1982. Die amerikanische Schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. Senckenbergiana Maritima 14, 147-173.

von Cosel, R., 2009. The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). Basteria 73, 9-56.

Wackernagel, H., 2003. Multivariate Geostatistics. An Introduction with Applications, Third Edition, Springer Verlag, Berlin, Germany, 390 pp.

Walczak, S., Cerpa, N., 1999. Heuristic principles for the design of artificial neural networks. Information and Software Technology 41, 107-117.

Walker, J.S., Balling, R.C., Briggs, J.M., Katti, M., Warren, P.S., Wentz, E.A., 2008. Birds of a feather: Interpolating distribution patterns of urban birds. Computers Environment and Urban Systems 32, 19-28.

Webb, C.O., 2000. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. The American Naturalist 156, 145-155.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists. John Wiley & Sons, New York, USA, 315 pp.

Weiher, E., Keddy, P.A., 1995. Assembly rules, null models, and trait dispersion: new questions front old patterns. Oikos 74, 159-164.

Weiher, E., Keddy, P.A. (eds.), 2001. Ecological Assembly Rules: Perspectives, Advances, Retreats. First paperback edition with corrections. Cambridge University Press, Cambridge, UK, 418 pp.

Weinsheimer, F., Mengistu, A.A., Rödder, D., 2010. Potential distribution of threatened *Leptopelis* ssp. (Anura, Arthroleptidae) in Ethiopia derived from climate and land-cover data. Endangered Species Research 9, 117-124.

Whitfield, J., 2002. Neutrality versus the niche. Nature 417, 480-481.

Whittaker, R.H., 1972. Evolution and measurement of species diversity. Taxon 21, 213-251.

Widdicombe, S., Austen, M.C., 1998. Experimental evidence for the role of *Brissopsis lyrifera* (Forbes, 1841) as a critical species in the maintenance of benthic diversity and the modification of sediment chemistry. Journal of Experimental Marine Biology and Ecology 228, 241-255.

Wieser, W., 1953. Die Beziehung zwischen Mundhohlengestalt, Ernahrungsweise und Vorkommen bei freilebenden marinen nematoden. Arkiv för Zoologi 4, 439-484.

Wieser, W., 1960. Benthic studies in Buzzards Bay. II. The meiofauna. Limnology and Oceanography 5, 121-137.

Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R., Stewart, A.M., 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16, 120-127.

Willems, W., Goethals, P., Van den Eynde, D., Van Hoey, G., Van Lancker, V., Verfaillie, E., Vincx, M., Degraer, S., 2008. Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*. Ecological Modelling 212, 74-79.

Wilson, J.B. 1990. Mechanisms of species coexistence: twelve explanations for Hutchinson's 'paradox of the plankton': evidence from New Zealand plant communities. New Zealand Journal of Ecology 13, 17-42.

Wilson, J.B., 2001. Assembly rules in plant communities. In: Weiher, E., Keddy, P.A. (eds.), Ecological assembly rules: perspectives, advances, retreats. Cambridge University press, Cambridge, UK, 418 pp.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Distribut, N.P.S., 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763-773.

Witt, J., Schroeder, A., Knust, R., Arntz, W.E., 2004. The impact of harbour sludge disposal on benthic macrofauna communities in the Weser estuary. Helgoland Marine Research 58, 117-128.

Witten, I.H., Frank, E., 2000. DataMining: Practical machine learning tools and techniques with Java implementations, 1st edition, Morgan Kaufmann Publishers, San Francisco, USA, 416 pp.

Witthöft-Mühlmann, A., Traunspurger, W., Rothhaupt, K.O., 2007. Combined influence of river discharge and wind on littoral nematode communities of a river mouth area of Lake Constance. Aquatic Ecology 41, 231-242.

Wollan, A.K., Bakkestuen, V., Kauserud, H., Gulden, G., Halvorsen, R., 2008. Modelling and predicting fungal distribution patterns using herbarium data. Journal of Biogeography 35, 2298-2310

Worm, B., Barbier, E.B., Beaumont, N., Duffy, J.E., Folke, C., Halpern, B.S., Jackson, J.B.C., Lotze, H.K., Micheli, F., Palumbi, S.R., Sala, E., Selkoe, K.A., Stachowicz, J.J., Watson, R., 2006. Impacts of biodiversity loss on ocean ecosystem services. Science 314, 787-790.

Yao, J., Teng, N., Poh, H.L., Tan, C.L., 1998. Forecasting and analysis of marketing data using neural networks. Journal of Information Science and Engineering 14, 843-862.

Yodnarasri, S., Montani, S., Tada, K., Shibanuma, S., Yamada, T., 2008. Is there any seasonal variation in marine nematodes within the sediments of the intertidal zone? Marine Pollution Bulletin 57, 149-154.

Yount, J.L., 1956. Factors That Control Species Numbers in Silver Springs, Florida. Limnology and Oceanography 1, 286-295.

Ziegelmeier, E., 1952. Beobachtungen über den Röhrenbau von *Lanice conchilega* (Pallas) im Experiment und am natürlichen Standort. Helgoländer Meeresuntersuch 4, 107-129.

Zühlke, R., 2001. Polychaete tubes create ephemeral community patterns: *Lanice conchilega* (Pallas, 1766) associations studied over six years. Journal of Sea Research 46, 261-272.

ADDENDUM 1

TECHNICAL DESCRIPTION OF ARTIFICIAL NEURAL NETWORKS, GEOSTATISTICS AND MAXIMUM ENTROPY MODELLING

ADDENDUM 1

TECHNICAL DESCRIPTION OF ARTIFICIAL NEURAL NETWORKS, GEOSTATISTICS AND MAXIMUM ENTROPY MODELLING

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) are non-linear mapping structures based on the functioning of the human brain. They have been shown to be highly flexible approximators for any data. ANNs make powerful tools for modelling complex relationships, especially when the underlying data relationships are unknown (Lek and Guégan, 1999).

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. By adjusting the values of the connections (weights) between elements, neural networks can be trained to perform a particular function. Commonly neural networks are adjusted or trained, to assure that a particular input leads to a specific target output. During the learning process, network outputs and targets are compared; networks are adjusted until the output matches the target. Typically, many such input/target pairs are used in this supervised learning to train a network (Demuth and Beale, 1998). In this research the input variables are the environmental variables and the target values are the diversity indices. The relationships between variables in ecology are often very complicated and highly non-linear. Therefore, neural networks are considered to be powerful tools to investigate these relationships. ANNs have indeed the capacity to predict the output variable but the mechanisms that occur within the network are often ignored. Therefore, ANNs are often considered as black boxes (Gevrey *et al.*, 2003). Several methods have been described to unravel these connections.

Architecture of an Artificial neural network

A Simple Neuron

A neuron with a single scalar input is shown in Fig. A1.1. The input p is multiplied by the weight w (this is a so called 'connection' in the neural network), to form the product $w.p$. This forms the scalar output n of the neuron and n is the argument of the transfer function$f$, which produces the output $a.$

The neuron in the right panel of Fig. A1.1 has a bias b. This bias is simply added to the product $w.p$. The bias is like a weight, except that it is multiplied with 1. The input of the transfer function is the sum of the weighted input $w.p$ and the bias b. This sum, called n, is the argument of the transfer function *f.* Here f is a transfer function that takes the argument n and produces the output a.

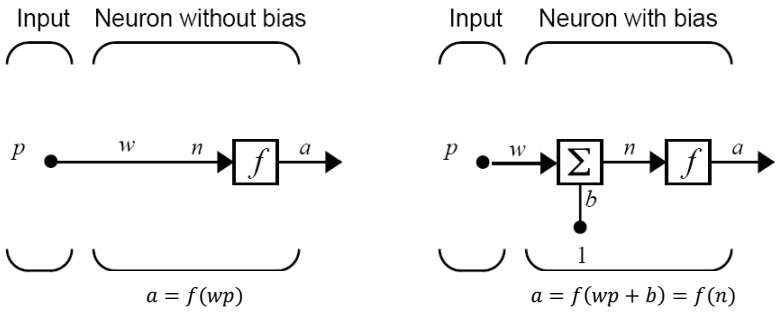


*Fig. A1.1. A neuron with a single scalar input and with and without bias (*from *Demuth and Beale, 1998)*

Note that w and b are both adjustable parameters of the neuron. The central idea of neural networks is that these parameters can be adjusted so that the network exhibits a desired behaviour. Thus, the network can be trained to do a particular job by adjusting the weight or bias parameters, or the network itself (Demuth and Beale, 1998).

Transfer Functions

Three of the most commonly used transfer functions are shown in Fig. A1.2. The hard limit transfer function limits the output of the neuron to either 0, if the net input argument n is less than 0, or 1, if n is greater than or equal to 0. The linear transfer function allows the output to take any value. The log-sigmoid transfer function transforms the input to any value between 0 and 1. It is commonly used in back propagation networks, in part because it is differentiable (Beale *et al.*, 2010).

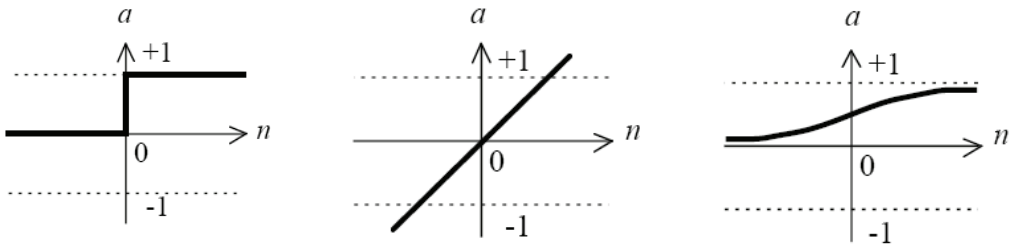


*Fig. A1.2. Three types of transfer functions: Hard limit transfer function, linear transfer function and log-sigmoid transfer function (*from *Beale* et al.*, 2010).*

Neuron with Vector Input

A neuron with an input vector with R elements is shown below (Fig. A1.3).The scalar product of the (single row) matrix $\boldsymbol{W}$ and the vector $\boldsymbol{p}$ results in $\boldsymbol{W}.\boldsymbol{p}$. The neuron has a bias b, which is summed with $\boldsymbol{W}.\boldsymbol{p}$ to form the input n which is the argument of the transfer function f. Two or more of the neurons shown above may be combined in a layer, and a particular network might contain one or more such layers. First consider a single layer of neurons.

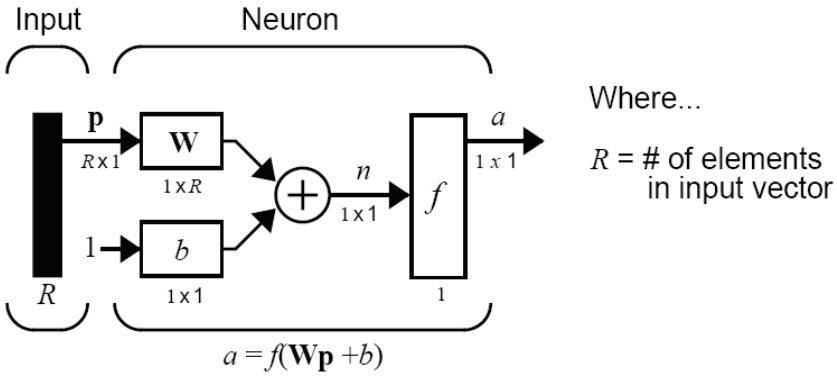


*Fig. A1.3. One neuron with an input vector with R elements (*from *Beale* et al*., 2010).*

One layer of Neurons

A one layer network with R input elements and S neurons is shown in Fig. A1.4. A layer includes the combination of the weights, the multiplication and summing operation, the bias $\boldsymbol{b}$, and the transfer function f. The array of inputs, vector $\boldsymbol{p}$, is not included in a layer.

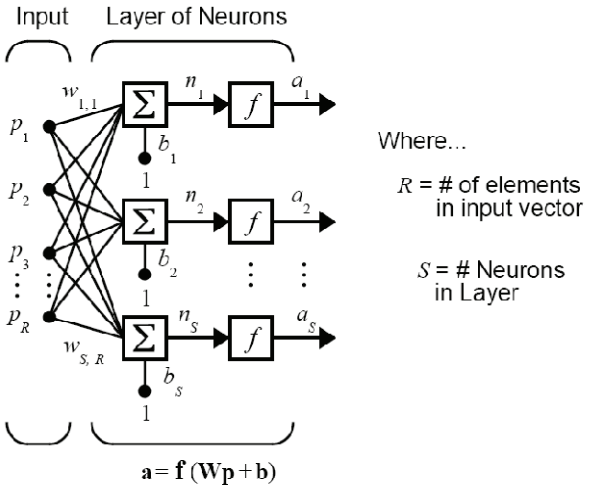


*Fig. A1.4. One layer network with R input elements and S neurons (*from *Demuth and Beale, 1998).*

In this network, each element of the input vector $\boldsymbol{p}$ is connected to each neuron input through the weight matrix $\boldsymbol{W}$, which has the dimensions SxR. The i^{th} neuron has a summation operator that gathers its weighted inputs and bias to form its own scalar output n_i. The n_i form a vector with S elements, called $\boldsymbol{n}$, which is fed through the transfer function and results in the neuron layer output: a vector $\boldsymbol{a}$ with dimension $Sx1$. A layer is not constrained to have the number of its inputs and the same number of neurons.

Multiple Layers of Neurons

A network can have several layers. Each layer has a weight matrix $\boldsymbol{W}$, a bias vector $\boldsymbol{b}$, and an output vector $\boldsymbol{a}$.

The layers of a multilayer network play different roles. A layer that produces the network output is called an output layer. All other layers are called hidden layers. The three layer network shown in Fig. A1.5 has one output layer (layer 3) and two hidden layers (layer 1 and layer 2). Some authors may refer to the inputs as a fourth layer. Weight matrices connected to inputs are called input weights and weight matrices coming from layer outputs are called layer weights (Demuth and Beale, 1998).

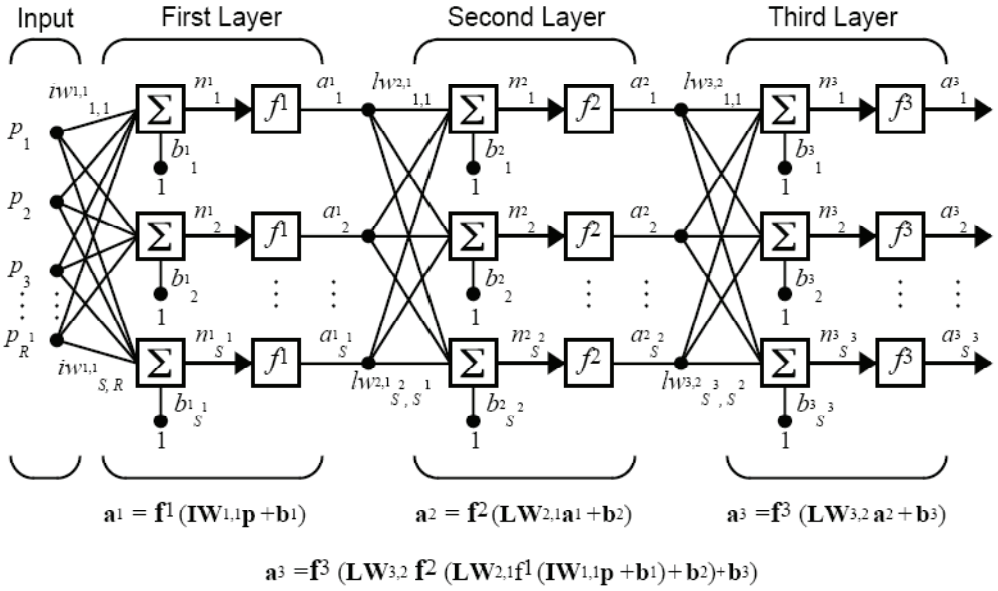


*Fig. A1.5. Three layer neural network (*from *Demuth and Beale, 1998)*

Multiple layer networks are quite powerful. For instance, a network of two layers, where the first layer is sigmoid and the second layer is linear, can be trained to approximate any function (with a finite number of discontinuities) arbitrarily well (Beale *et al.*, 2010).

Training a neural network

The process of optimizing the connection weights is known as 'training' or 'learning'. This is equivalent to the parameter estimation in conventional statistical models (Maier and Dandy, 2000). The initial network weights and biases are randomly initialised. After this initialisation, the network is ready for training. The training process requires a set of examples of proper network behaviour - network inputs p and target outputs t. During training the weights and biases of the network are iteratively adjusted to minimise the network performance function. The default performance function is the mean square error (MSE) which is the average squared error between the network outputs a and the target outputs t (Demuth and Beale, 1998). In many cases there are practical difficulties in

optimizing the error function, because the error surface may be complicated and have many local minima (Cheng and Titterington, 1994).

There are many variations of the back propagation algorithm. The simplest implementation of back propagation learning updates the network weights and biases in the direction in which the performance function decreases most rapidly - the negative of the gradient. One iteration of this algorithm can be written as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - a_k \boldsymbol{g}_k$ where $\boldsymbol{x}_k$ is a vector of current weights and biases, $\boldsymbol{g}_k$ is the current gradient, and a_k is the learning rate (or step size).

There are different training algorithms. In batch mode, the weights and biases of the network are updated only after the entire training set has been applied to the network. The gradients calculated at each training example are added together to determine the change in the weights and biases. For this research we applied the Levenberg-Marquardt training algorithm. The Levenberg-Marquardt algorithm is faster than other algorithms by a factor 10 to 100 (Beale *et al.*, 2010). For most situations, the Levenberg-Marquardt algorithm is recommended. The only drawback of this algorithm is that it may require too much computer memory. If this is a problem, one can make use of a variety of other fast algorithms available (Beale *et al.*, 2010). One iteration of this algorithm can be written as:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\boldsymbol{J}^T\boldsymbol{J} + \mu\boldsymbol{I}]^{-1}\boldsymbol{J}^T\boldsymbol{e} \qquad \text{(Eq. A1.1)}$$

where μ is the learning rate and $\boldsymbol{I}$ the identity matrix, $\boldsymbol{J}$ is the Jacobian matrix which contains first derivatives of the network errors with respect to the weight and biases, and $\boldsymbol{e}$ is a vector of network errors. μ is decreased after each successful step and is increased when an individual step increases the performance function (Karul *et al.*, 2000). Since the weights and biases are initialised before training, training the network several times, may result in different resulting networks.

Improving Generalisation

One of the problems occurring during neural network training is called overfitting. An overly complex neural network may have too many adjustable parameters and the error on the training set is driven to a very small value, but when new data is presented to the network, the error is large. The network has memorised the training examples, but it has not learned to generalise to new situations.

One method for improving network generalisation is to use a network which is just large enough to provide an adequate fit. The larger a network is, the more complex the functions that the network can create. If a small enough network is used, it will not have enough power to overfit the data. Overfitting is linked to the ratio of the number of training samples to the number of connection weights. Amari *et al.* (1997) show that overfitting does not occur if the above ratio exceeds 30. When the above condition is not met, there are clear benefits in using cross-validation.

Early Stopping with validation

A method for improving generalisation is called early stopping. In this technique the available data is divided into three subsets: the training set, the validation set and the test set. The first subset is the training set which is used for computing the gradient and updating the network weights and biases. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error will normally decrease during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error on the validation set will typically begin to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are returned (Fig. A1.6).

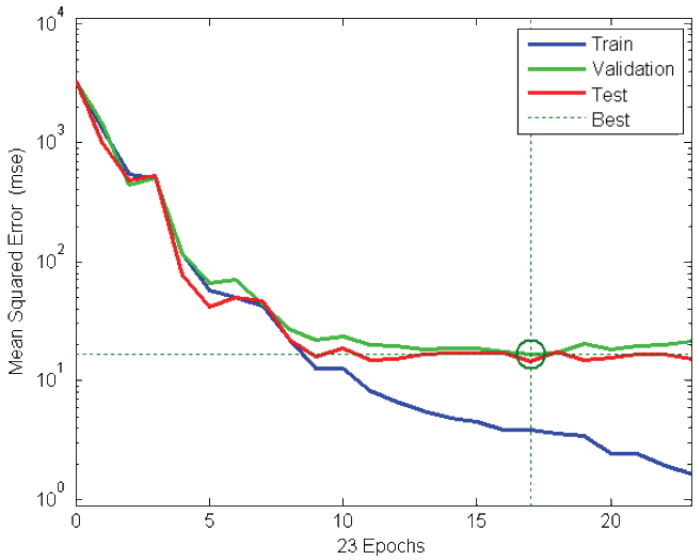


*Fig. A1.6. Evolution of the mean squared error of the training set, the validation set and the test set during training (*from *Beale* et al*., 2010). The training is stopped when the error on the validation set reaches a minimum.*

It is vital that the third set, the test set, is not used during the training process (Maier and Dandy, 2000). The test set error is only used to compare the performance of different models, so it is a truly independent dataset. It is also useful to plot the test set error during the training process. If the error in the test set reaches a minimum at a significantly different iteration number than the validation set error, this may indicate that the model has been overfitted or the two datasets (test and validation) are not representative of the same population (Masters, 1993). A solution to the latter problem is using stratified datasets, this means that the classes of the predicted variable are evenly distributed over the subsets. In some cases also stratification of the independent variables can be useful (Goethals, 2005).

Cross-validation

Cross-validation is a technique that is frequently used in ANN modelling. In cross-validation a validation set is used to assess the performance of the model at various stages of learning. A different test set is needed to assess the generalisation ability of the different models (Maier and Dandy, 2000). Fig. A1.7 shows how data is divided in a 10-fold cross-validation. The model is built by using 80% of the data as training data. Early stopping of the iterations is done by the use of a validation set. The generalisation performance of the model is screened by applying the test set on the resulting model and calculate the error function. The error function is then averaged over the ten models. By repeating this for different neural network architectures, the optimal architecture can be found. This is the architecture with the lowest average error.

train	train	train	train	train	train	train	train	val	test	→model1
test	train	train	train	train	train	train	train	train	val	→model2
val	test	train	train	train	train	train	train	train	train	→model3
train	val	test	train	train	train	train	train	train	train	→model4
train	train	val	test	train	train	train	train	train	train	→model5
train	train	train	val	test	train	train	train	train	train	→model6
train	train	train	train	val	test	train	train	train	train	→model7
train	train	train	train	train	val	test	train	train	train	→model8
train	train	train	train	train	train	val	test	train	train	→model9
train	train	train	train	train	train	train	val	test	train	→model10

Fig. A1.7. Data division in a tenfold cross-validation with validation and independent test set. The eight individual training sets are joined in one training set (val is the validation set, train is the training set).

Preprocessing

Neural network training can be made more efficient if certain preprocessing steps are performed on the network inputs and targets. Below only those optimisation methods used in this research are mentioned.

The network inputs

Generally, different input variables span different ranges. In order to ensure that all variables receive equal attention during the training process, they should be standardised (Maier and Dandy, 2000). The network inputs can be scaled by normalizing the mean and standard deviation of the training set. It normalises the inputs so that they will have zero mean and unity standard deviation.

In some situations the dimension of the input vector is large, and the components of the vectors may be highly correlated (redundant variables). It is useful in this situation to reduce

the dimension of the input vectors. An effective procedure for performing this operation is principal component analysis (PCA). This technique has three effects: it orthogonalises the components of the input vectors (so that they are uncorrelated with each other); it orders the resulting orthogonal components (principal components) so that those with the largest variation come first; and it eliminates those components which contribute the least to the variation in the dataset (Demuth and Beale, 1998).

The targets

As the output of the logistic transfer function is between 0 and 1, the data are generally scaled between 0.1-0.9 or 0.2 and 0.87. If the data are scaled to the extreme limits of the logistic transfer function, the size of the weight updates is extremely small and flat spots in training are likely to occur (Maier and Dandy, 2000). When the transfer function in the output layer is unbounded, as is the case for a linear transfer function, scaling is not strictly required. However, scaling to uniform ranges is still recommended (Masters, 1993). In this research only a linear transfer function was used in the output layer and the targets were scaled by scaling minimum and maximum values to [-1 1].

Input variables contribution methods

Neural networks are generally considered to be a 'black box'. Gevrey *et al.* (2003) gives an overview of six different methods to unravel this 'black box'. These methods can be mainly divided in three groups:

1) Changing the input variables and monitoring the effect on the output (PaD method, Profile and Perturb method). The techniques we applied in Chapter 3 belong to this group;
2) removing input variables and monitor the effect on the output (Stepwise method, Improved stepwise method);
3) reveal the relative importance of the various inputs on the connection weights (Weights method).

GEOSTATISTICS

The term geostatistics refers to the statistical analysis of phenomena which vary in a continuous and spatial way. Classical statistical methods, such as linear regression analysis and analysis of variance, are used intensively for geo-referenced data as well. However, there are limitations associated with these methods due to their underlying assumption of independency of observations. Therefore, there should be no spatial autocorrelation between the observations or between the residuals of the model. However, this condition is rarely met in geographical information. Such data is almost always correlated to some degree in relationship with the distance between observations (Van Meirvenne, 2007).

Geostatistics is related to interpolation methods, but extends beyond simple interpolation problems. It consists of a collection of numerical and mathematical techniques dealing with the characterisation of spatial phenomena. Geostatistical techniques model the uncertainty associated with spatial estimation (Van Meirvenne, 2007).

Regionalised variables

A regionalised variable Z is a variable which exists in a spatial continuous way, thus its value is a function of its spatial position **x**. This variable has a unique value at every location $Z = f(\mathbf{x})$. Assume that Z was observed at a number of locations and a prediction is needed at the unvisited location $\mathbf{x}_0$. In classical statistics such a prediction is modelled as a combination of two components: a mean m plus a random error ε which represents the spatially independent fluctuations around the mean: $Z(\mathbf{x_0}) = m + \varepsilon$ with ε normally distributed and zero mean. In the case of a regionalised variable the hypothesis that the error term is spatially independent is unrealistic. Therefore the error term can be split into two: $\varepsilon'(\mathbf{x})$, an error term which represents the spatial structured component of the spatial variance and ε" a term which can be considered to be spatially uncorrelated:

$$Z(\mathbf{x}_0) = m + \varepsilon'(\mathbf{x}) + \varepsilon" \qquad \text{(Eq. A1.2)}$$

In geostatistics the aim is to characterise $\varepsilon'(\mathbf{x})$ as complete as possible in order to use it for improving the prediction of m (Van Meirvenne, 2007).

Variogram modelling

Variogram modelling is the key to geostatistical modelling. It is a summarizing function and has a strong descriptive and interpretative power about the structure of the spatial variability of a regionalised variable. The variogram is estimated by (Journel and Huijbregts, 1978):

$$\gamma(\mathbf{h}) = \frac{1}{2\mathrm{N}(\mathbf{h})} \sum_{\alpha=1}^{\mathrm{N}(\mathbf{h})} \{\mathrm{z}(\mathbf{x}_\alpha + \mathbf{h}) - \mathrm{z}(\mathbf{x}_\alpha)\}^2 \qquad \text{(Eq. A1.3)}$$

with $\gamma(\mathbf{h})$ the variogram for a distance vector (lag) $\mathbf{h}$ between observations $\mathrm{z}(\mathbf{x}_\alpha)$ and $\mathrm{z}(\mathbf{x}_\alpha + \mathbf{h})$ of the variable at the locations $\mathbf{x}_\alpha$ and $\mathbf{x}_\alpha + \mathbf{h}$, and with $N(\boldsymbol{h})$ the number of pairs separated by $\mathbf{h}$.

A plot of the calculated $\gamma(\mathbf{h})$ values versus the lag $\mathbf{h}$ is called an experimental variogram. To this experimental variogram a theoretical variogram model is fit yielding a continuous function of $\gamma(\mathbf{h})$ versus $\mathbf{h}$. The fitting of a variogram model is an interactive and iterative process (Webster and Oliver, 2007).

Four important characteristics can be derived (Fig. 4.2): the sill, the range, the nugget and the model type. The 'sill' represents the total variance of the variable and is the maximum of the variogram model. The 'range' is the maximal spatial extent of spatial correlation between observations of the variable. At lags larger than the range, the expected difference between observations is maximal (being the sill) and independent of the distance. At

distances smaller than the range a dependency exists between the observations which increases as the observations are situated closer to each other. The extrapolation of the model to lags approaching zero is called the 'nugget variance' and represents sources of random noise such as sampling errors and variability at distances closer than the smallest sampling lag (Van Meirvenne, 2007). The smaller the nugget variance or pure random noise, the smaller ε" (i.e. the part of the error term which is spatially independent (Eq. A1.2), while the difference between sill and nugget relates to the spatially structured error term $\varepsilon'(\mathbf{x})$. Sometimes the nugget equals the sill. This situation is called a pure nugget effect. All variability is purely random and unstructured and has no spatial structure, thus no spatial autocorrelation is observable in the data. This may indicate that the sources of variability are too large and they mask the underlying spatial pattern or the smallest sampling distance is larger than the range.

The theoretical variogram can be composed of nested models or structures. Common models are the spherical model, the exponential model, the Gaussian model and the power model.

In isotropic situations, the regionalised variable shows the same range and variability in all directions. However, in an anisotropic situation the regionalised variable may display different ranges or higher variability in different directions. In that case, directional variograms should be derived (Wackernagel, 2003).

To obtain a reliable representation of the average structure of the spatial variability, it is necessary to have sufficient observation points. About 100 observations are a minimum in isotropic situations (Webster and Oliver, 2007). This represents some limitation to the applicability of geostatistical analysis. The minimum number of pairs $N(\mathbf{h})$ that is required to obtain a reliable estimation of $\gamma(\mathbf{h})$ is 30 to 50 per $\mathbf{h}$ class. For each $\mathbf{h}$ class one point in the experimental variogram is derived.

Kriging

Kriging is a collection of generalised linear regression techniques for minimizing an estimation variance defined from a prior covariance model (Deutsch and Journel, 1992).

A local interpolation can be written as a weighted linear combination of measurements points located within a neighbourhood around $\mathbf{x}_0$:

$$Z^*(\mathbf{x}_0) = \sum_{\alpha=1}^{n(\mathbf{x}_0)} \lambda_\alpha \, Z(\mathbf{x}_\alpha) \qquad \text{(Eq. A1.4)}$$

with λ_α being the weight attributed to the observation $Z(\boldsymbol{x}_\alpha)$ to estimate the value of $Z^*(\mathbf{x}_0)$.

In geostatistics the focus is on the stochastic part of a regionalised variable, the part which is modelled by the variogram: $\varepsilon'(\mathbf{x}) = Z(\mathbf{x}) - m$. Hence the equation is modified to:

$$\varepsilon'(\mathbf{x}_0) = \sum_{\alpha=1}^{n(\mathbf{x}_0)} \lambda_\alpha \, \varepsilon'(\mathbf{x}_\alpha) \qquad \text{(Eq. A1.5)}$$

Yielding the general kriging equation:

$$Z^*(\mathbf{x}_0) - m(\mathbf{x}_0) = \sum_{\alpha=1}^{n(\mathbf{x}_0)} \lambda_\alpha \, [Z(\mathbf{x}_\alpha) - m(\mathbf{x}_\alpha)] \qquad \text{(Eq. A1.6)}$$

Based on different assumptions and objectives, a number of variants of kriging are available. Here ordinary kriging, simple kriging and regression kriging are handled shortly (*modified from* Van Meirvenne, 2007).

Around $\mathbf{x}_0$ an interpolation window, called search neighbourhood , must be specified. Usually this neighbourhood has the shape of a circle in an isotropic situation or an ellipse in anisotropic situations. The search radius is chosen in respect to the range, since observations taken at a larger distance than the range are considered to be uncorrelated with $\mathbf{x}_0$ and will not attribute to its value.

Ordinary kriging (OK)

OK applies to the situation where $m(\mathbf{x})$ is unknown. An underlying assumption for OK is that the mean, although unknown, is locally stationary. Thus, it has a constant value inside the interpolation window.

The algorithm can be written as:

$$Z^*_{OK}(\mathbf{x}_0) = \sum_{\alpha=1}^{n(\mathbf{x}_0)} \lambda_\alpha \, Z(\mathbf{x}_\alpha) \qquad \text{with } \sum_{\alpha=1}^{n(\mathbf{x}_0)} \lambda_\alpha = 1 \qquad \text{(Eq. A1.7)}$$

with $n(\mathbf{x}_0)$ being the total number of observations in the interpolation window around $\mathbf{x}_0$ and λ_α being the interpolation weight attributed to the observation $Z(\mathbf{x}_\alpha)$. The weights λ_α are obtained by solving a set of equations involving knowledge of the variogram and they are chosen in such a way that the prediction error variance is minimised (Webster and Oliver, 2007). Observations closer to $\mathbf{x}_0$ get a higher weight than observations further away.

Simple kriging (SK)

Simple kriging differs from OK in the way $m(\mathbf{x})$ is handled. In SK $m(\mathbf{x})$ is supposed to be known and globally stationary, thus the estimated value of $Z(\mathbf{x})$ equals m. Since m is known, SK works on the residuals R: $R(\mathbf{x}_\alpha) = Z(\mathbf{x}_\alpha) - m$

$$R^*_{SK}(\mathbf{x}_0) = \sum_{\alpha=1}^{n(x_0)} \lambda_\alpha \, R(\mathbf{x}_\alpha) \qquad \text{(Eq. A1.8)}$$

Thus SK starts with subtracting the global mean from all the observations. Then, the residuals are interpolated using the kriging system. And finally the global mean is added back to the estimations of the kriging system. OK is more often used then SK because the global mean m is rarely known (Van Meirvenne, 2007). However, in case of regression kriging (see 2.3.3) the mean is deduced from external secondary information and the mean of the residuals can be considered to equal zero (Hengl *et al.*, 2007).

Regression kriging (RK)

An alternative to ordinary kriging is regression kriging. Predictions by RK involve fundamentally two steps: first the relationship between the dependent variable and the

independent environmental variables at the sampling locations are modelled by a linear regression and this model is then applied to the unseen locations using the environmental variables at this location. Secondly, the residuals of this linear model are subjected to simple kriging with an expected mean of zero (Deutsch and Journel, 1992).

First, the linear model can be written as a linear combination of the environmental variables:

$$\hat{Z}(\mathbf{x}_0) = \sum_{k=0}^{p} \widehat{\beta}_k \, q_k(\mathbf{x}_0) \qquad \text{with } q_0(\boldsymbol{x}_0) \equiv 1 \qquad \text{(Eq. A1.9)}$$

where $q_k(\mathbf{x}_0)$ is the value of the independent variable k at the location $\mathbf{x}_0$, $\widehat{\beta}_k$ is the estimated regression coefficient of the variable k and p is the number of dependent variables.

Secondly, simple kriging with expected mean 0 is used to fit the residuals of the linear model. This results in (Hengl *et al.*, 2007):

$$\hat{Z}(\mathbf{x}_0) = \sum_{k=0}^{p} \widehat{\beta}_k \, q_k(\mathbf{x}_0) + \sum_{\alpha=1}^{n(s_0)} \lambda_\alpha \, e(\mathbf{x}_\alpha) \qquad \text{with } q_0(\boldsymbol{x}_0) \equiv 1 \qquad \text{(Eq. A1.10)}$$

where $e(\mathbf{x}_\alpha)$ is the residual of the linear model at location $\mathbf{x}_\alpha$. The first term describes the linear regression, and the second term describes the simple kriging algorithm of the residuals of the linear regression.

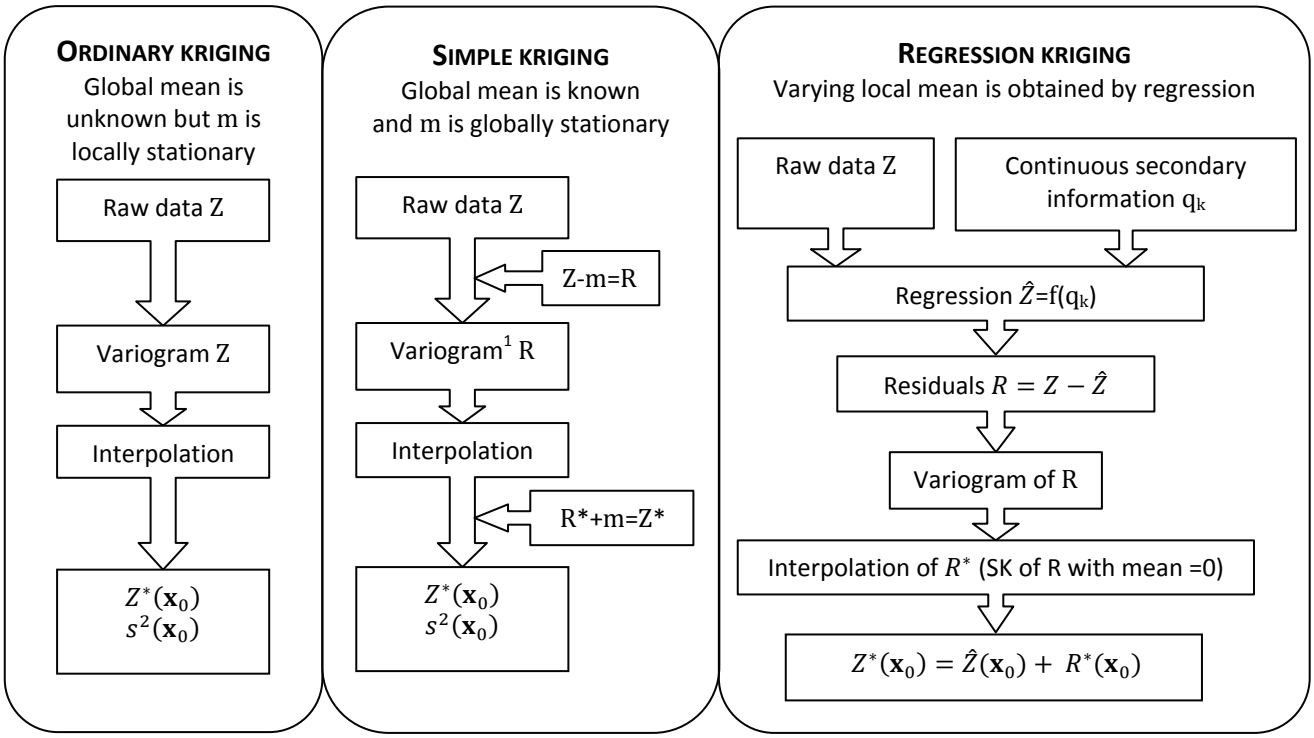


*Fig. A1.8. Schematic representation of the three kriging techniques: ordinary kriging, simple kriging and regression kriging (*modified from *Van Meirvenne, 2007).*

[3] In case of SK the variogram is converted into the covariance of the residuals using the relationship $C(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h})$ with $C(0)$ the covariance at lag 0.

Comparison of the three kriging techniques

Fig. A1.8 shows a schematic overview of the three kriging techniques. As mentioned before, the main difference between these three kriging techniques lies within the assumption concerning the mean of the regionalised variable. In case of OK and SK both a map of the regionalised variable and a map of the variance of this variable can be deduced. In case of regression kriging only a map of the regionalised variable can be inferred, an estimation of the local variance is not possible.

MAXIMUM ENTROPY MODELLING

Maximum entropy modelling is based on the second law of thermodynamics, stating that in systems without outside influences, processes move towards maximum entropy. This theory can be applied in habitat suitability modelling (HSM): in the absence of influences other than those included as constraints in the model, the geographic distribution of a species will evolve to a maximum entropy distribution (Phillips *et al.*, 2006), thus the target is to find the probability distribution of maximum entropy (i.e., the distribution that is most spread out, or closest to uniform), but subject to a set of constraints.

Maxent is a software program for maximum entropy modelling of species' geographic distributions. When Maxent is applied to presence-only species distribution modelling, the pixels of the study area are the space where the Maxent probability distribution is defined. Pixels with known species occurrence records constitute the sample points (Phillips *et al.*, 2006).

Data

Species data

Common statistical methods generally use presence/absence data, however in most cases data can only be considered as presence-only data, e.g. data from inconspicuous species such as nematodes, species with patchy distributions, in case of invasive species which have not yet occupied their potential niche or data from natural history museums (Phillips *et al.*, 2004). In such cases techniques working with presence-only data, such as Maxent, can be very useful. Earlier research pointed out that Maxent is a reliable presence-only modelling technique and it performs well compared to other presence-only modelling techniques (Hernandez *et al.*, 2006; Ortega-Huerta and Peterson, 2008) and it may compete with or even outcompete presence/absence modelling techniques (Elith *et al.*, 2006; Wisz *et al.*, 2008).

Environmental data

The environmental layers should be in raster format all pertaining to the same geographic area (i.e. the study area) which has been partitioned into a grid of pixels with raster cells having the same resolution. The environmental variables or functions thereof are called the 'features'.

Six feature types are used in Maxent:

1) A continuous variable *f* is a 'linear feature' (Fig. A1.9).
2) The square of a continuous variable *f* is a 'quadratic feature'. It models the species' tolerance for variation from its optimal conditions (Fig. A1.9) (Phillips *et al.*, 2006).
3) The product of two continuous environmental variables f_i and f_j is a 'product feature'. Product features incorporate interactions between predictor variables (Phillips *et al.*, 2006).
4) For a continuous environmental variable *f*, a 'threshold feature' is equal to 1 when *f* is above a given threshold, and 0 otherwise (Phillips *et al.*, 2006).
5) The forward hinge feature is 0 if $f(x) \leq h$ and then increases linearly to 1 at the maximum value of $f(x)$. In a similar way, a reverse hinge feature is defined , which is 1 at the minimum value of f. It drops linearly to 0 at $f(x) = h$ and remains 0. Forward and reverse hinge features are collectively referred to as hinge features (Fig. A1.9) (Phillips and Dudík , 2008).
6) Category indicator features are derived from categorical variables. Specifically, if a categorical variable has k categories, it is used to derive k categories of indicator features. For each of the k categories, the corresponding category indicator equals 1 if the variable has the corresponding value and 0 if it has any of the remaining $k - 1$ values (Phillips and Dudík , 2008). The category indicator was not used in this research, since all our environmental variables were continuous variables.

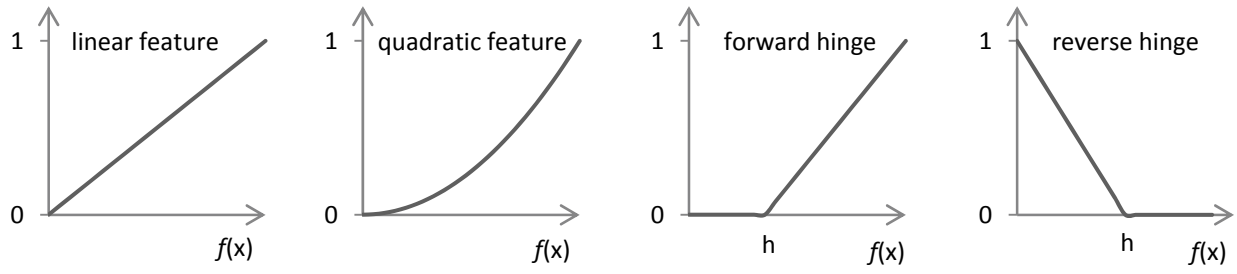


Fig. A1.9. Schematic representation of the features: linear, quadratic, forward and reverse hinge.

Maxent modelling

The basic idea of maximum entropy modelling

A space X represents the geographic region of interest. Typically, X is a set of discrete grid cells. Also a set of points $x_1, \ldots, x_m$ in X are given, each representing a locality where the species has been observed and recorded. In addition, a set of environmental variables defined on X is known. Based on this information, the goal is to estimate the range of the given species (Phillips *et al.*, 2004).

The unknown probability distribution is denoted π over X. The distribution π assigns a non-negative probability $\pi(x)$ to each point x of the area, and these probabilities sum to 1. The approximation of π is also a probability distribution, and is denoted as $\hat{\pi}$. The entropy of the set of probabilities $\hat{\pi}(x)$ is defined as (Phillips *et al.*, 2004):

$$H(\hat{\pi}) = -\sum_{x \in X} \hat{\pi}(x).\log_e\left(\hat{\pi}(x)\right) \qquad \text{(Eq. A1.11)}$$

The entropy is nonnegative and is at most the natural log of the number of elements in X. Entropy is a fundamental concept in information theory. The quantity H has a number of interesting properties:

- H = 0 if and only if all the $\hat{\pi}(x)$ but one are zero, this one having the value unity. Thus only when the outcome is certain, there is no entropy and H vanishes. Otherwise H is positive.
- H reaches the maximum when all $\hat{\pi}(x)$ are equal. This is the most uncertain situation (Shannon, 1948). This is the case when 'a species' shows maximum entropy and has the same likelihood across the whole region.
- Entropy is thus a measure of how much 'choice' is involved in the selection of the event or of how uncertain we are of the outcome (Shannon, 1948). Thus a distribution with higher entropy involves more choices (i.e., it is less constrained) (Phillips *et al.*, 2006). Therefore, the maximum entropy principle can be interpreted as saying that no unfounded constraints should be placed on $\hat{\pi}$, or alternatively, it agrees with everything that is known, but carefully avoids assuming anything that is not known (Jaynes, 1990). At the foundation of the Maxent approach lays the premise that the distribution (or a thresholded version of it) coincides with the biologists' concept of the species' potential distribution (Phillips *et al.*, 2006). However, it ignores the fact that some localities are more likely to have been visited than others. Preferential sampling may bias the model towards areas and environmental conditions that have been better sampled (Phillips *et al.*, 2004).

The problem becomes one of density estimation: given $x_1, \ldots, x_m$ chosen independently from some unknown distribution π, a distribution $\hat{\pi}$ that approximates π has to be constructed. The maximum entropy principle consists of defining the constraints on the unknown probability distribution π in the following way. A set of n environmental variables

$f_1, \ldots, f_n$ on X, the so called 'features', are known for the entire area. The information known about π is characterised by the expectations (averages) of the features under π. Here, each feature f_j assigns a real value $f_j(x)$ to each point x in X. The expectation of the feature f_j under π is defined as $\sum_{x \in X} \pi(x).\, f_j(x)$ and denoted by $\pi[f_j]$ (Phillips *et al.*, 2006).

The feature expectations $\pi[f_j]$ can be approximated using a set of sample points $x_1, \ldots, x_m$ drawn independently from X (with replacement) according to the probability distribution π. The empirical average of f_j is $\frac{1}{m}\sum_{i=1}^{m} f_j(x_i)$, which can be written as $\tilde{\pi}[f_j]$ where $\tilde{\pi}$ is the uniform distribution on the sample points, and is an estimation of $\pi[f_j]$. The probability distribution $\hat{\pi}$ of maximum entropy is subject to the constraint that each feature f_j has the same mean under $\hat{\pi}$ as observed empirically, i.e. $\hat{\pi}[f_j] = \tilde{\pi}[f_j]$ (Eq. A1.12)

for each feature f_j. It can be shown that this characterisation uniquely determines $\hat{\pi}$ (Phillips *et al.*, 2006).

Consider all probability distributions of the form:

$$q_\lambda(x) = \frac{e^{\lambda.f(x)}}{Z_\lambda} \qquad \text{(Eq. A1.13)}$$

where λ is a vector of n real-valued coefficients or feature weights, f denotes the vector of all n features, and Z_λ is a normalizing constant that ensures that q_λ sums to 1. Such distributions are known as Gibbs distributions. It can be shown that the Maxent probability distribution $\hat{\pi}$ is exactly equal to the Gibbs probability distribution q_λ that maximises the likelihood (i.e., the probability) of the m sample points. Equivalently, it minimises the negative log likelihood of the sample points $\tilde{\pi}[-\ln(q_\lambda)]$ which can be written as

$$\log_e Z_\lambda - \frac{1}{m}\sum_{i=1}^{m} \lambda \,.\, f(x_i) \qquad \text{(Eq. A1.14)}$$

and is named the 'log loss' (Phillips *et al.*, 2006).

Maxent can severely overfit training data when the constraints on the output distribution are based on feature expectations as described above, especially if there are a large number of features (Dudík *et al.*, 2004). The problem derives from the fact that the empirical feature means will typically not equal the true means; they will only approximate them. Therefore the means under $\hat{\pi}$ should only be restricted to be close to their empirical values. One way this can be done is to relax the constraint in Eq. A1.12, replacing it with

$$|\hat{\pi}[f_j] - \tilde{\pi}[f_j]| \leq \beta_j \qquad \text{(Eq. A1.15)}$$

for each feature f_j for some constants β_j resulting in a form of ℓ_1-regularisation. The Maxent distribution can now be shown to be the Gibbs distribution that minimises

$$\tilde{\pi}[-\log_e n(q_\lambda)] + \sum_{j=1}^{n} \beta_j |\lambda_j| \qquad \text{(Eq. A1.16)}$$

where the first term is the log loss, while the second term penalises the use of large values for the weights λ_j. Regularisation forces Maxent to focus on the most important features, and ℓ_1-regularisation tends to produce models with few nonzero λ_j values. Such models are

less likely to overfit, because they have fewer parameters. As a general rule, the simplest explanation of a phenomenon is usually best (the principle of parsimony, Occam's Razor) (Phillips *et al.*, 2006).

The Maxent probability distribution is found by starting from the uniform probability distribution, for which $\boldsymbol{\lambda} = (0, \ldots, 0)$, then repeatedly make adjustments to one or more of the weights in such a way that the regularised log loss decreases. It can be shown that the regularised log loss is a convex function of the weights, so no local minima exist and the weights can be adjusted in a way that guarantees convergence to the global minimum (Phillips *et al.*, 2006). Practically, the 'regularisation multiplier' has a default value of one in the software. This value can be changed but a smaller value than the default of 1 will result in a more localised output distribution that is a closer fit to the given presence records; however this can result in overfitting. A larger regularisation multiplier will give more spread out, thus less localised, predictions (Phillips, 2010).

Quality measures of habitat suitability models

A range of techniques for measuring error in presence/absence models exists. Most of these accuracy measures are calculated from a confusion matrix (Fig. A1.10).

		actual	
predicted		+	−
	+	a	b
	−	c	d

Fig. A1.10. A confusion matrix

In the confusion matrix four parameters are available:

1. a represents the number of true positives;
2. b represents the number of false positives. This error is related to Type I error (the error of rejecting a null hypothesis when it is actually true);
3. c represents the number of false negatives. This error is related to Type II error (the error of failing to reject a null hypothesis when it should be rejected);
4. d represents the number of true negatives.

Several parameters can be calculated from this matrix. In Table A1.1 the most relevant ones are shown. These parameters serve different purposes. If the aim is to assess the effectiveness of the model a parameter such as Kappa, which assesses improvement over chance, is appropriate. This can be important because in case of very high or very low prevalence, the overall accuracy may be high. For instance, if the species is found in 95% of the cases, a model which predicts the species over the complete area will classify 95% of the data correctly. Sensitivity measures the proportion of actual positives which are correctly identified, while specificity measures the proportion of negatives which are correctly identified. An optimal prediction would achieve 100% sensitivity and 100% specificity.

However, in reality there is usually a trade-off between the measures. The use of receiver operating characteristic (ROC) plots visualises this trade-off (Fig. A1.11) (Fielding and Bell, 1997).

Parameter	Formula
Prevalence	$\frac{a+c}{N}$
Overall diagnostic power	$\frac{b+d}{N}$
Sensitivity	$\frac{a}{a+c}$
Specificity	$\frac{d}{b+d}$
Kappa	$\frac{\mathrm{a+d}-\frac{(\mathrm{a+c})(\mathrm{a+b})+(\mathrm{b+d})(\mathrm{c+d})}{\mathrm{N}}}{N-\frac{(a+c)(a+b)+(b+d)(c+d)}{N}}$

*Table A1.1. parameters of classification accuracy derived from a confusion matrix (*from *Fielding and Bell, 1997). For description of the parameters a, b, c, d see Fig. A1.10. N is the total number of samples and equals $a+b+c+d$.*

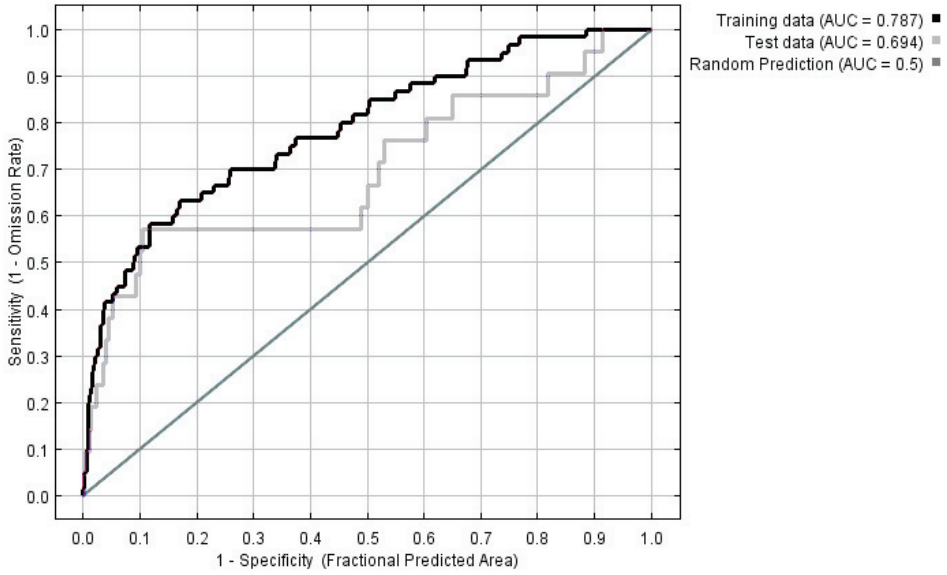


Fig. A1.11. ROC plot for test and training set. The straight line indicates a random prediction.

All of the measures described in Table A1.1 depend on the values assigned to *a*, *b*, *c* and *d* in the confusion matrix. These values are obtained by applying a threshold to the continuous output of the model. One problem with the threshold dependent measures is their failure to use all of the information provided by the model, although dichotomous classifications can be convenient for decision making (Fielding and Bell, 1997). A threshold independent measure is a ROC plot. It is obtained by plotting all sensitivity values (true positive fraction) on the *y* axis against their equivalent (1 - specificity) values (false positive fraction) for all

available thresholds on the *x* axis, as in the example shown in Fig. A1.11. The area under the ROC function (AUC) is an important index because it provides a single parameter of overall accuracy that is not dependent of a particular threshold. The value of the AUC is between 0.5 and 1.0. If the value is 0.5, the model is a random model, while a score of 1.0 indicates a perfect model. A value of 0.8 for the AUC means that for 80% of the time a random selection from the positive group will have a score greater than a random selection from the negative class (Fielding and Bell, 1997). The AUC has thus an intuitive interpretation, namely the probability that a random positive instance and a random negative instance are correctly ordered by the model.

Maxent uses presence-only data and it would appear that ROC curves are inapplicable, since there are no absences, and thus it seems impossible to calculate specificity. However, this problem can be circumvented by considering a different classification problem, namely, distinguishing presence from random, rather than presence from absence. More formally, for each pixel x in the study area, a negative instance x_{random} is defined. Similarly, for each pixel x that is included in the species' true geographic distribution, a positive instance $x_{presence}$is defined. The species distribution model can then make predictions for the pixels corresponding to these instances, without seeing the labels random or presence. Thus, predictions are made for both a sample of positive instances ($x_{presence}$) and a sample of negative instances (x_{random} which are background pixels chosen uniformly at random). Together these are sufficient to define an ROC curve. This process can be interpreted as using pseudo-absences instead of real absences in the ROC analysis. For each ROC analysis, all the test localities for the species are used as presences, and a sample of 10 000 pixels drawn randomly from the study region as random instances (Phillips *et al.*, 2006), called the 'fractional predicted area' (the fraction of the total study area predicted present). AUC values tend to be higher for species with narrow ranges, relative to the study area. This does not necessarily mean that the models are better; instead this behaviour is an artefact of the AUC statistic (Phillips, 2010).

The above treatment differs from the use of ROC analysis on presence/absence data in one important respect: with presence-only data, the maximum achievable AUC is less than 1 (Wiley *et al.*, 2003). If the species' distribution covers a fraction a of the pixels, then the maximum achievable AUC is exactly $1 - a/2$. Unfortunately, the value of a is most of the time not known, so it is impossible to say how close to optimal a given AUC value is. Random prediction still corresponds to an AUC of 0.5.

Feature contribution

While the Maxent model is trained, each step of the Maxent algorithm increases the gain of the model by modifying the coefficient for each feature. The program registers for each environmental variable(s) the increase in gain. At the end of the training process this is converted to percentages. These contribution values are heuristically determined: thus they depend on the path that the Maxent algorithm followed to find the optimal solution, and a

different algorithm could get to the same solution through a different path, thus resulting in different contribution values. In addition, when the environmental variables are highly correlated, the percent contributions should be interpreted with caution. To get alternate estimates of which variables are most important in the model, a jackknife test can be run. Several models are created: each variable is excluded in turn, and a model is created with the remaining variables and alternatively, a model is created using each variable in isolation. The contribution of each variable to the original model can be monitored in this way (Phillips, 2010).

Relationships with other modelling approaches

The Maxent modelling technique is an 'unconditional' maximum entropy model, it uses only presence data. 'Conditional' models require both presence and absence data. Maxent has strong similarities to some existing methods for modelling species distributions, in particular, generalised linear models (GLMs) and generalised additive models (GAMs) (Phillips *et al.*, 2006). When GLM/GAMs are used to model probability of occurrence, absence data are required. When applied to presence-only data, background pixels must be used instead of true absences (Ferrier and Watson, 1996). However, the interpretation of the result is less clear-cut. It must be interpreted as a relative index of environmental suitability. In contrast, Maxent generates a probability distribution over the pixels in the study region, and in no sense are pixels without species records interpreted as absences. In addition, Maxent is a generative approach, whereas GLM/GAMs are discriminative. The latter approach is generally preferred. However, generative methods may give better predictions when the amount of training data is small (Ng and Jordan, 2001).

ADDENDUM 2

SUBSETS IN THE UGENT DATABASE

&

SAMPLING TECHNIQUES OF THE MANUELA AND UGENT DATA

ADDENDUM 2

SUBSETS IN THE UGENT DATABASE & SAMPLING TECHNIQUES OF THE MANUELA AND UGENT DATA

The UGent database consists of historical data collected at the Marine Biology Section at Ghent University. It consists of 22 subsets with data collected in the framework of PhD research, BSc dissertations and funded research projects. In the framework of this research hardcopy data were scanned and compiled to a database together with the digital data. An overview of these 22 subsets is given in table A2.1.

Building a consistent database from hard copy data and digital sources is often labour intensive. Firstly, the hardcopy data need to be scanned and after applying optical character recognition the data needs to be corrected and double checked for potential errors. This data is then transformed in a consistent database. This involves different quality checks. The list below shows a brief overview concerning the UGent database:

- Start with designing a hierarchical database which can cope with the different types of data collected for different research topics:
    - Tables: metadata, stations, samples, slices, species data, biometrics, abiotic data, species list;
    - Add the data in a consistent way: Check referential integrity in the database: each species observation should be linked to one slice, each slice should be linked to just one sample, each sample belongs to only one station, each station belongs to one researcher in the database;
    - Check for duplicate entries: identify overlap between data from PhD, MSc thesis and projects;
    - Add meta data about the data supplier and the topic of the research;
    - Add information about each field of the table in the table description;
- Check station data
    - Identify the coordinate system of the supplied data (e.g. ED50, WGS85);
    - Identify faulty coordinates e.g. N>90°, E>180°, swapping of E and N;
    - Identify missing data e.g. 0°N, 0°E is most probably incorrect;
    - In case of marine data: coordinates should be situated at sea;
    - Spatial references which are plausible, but incorrectly described in the original source, are often hard to trace (e.g. in the UGent database there was one station which always represented an outlier in the analyses. Eventually, while double checking the original literature source, it was clear that the coordinates in the table and the coordinates on the map represented different coordinates. The map showed the correct position of the station. By replacing the coordinates by those represented in the map, the data were no longer outliers.)

- Add sample data:
    - Date of sampling (if known: hour of sampling);
    - Sample gear, sample size, subsample size;
    - Identify replicate samples;
    - Add total density of the core sample;
- Add slice data:
    - In order to construct a consistent database identify the whole core as a slice if no separate slices where taken;
    - Add total density of the slices;
- Add species data
    - Check validity of nomenclatural and taxonomic classification
        - check validity of names
        - identify synonyms (check literature sources and taxonomical websites such as NeMys, ERMS, WoRMS, …)
        - identify misspellings (find similar names and identify the correct spelling)
        - standardise records with an uncertainty level (e.g. aff., n.sp.1, …) and identify the closest most specific taxonomical level. However, keep the original data. It often holds additional information.
        - identify misidentifications (hard, if not impossible, to trace)
    - Check species numbers
        - Include as much as possible the original data:
            - Species counts are preferred to abundances if the total abundance in the sample is known;
            - Abundances are preferred to relative abundances;
            - Relative abundances are preferred to presence/absence.
        - Number of counts should never be larger than the total species density in the sample;
        - The sum of the abundances per species should never be larger than the total density of the sample;
        - The relative abundances should never add up to more than 100%;
- Add environmental data
    - Add information on how the environmental data was measured. Check if the different methods result in different accuracies (often hard to trace);
    - Specify matrix where the environmental data is measured: water, sediment (e.g. chl *a*);
    - Add a unit to each variable;
    - Standardize units as much as possible (e.g. µg/l ~ g/m³);
    - Describe units as specifically as possible (e.g. weight% or volume% for sediment fractions);

    - Identify improbable or impossible values of the environmental variables (e.g. sediment fractions add up to more than 100%; sort from largest to smallest values and identify extreme outliers);
    - Write percentages in the same way e.g. 1%=0.01, the value may be entered as 1 or as 0.01 in the database. Specify in the description of the table;
    - Add environmental data extracted from maps, but give them a different label;
- Add additional data to the database:
    - Feeding type for each species;
    - Taxonomical tree: family, order, class;
    - Information known from some data sources but not specified in others (such as sampling design, species counts, total densities, …) was traced back in original descriptions as much as possible.
- When in doubt always contact the data supplier and ask feedback!

Data provider	Data source	Description	Time frame
Ann Vanreusel	Hardcopy	PhD: Ecology of the free-living marine nematodes from the Voordelta (Southern Bight of the North Sea).	1988-1989
Carlo Heip	Digital	Nematodes from the NSBS. Data collected in the framework of the 1986 North Sea Benthos Survey, an activity of the Benthos Ecology Working Group of ICES.	1986
Chen Guotong	Hardcopy	BSc Thesis: Study of the meiobenthos in the Southern Bight of the North Sea and its use in ecological monitoring.	1987
Gonda Bisschop	Hardcopy	BSc Thesis: Study of the nematode fauna of the North Sea and the mouth of the Western Scheldt estuary.	1976-1977
Jan Vanaverbeke	Digital	Nematodes from station 330: structural and functional biodiversity on the Belgian Continental Shelf (TROPHOS project).	1999
Jan Vanaverbeke	Digital	Study of the ecological effects of sand extraction on the Kwintebank: evaluation of past extraction effects. (SPEEK project).	2003-2006
Jian Li	Hardcopy	The temporal variability of free-living nematodes in a brackish tidal flat of the Western Scheldt with emphasis on the use of an ecological model.	1992-1993
Jyotsna Sharma	Hardcopy	PhD: A study of the nematode fauna of three estuaries in the Netherlands.	1985
Maaike Steyaert	Digital	Meiobenthos at the stations 115, 702, 790 on the Belgian Continental Shelf (IMPULS-project).	1994
Maaike Steyaert	Digital	Spatial heterogeneity of nematodes on an intertidal flat in the Western Scheldt Estuary. (Ecoflat project).	1997
Maaike Steyaert	Digital	Nematode data of a station at the German Bight.	2002
Maaike Steyaert	Digital	Meiobenthos station 115bis - benthic-pelagic coupling (TROPHOS/PODO-I work-database I 23/01/2004):	2004
Maarten Raes	Digital	PhD: An ecological and taxonomical study of the free-living marine nematodes associated with cold-water and tropical coral structures.	2006

Data provider	Data source	Description	Time frame
Magda Vincx	Hardcopy	PhD: Free-living marine nematodes from the Southern Bight of the North Sea.	1987
Matthew Lammertyn	Digital	BSc Thesis: Live on and around ship wrecks on the Belgian Continental Shelf (BEWREMABI project).	2005
Preben Jensen	Hardcopy	BSc Thesis: Nematode fauna on the silty and sandy sea floor in the southern North Sea (I.C.W.B.).	1974
Regine Vandenberghe	Hardcopy	BSc Thesis: The meiobenthos in a dumping site at the Southern Bight of the North Sea, with emphasis on the free-living marine nematodes.	1987
Sandra Vanhove	Digital	Data collected in the framework of the ANDEEP 2 project.	2002
Sandra Vanhove	Digital	Data collected in the framework of the LAMPOS-ANDEEP-project.	2002
Sandra Vanhove	Digital	Data collected in the framework of the EPOS-leg3-project.	1989
Saskia Van Gaever	Digital	PhD: Biodiversity, distribution patterns and trophic position of meiobenthos associated with reduced environments at continental margins.	2008
Tom Gheskiere	Digital	PhD: Nematode assemblages from European sandy beaches diversity, zonation patterns and tourist impacts.	2005
Zhang Derong	Hardcopy	BSc Thesis: Evaluation of the meiofauna and nematode community at a TiO_2-dumping site after recovery.	1995

Table A2.1 Data providers, source, description and time frame of the 22 subsets in the UGent database.

Data provider	Sample device	Sample area (cm²)
Magda Vincx (Chen Guotong)	Box corer	10
Maaike Steyaert & Li Jian	Cores	10
Magda Vincx (Jyotsna Sharma)	Divers and Reineck box corer	10
Jan Vanaverbeke & Magda Vincx	Reineck box corer	10
Ann Vanreusel	Van Veen or box corer	10
Michaela Schratzberger	Bowers & Connelly multi corer	23.76
John Lambshead	Core	3.8
Tim Ferrero	50 ml syringe corer	5.31
Michaela Schratzberger	Bulk	unknown
Magda Vincx	Van Veen, box corer or diver	7 to 18.4
Andrea McEvoy	Craib cores & Day grab	unknown

Table A2.2. Sampling techniques of the MANUELA data used in Chapter 2.

Data provider	Sample device	Sample area (cm²)
Preben Jensen - BSc Thesis	Van Veen	11
Jan Vanaverbeke - Trophos	Reineck	10
Maaike Steyaert - Trophos	Reineck	10
Magda Vincx – PhD	Diver, Reineck or Van Veen	7 to 18.4 cm²
Jan Vanaverbeke - Speek	Reineck	10
Chen Guotong - BSc Thesis	Box corer	10

Table A2.3. Sampling techniques of the UGent data used in Chapter 3.

Data provider	Sample device	Sample area (cm²)
Ann Vanreusel (PhD)	Box corer	10
Ann Vanreusel (PhD)	Van Veen	10
Preben Jensen (BSc Thesis)	Van Veen	11
Jan Vanaverbeke (Trophos)	Reineck	10
Maaike Steyaert (Impuls)	Box corer	10
Maaike Steyaert (Trophos)	Reineck	10
Magda Vincx (PhD)	Box corer	10
Magda Vincx (PhD)	Diver	10
Magda Vincx (PhD)	Reineck	7 to 18.4 cm²
Magda Vincx (PhD)	Van Veen	7 to 18.4 cm²
Jan Vanaverbeke (Speek)	Reineck	10
Chen Guotong (BSc Thesis)	Box corer	10

Table A2.4. Sampling techniques of the UGent data used in Chapter 4.

Data provider	Sample device	Sample area (cm²)
Ann Vanreusel (PhD)	Box corer	10
Ann Vanreusel (PhD)	Van Veen	10
Regine Vandenberghe	Box corer	10
Jan Vanaverbeke (Trophos)	Reineck	10
Maaike Steyaert (Impuls)	Box corer	10
Maaike Steyaert (Trophos)	Reineck	10
Magda Vincx (PhD)	Diver	10
Magda Vincx (PhD)	Reineck	7 to 18.4 cm²
Magda Vincx (PhD)	Van Veen	7 to 18.4 cm²
Magda Vincx (PhD)	Box corer	10
Sharma Jyotsna (PhD)	Reineck	10
Jan Vanaverbeke (Speek)	Reineck	10
Zhang Derong (BSc Thesis)	Box corer	10
Chen Guotong (BSc Thesis)	Box corer	10

Table A2.4. Sampling techniques of the UGent data used in Chapter 5 and 6.

ADDENDUM 3

Nematode habitat suitability models

ADDENDUM 3

NEMATODE HABITAT SUITABILITY MODELS

This addendum supplies an overview of the final species models. Only those species models performing better than random are given (Chapter 5). This addendum holds three sections with species models passing the preferential sampling and cross-validation test (Chapter 5) according to the minimum distance between test and training set:

- no minimum distance between test and training set (designated by '0 km' in page header);
- a minimum distance of 5 km between test and training set (designated by '5 km' in page header);
- a minimum distance of 10 km between test and training set (designated by '10 km' in page header).

This resulted in respectively 111, 76 and 63 species models (Chapter 5). Since overfitting is still an issue, these models were further optimised by a backward and forward variable selection based on a fivefold cross-validation. On average the number of variables selected in the model decreases with increasing distance between the datasets.

Each of the three sections is subdivided in two parts:

- the first part holds a table with the variable contributions. The variable contribution is an estimate of the relative contribution (%) of the environmental variable to the Maxent model. These variable contributions should be interpreted with caution when the predictor variables are correlated.
- the second part shows the resulting habitat suitability maps. Although the models perform better than random, preferential sampling may still have an influence on the final model. In practice, the probability distribution should be interpreted more conservatively as a relative index of environmental suitability, where higher values represent a prediction of better conditions for the species (Phillips *et al.*, 2006) and the maps may represent an underestimation of the true geographical range of the species (Raes and ter Steege, 2007).

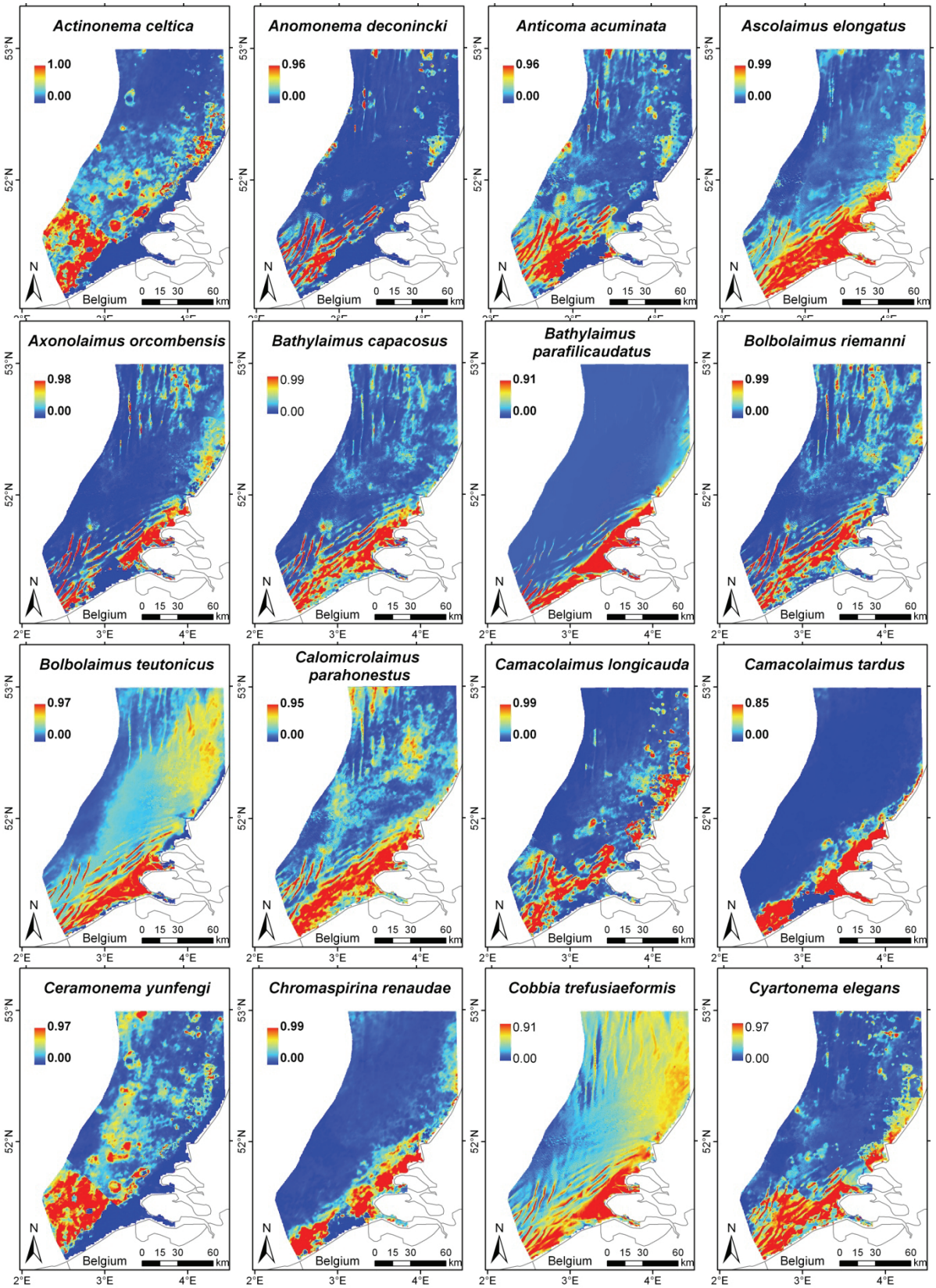
The response curves show the relation between the variable and the model output. In total there are 1017 response curves. These curves can be found on the DVD annexed to this thesis. Each html-file of a species shows the corresponding output of the Maxent model with the response curves, the AUC and the data points used in the analyses.

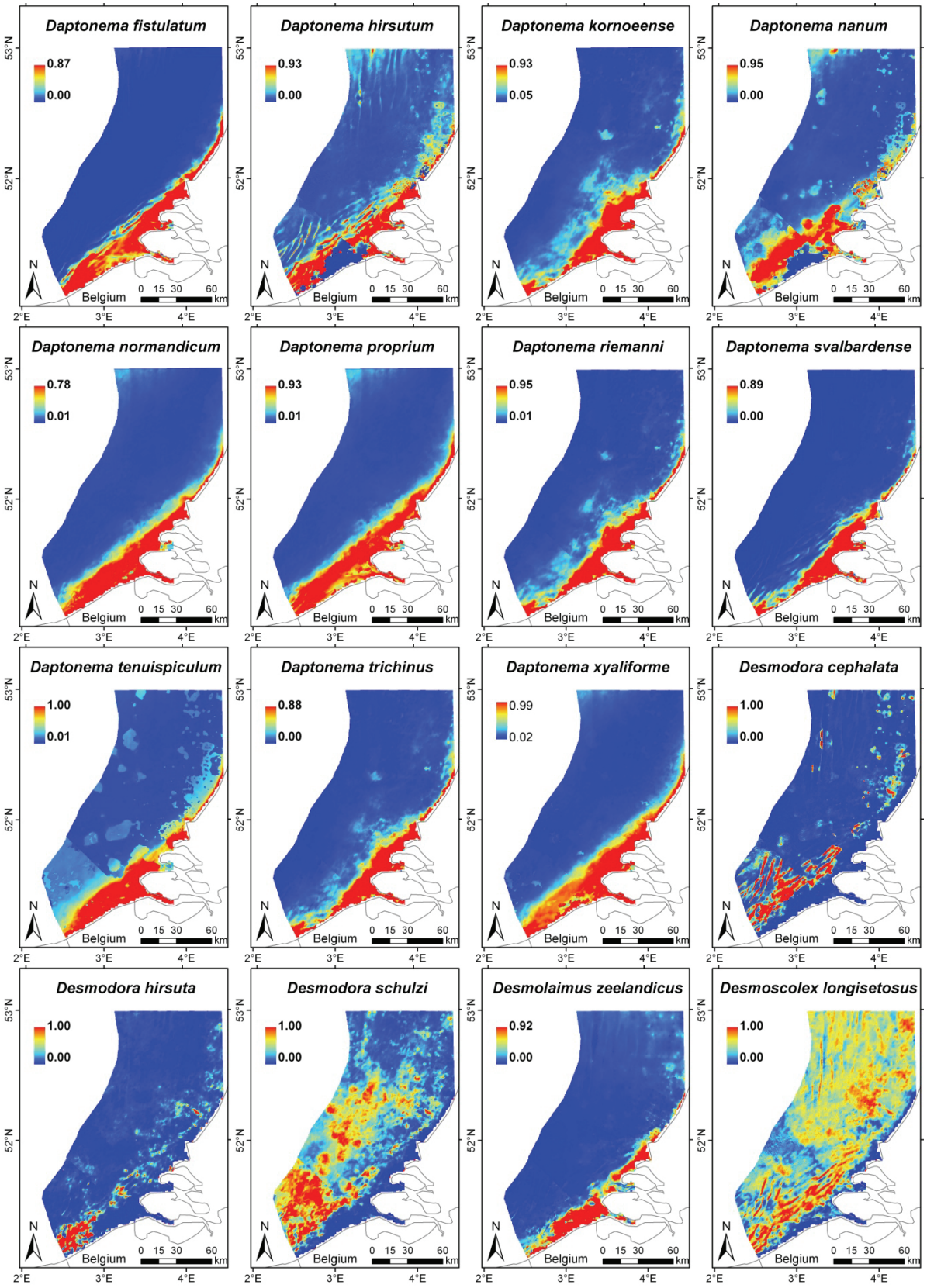
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Actinonema celtica	0.6	0	23.7	17.2	0	48.8	0	0	9.6
Anomonema deconincki	17.8	0	0	14.5	25.2	36.2	4.9	0	1.3
Anticoma acuminata	18.3	2.7	0	10.9	19.7	39.8	4.9	0	3.6
Ascolaimus elongatus	7.1	0	0	2.7	9.2	24.2	56.8	0	0
Axonolaimus orcombensis	17.8	3.6	5.1	2.3	55.3	15.8	0	0	0
Bathylaimus capacosus	17.3	3.3	10	4.3	57.4	0	5.8	1.8	0
Bathylaimus parafilicaudatus	3.5	0	0	0.8	95.7	0	0	0	0
Bolbolaimus riemanni	21.2	3.8	9	6.8	56.7	0	2.4	0	0
Bolbolaimus teutonicus	30.8	0	0	0	69.2	0	0	0	0
Calomicrolaimus parahonestus	10.4	4.2	7.1	0	13.8	0	49.4	9.1	5.9
Camacolaimus longicauda	0	0	23.2	10.4	14.1	48.8	3.5	0	0
Camacolaimus tardus	0	0	0	0	0	10.2	0	89.8	0
Ceramonema yunfengi	27.4	0	10.1	0	0	50.9	5.3	0	6.2
Chromaspirina renaudae	0	0	36	0	9.9	7.3	0	46.8	0
Cobbia trefusiaeformis	9.5	0	0	0	87.9	2.6	0	0	0
Cyartonema elegans	9.3	12	4.8	0	20.9	51.1	2	0	0
Daptonema fistulatum	0	0.5	0	0	4	0	91.4	0	4.1
Daptonema hirsutum	2.7	0	0	0	11.3	24.2	0	56.8	5
Daptonema kornoeense	0.9	99.1	0	0	0	0	0	0	0
Daptonema nanum	11.2	7	0	0	0	45.7	36.1	0	0
Daptonema normandicum	0	0	0	0	0	0	96.3	3.7	0
Daptonema proprium	0	3.7	0	0	1.5	0	93	1.7	0
Daptonema riemanni	0	91.9	0	0	0	8.1	0	0	0
Daptonema svalbardense	5.1	81.2	0.5	0.1	12.9	0	0	0	0.1
Daptonema tenuispiculum	0	0	0.4	2.3	0	20.6	76.7	0	0
Daptonema trichinus	0	98.4	0	1.6	0	0	0	0	0
Daptonema xyaliforme	0	10.5	0.7	0	0	4.6	84.2	0	0
Desmodora cephalata	11.2	0	18	14.4	17.2	39.2	0	0	0

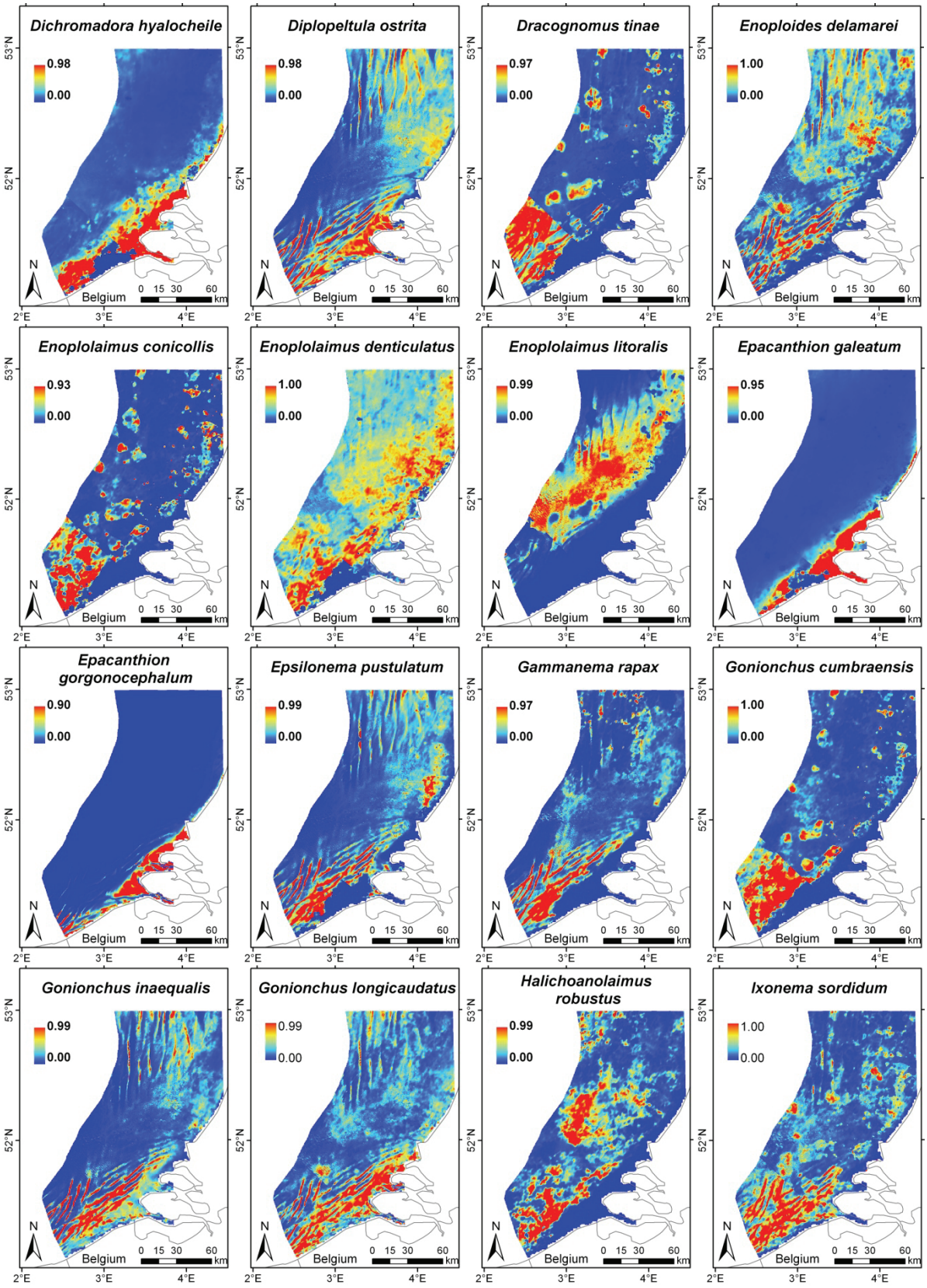
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Desmodora hirsuta	0	4.5	19.5	16.5	1.9	0.4	0.4	47.4	9.3
Desmodora schulzi	5.6	12.7	15.4	28.4	0	27.9	0	10	0
Desmolaimus zeelandicus	0	0	0	11	0	0	86.9	0	2
Desmoscolex longisetosus	28.7	0	56.5	4.4	10.4	0	0	0	0
Dichromadora hyalocheile	0	0	0	2.6	0	12.9	0	84.6	0
Diplopeltula ostrita	26	0	0	7.2	66.8	0	0	0	0
Dracognomus tinae	21.4	0	0	0	13.6	51.7	6.1	0	7.2
Enoploides delamarei	45.1	0	13.4	12.9	28.5	0	0	0	0
Enoplolaimus conicollis	7.8	8.5	18.8	0	0	49.1	9.9	0	5.8
Enoplolaimus denticulatus	0.1	2.6	64.2	8.8	6.5	17.9	0	0	0
Enoplolaimus litoralis	10.3	1.5	0	0	9.6	18.8	59.8	0	0
Epacanthion galeatum	0	0	0	0	0	18.6	76.4	0	4.9
Epacanthion gorgonocephalum	2.6	46.2	0.3	0.1	50.3	0	0	0	0.4
Epsilonema pustulatum	24.6	0	0	7.7	51.4	8.5	0.2	1.6	5.9
Gammanema rapax	22.4	20.3	0	2.5	31.3	0	10.7	0	12.8
Gonionchus cumbraensis	11.3	10	10.1	0	0	54.9	7.6	1.5	4.6
Gonionchus inaequalis	19.9	1	0	11.1	59.7	0	8.2	0	0
Gonionchus longicaudatus	18.8	1.5	8.8	6.3	56.1	0	6.2	0	2.3
Halichoanolaimus robustus	16.5	12.5	31.3	14.4	0	0	14.2	0	11.1
Ixonema sordidum	15.8	10.8	8.4	7.1	13.2	33.9	3.1	6	1.7
Leptolaimus venustus	29.1	0	25.7	0	23.6	0	7	0	14.6
Manunema annulatum	11.1	8.8	8	0.6	21.2	46.2	2.1	0	1.9
Mesacanthion diplechma	8.3	36	0	0.1	47.1	0	0	0	8.5
Metadesmolaimus aduncus	7.5	3.9	0	3.2	85.4	0	0	0	0
Metadesmolaimus gelana	29.3	5.5	6	2.3	56.9	0	0	0	0
Metadesmolaimus varians	0	0	0	0	90.7	0	9.3	0	0
Metalinhomoeus biformis	0	99.9	0	0	0	0	0	0	0.1
Metepsilonema emersum	20.7	0	0	13.2	47	7.5	3.1	0	8.4

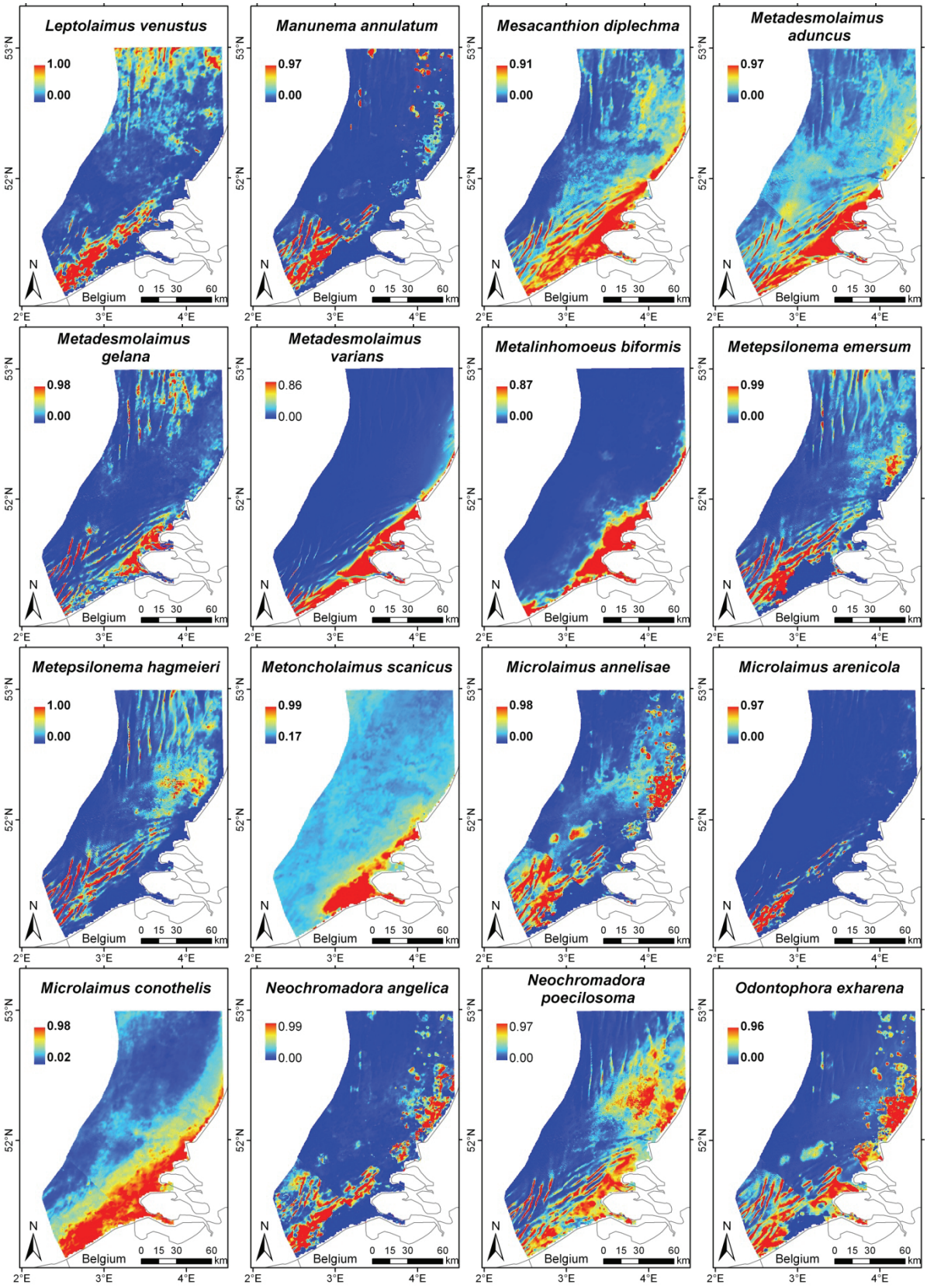
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Metepsilonema hagmeieri	31.2	0	0	11.3	52.3	0	0	0	5.2
Metoncholaimus scanicus	0	0	0	0	0	0.2	0	0	99.8
Microlaimus annelisae	0	17.6	0	7.8	15.8	44	14.7	0	0
Microlaimus arenicola	24.5	0	23.2	0	32.8	0.5	6.6	0	12.3
Microlaimus conothelis	4.9	0	0	2.8	0	0	89.4	2.9	0
Neochromadora angelica	0	6.5	27.4	4.8	9.7	44.1	5.4	2.1	0
Neochromadora poecilosoma	0	0	0	12.8	73.8	0	0	13.5	0
Odontophora exharena	0	0	0	7	28.1	57.4	7.5	0	0
Odontophora ornata	0	0	0	0	0	33.5	0	66.5	0
Odontophora rectangula	8.2	19.4	0	0	18	0	54.4	0	0
Oncholaimellus calvadosicus	0	0	0	0	94.3	5.7	0	0	0
Oxyonchus dentatus	0	6.6	15.5	9.6	14.4	49.9	0	4	0
Paracyatholaimoides labiosetosus	10.8	6.9	17	0	0	51	10.8	0	3.5
Paracyatholaimoides multispiralis	29.4	0	17.2	13.9	39.5	0	0	0	0
Pareurystomina acuminata	37.8	0	18.7	0	23.7	0	7.9	0	11.9
Pomponema elegans	0	0	31	6.4	25	37.5	0	0	0
Pomponema multipapillatum	13.7	0	0	13.7	14.4	42.5	12.8	0	2.9
Pomponema tesselatum	8.5	9.2	23.4	1	3.7	6.2	0.4	47.6	0
Prochromadorella ditlevseni	11	6.3	7.4	0	12.2	33.2	28.9	1.1	0
Prochromadorella longicaudata	0	98	2	0	0	0	0	0	0
Prochromadorella septempapillata	12	9.3	29.8	6.2	8.1	0	23	2.6	9.1
Pseudonchus decempapillatus	29.4	0	29.4	0	41.2	0	0	0	0
Ptycholaimellus ponticus	3.9	2.6	27.7	0	3	0.1	1.6	47.8	13.3
Ptycholaimellus vincxae	15	0	0	0	25.9	43.6	11.5	0	3.9
Rhabdodemania imer	0	13.5	7.1	0	27.4	40.6	0	11.4	0
Rhynchonema ceramotos	3.3	12	0	0	0	63.5	10.4	2.9	7.9
Rhynchonema lyngei	15.2	5	0	15.1	54.5	0	0	10.2	0
Rhynchonema moorea	2.8	15.7	35.8	0	0	45.7	0	0	0

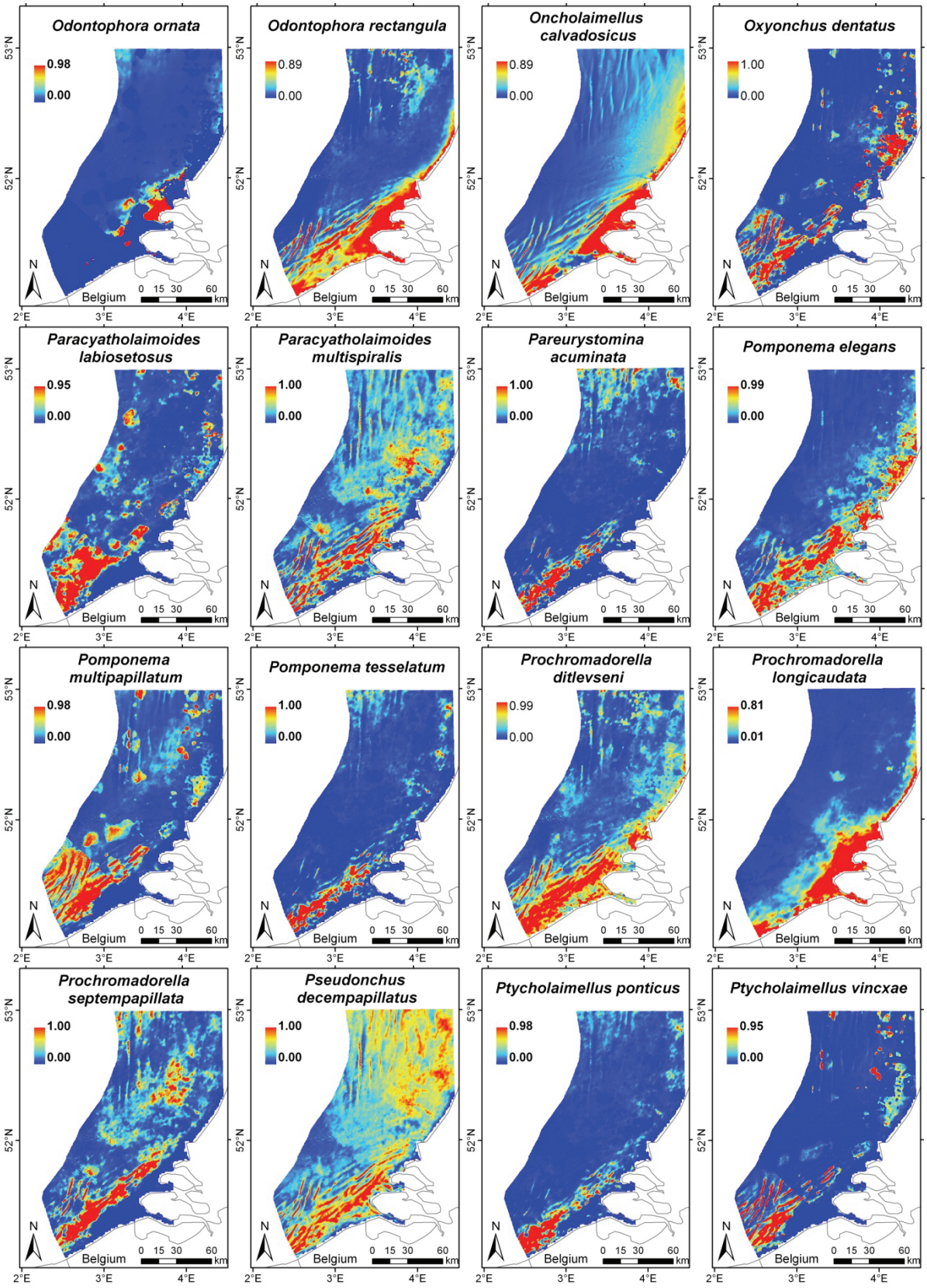
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Rhynchonema quemer	0	10.9	11.8	14.4	15.2	38.8	8.9	0	0
Rhynchonema scutatum	14.8	0	13	0	10	54	4.5	0	3.7
Richtersia inaequalis	8.8	0	3.4	4.4	0	0	53.8	10.6	19
Sabatieria elongata	35	8.2	56.9	0	0	0	0	0	0
Sabatieria punctata	0	0	0	0	0	0	0	96.8	3.2
Setosabatieria hilarula	0	0	4.8	7.9	0	0	87.4	0	0
Sigmophoranema rufum	16.2	0	0	12.4	58.3	0	0	13.1	0
Siphonolaimus ewensis	2.2	0	0	0	55.1	38.6	0	0	4.1
Sphaerolaimus gracilis	0	0	0	0	0	0	0	100	0
Spilophorella paradoxa	14.7	0	12.2	0	0	47.2	15.2	0	10.8
Stephanolaimus flevensis	5	0	0	0	95	0	0	0	0
Stephanolaimus gandavensis	22.5	0	0	37.1	31.1	0	0	9.3	0
Tarvaia angusta	22.8	0	0	18.9	38.5	19.8	0	0	0
Terschellingia longicaudata	0	0	0	0	0	100	0	0	0
Theristus acer	0	0	0	0	0	0	100	0	0
Theristus balticus	35	0	0	9.2	31.8	0	21.4	0	2.5
Theristus bastiani	9.6	0	33.8	0	11.5	31.4	0	0	13.6
Theristus denticulatus	22.2	0	15.2	0	16.6	44.3	1.7	0	0
Theristus longicollis	6.4	8.5	28.3	0	2.6	2.7	0	51.5	0
Theristus maior	11.3	6.3	20	2.8	16.9	38.4	2.4	0	2
Theristus pertenuis	2	0	0	0	5.9	0	83.8	0	8.4
Theristus profundus	3.5	0	21.6	15.3	2.6	0	1.1	44.1	11.8
Trefusia litoralis	0	0	0	0	47.2	7.1	0	45.6	0
Trefusia longicaudata	0	0	11.5	0	4.6	0	0.1	59.9	23.9
Viscosia langrunensis	0	54.6	0	1	34.7	9.7	0	0	0
Viscosia separabilis	0	100	0	0	0	0	0	0	0
Viscosia viscosa	0	0	0	0	0	0	100	0	0

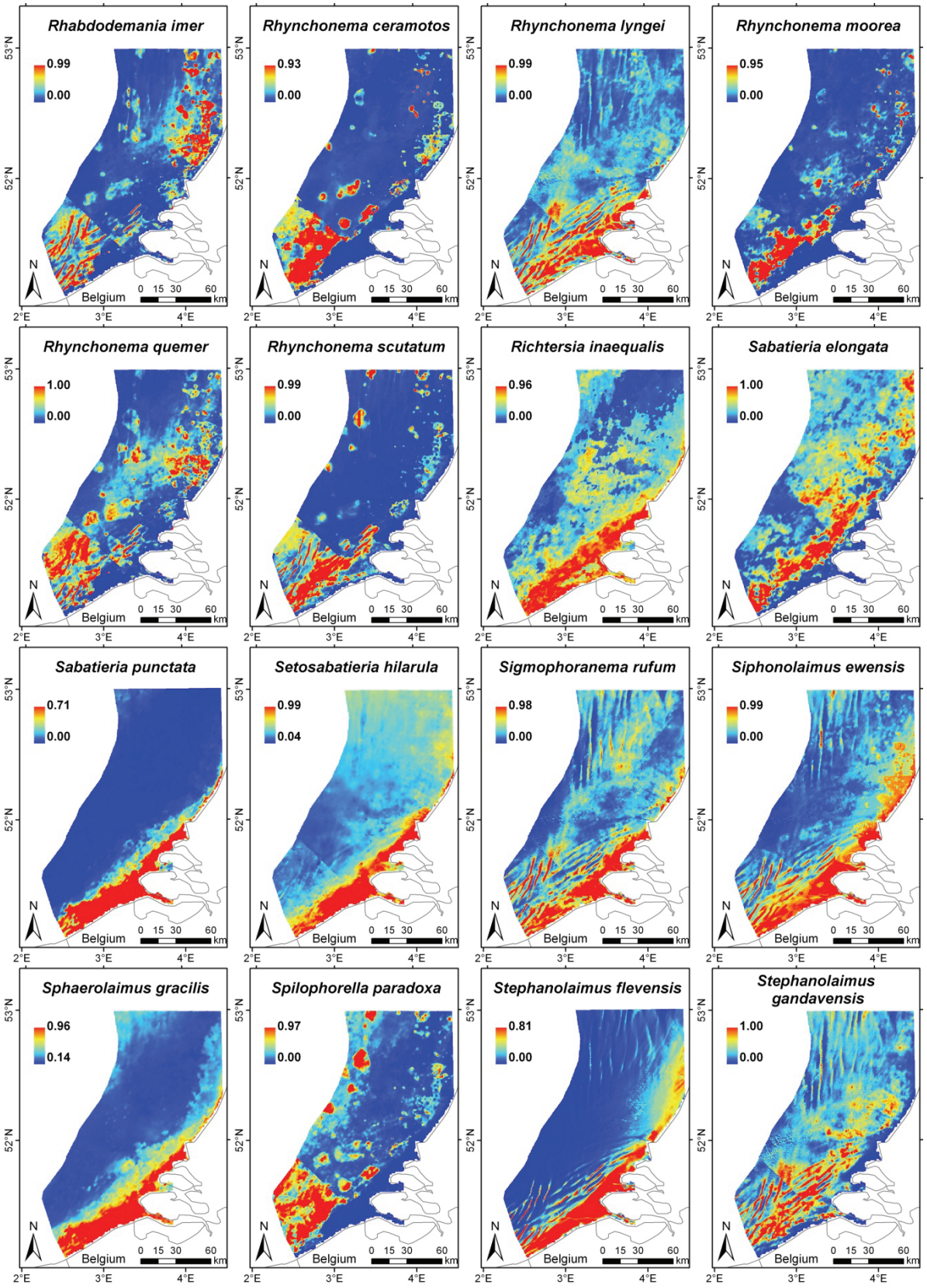

Actinonema celtica
1.00
0.00
Anomonema deconincki
0.96
0.00
Anticoma acuminata
0.96
0.00
Ascolaimus elongatus
0.99
0.00
Axonolaimus orcombensis
0.98
0.00
Bathylaimus capacosus
0.99
0.00
Bathylaimus parafilicaudatus
0.91
0.00
Bolbolaimus riemanni
0.99
0.00
Bolbolaimus teutonicus
0.97
0.00
Calomicrolaimus parahonestus
0.95
0.00
Camacolaimus longicauda
0.99
0.00
Camacolaimus tardus
0.85
0.00
Ceramonema yunfengi
0.97
0.00
Chromaspirina renaudae
0.99
0.00
Cobbia trefusiaeformis
0.91
0.00
Cyartonema elegans
0.97
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km

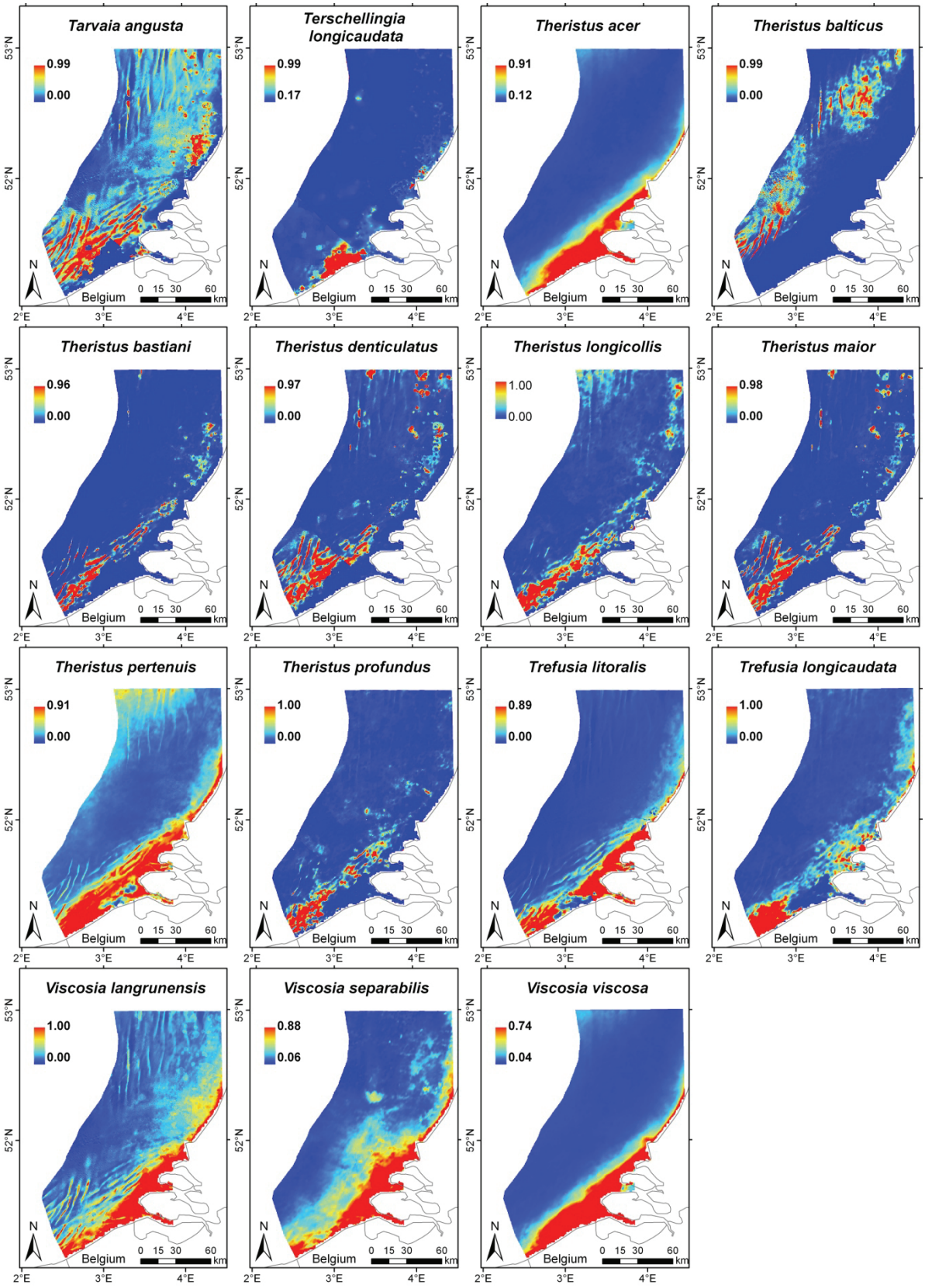

Daptonema fistulatum
0.87
0.00
Daptonema hirsutum
0.93
0.00
Daptonema kornoeense
0.93
0.05
Daptonema nanum
0.95
0.00
Daptonema normandicum
0.78
0.01
Daptonema proprium
0.93
0.01
Daptonema riemanni
0.95
0.01
Daptonema svalbardense
0.89
0.00
Daptonema tenuispiculum
1.00
0.01
Daptonema trichinus
0.88
0.00
Daptonema xyaliforme
0.99
0.02
Desmodora cephalata
1.00
0.00
Desmodora hirsuta
1.00
0.00
Desmodora schulzi
1.00
0.00
Desmolaimus zeelandicus
0.92
0.00
Desmoscolex longisetosus
1.00
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Dichromadora hyalocheile
0.98
0.00
Diplopeltula ostrita
0.98
0.00
Dracognomus tinae
0.97
0.00
Enoploides delamarei
1.00
0.00
Enoplolaimus conicollis
0.93
0.00
Enoplolaimus denticulatus
1.00
0.00
Enoplolaimus litoralis
0.99
0.00
Epacanthion galeatum
0.95
0.00
Epacanthion gorgonocephalum
0.90
0.00
Epsilonema pustulatum
0.99
0.00
Gammanema rapax
0.97
0.00
Gonionchus cumbraensis
1.00
0.00
Gonionchus inaequalis
0.99
0.00
Gonionchus longicaudatus
0.99
0.00
Halichoanolaimus robustus
0.99
0.00
Ixonema sordidum
1.00
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Leptolaimus venustus
1.00
0.00
Manunema annulatum
0.97
0.00
Mesacanthion diplechma
0.91
0.00
Metadesmolaimus aduncus
0.97
0.00
Metadesmolaimus gelana
0.98
0.00
Metadesmolaimus varians
0.86
0.00
Metalinhomoeus biformis
0.87
0.00
Metepsilonema emersum
0.99
0.00
Metepsilonema hagmeieri
1.00
0.00
Metoncholaimus scanicus
0.99
0.17
Microlaimus annelisae
0.98
0.00
Microlaimus arenicola
0.97
0.00
Microlaimus conothelis
0.98
0.02
Neochromadora angelica
0.99
0.00
Neochromadora poecilosoma
0.97
0.00
Odontophora exharena
0.96
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Odontophora ornata
0.98
0.00
Odontophora rectangula
0.89
0.00
Oncholaimellus calvadosicus
0.89
0.00
Oxyonchus dentatus
1.00
0.00
Paracyatholaimoides labiosetosus
0.95
0.00
Paracyatholaimoides multispiralis
1.00
0.00
Pareurystomina acuminata
1.00
0.00
Pomponema elegans
0.99
0.00
Pomponema multipapillatum
0.98
0.00
Pomponema tesselatum
1.00
0.00
Prochromadorella ditlevseni
0.99
0.00
Prochromadorella longicaudata
0.81
0.01
Prochromadorella septempapillata
1.00
0.00
Pseudonchus decempapillatus
1.00
0.00
Ptycholaimellus ponticus
0.98
0.00
Ptycholaimellus vincxae
0.95
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Rhabdodemania imer
0.99
0.00
Rhynchonema ceramotos
0.93
0.00
Rhynchonema lyngei
0.99
0.00
Rhynchonema moorea
0.95
0.00
Rhynchonema quemer
1.00
0.00
Rhynchonema scutatum
0.99
0.00
Richtersia inaequalis
0.96
0.00
Sabatieria elongata
1.00
0.00
Sabatieria punctata
0.71
0.00
Setosabatieria hilarula
0.99
0.04
Sigmophoranema rufum
0.98
0.00
Siphonolaimus ewensis
0.99
0.00
Sphaerolaimus gracilis
0.96
0.14
Spilophorella paradoxa
0.97
0.00
Stephanolaimus flevensis
0.81
0.00
Stephanolaimus gandavensis
1.00
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km

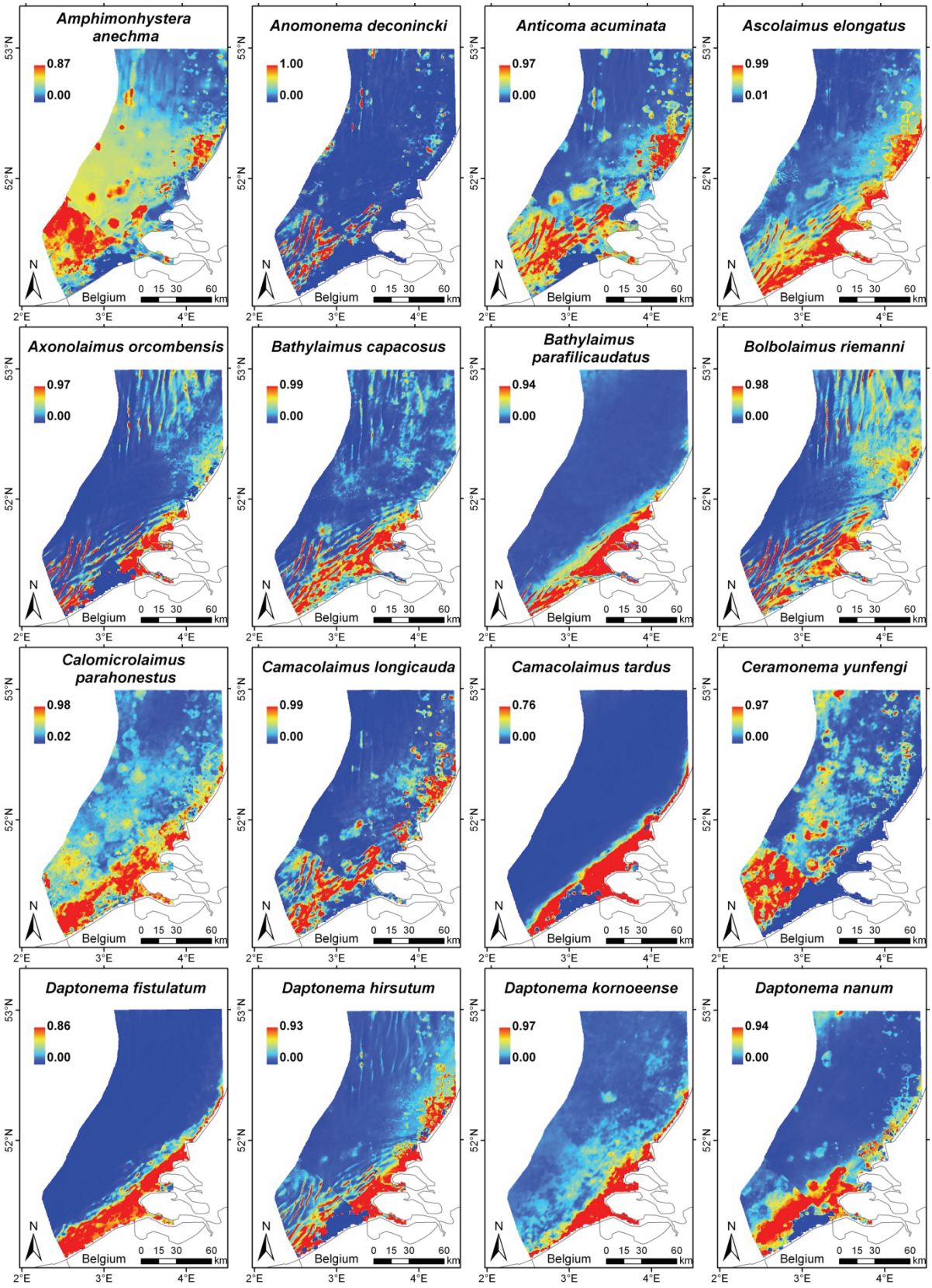

Tarvaia angusta
0.99
0.00
Terschellingia longicaudata
0.99
0.17
Theristus acer
0.91
0.12
Theristus balticus
0.99
0.00
Theristus bastiani
0.96
0.00
Theristus denticulatus
0.97
0.00
Theristus longicollis
1.00
0.00
Theristus maior
0.98
0.00
Theristus pertenuis
0.91
0.00
Theristus profundus
1.00
0.00
Trefusia litoralis
0.89
0.00
Trefusia longicaudata
1.00
0.00
Viscosia langrunensis
1.00
0.00
Viscosia separabilis
0.88
0.06
Viscosia viscosa
0.74
0.04
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km

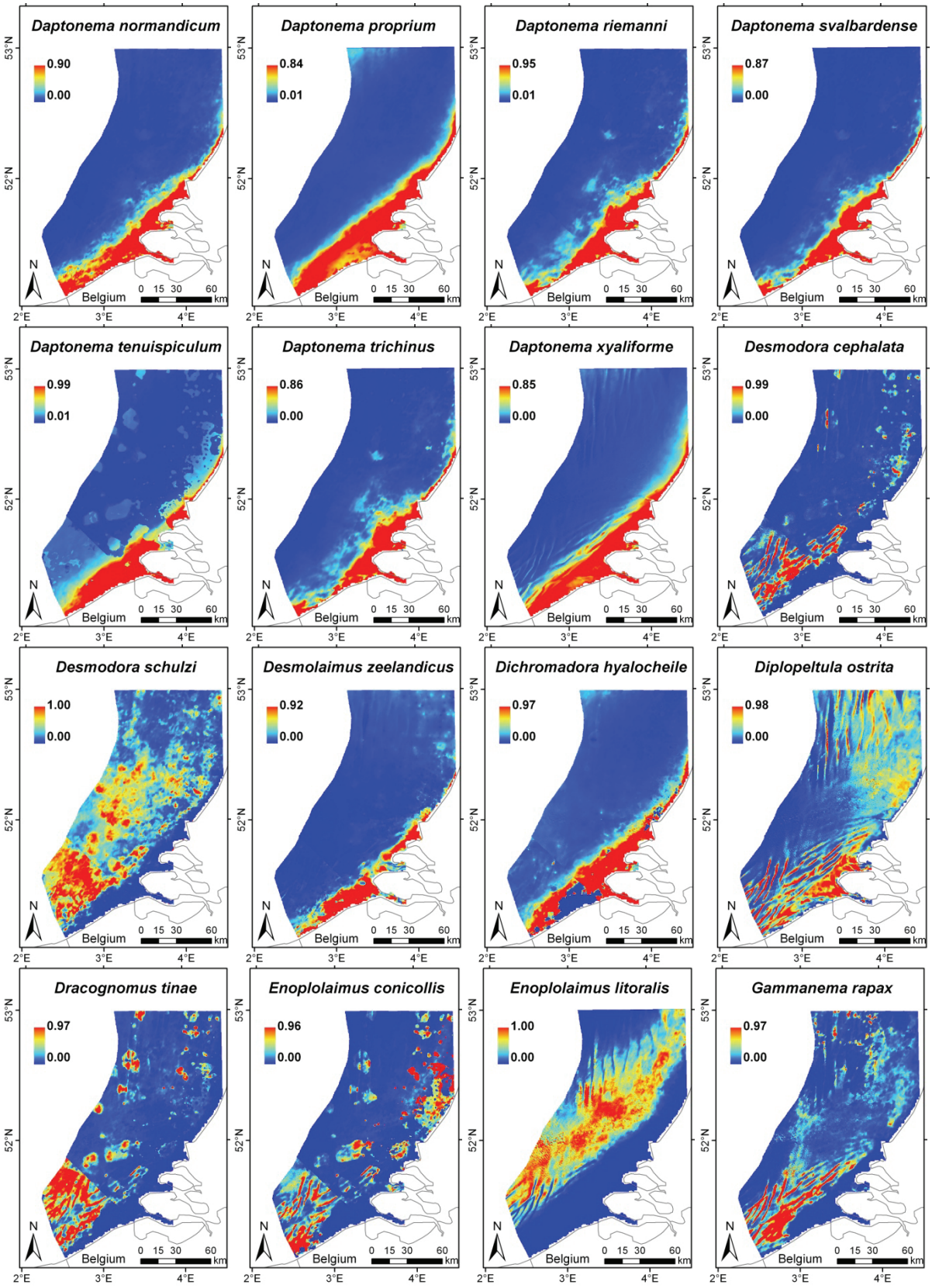
5 KM

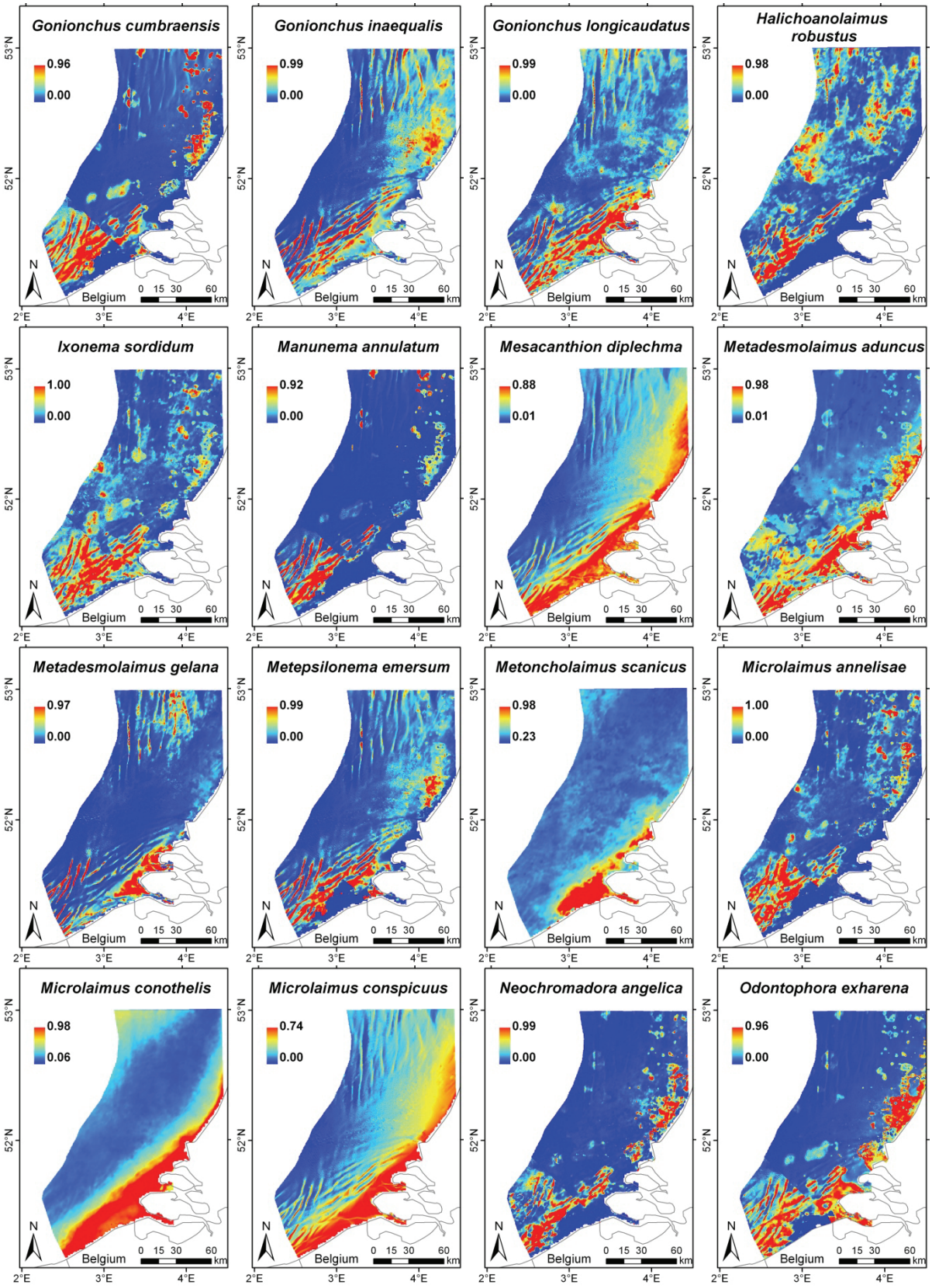
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Amphimonhystera anechma	0	0	0	56	0	44	0	0	0
Anomonema deconincki	19.8	0	6.9	13.9	24.2	35.2	0	0	0
Anticoma acuminata	0	0	0	27.7	18	54.4	0	0	0
Ascolaimus elongatus	0	16.3	0	2.6	38.8	42.3	0	0	0
Axonolaimus orcombensis	12.4	0	0	2.2	61.7	18.6	0	5	0
Bathylaimus capacosus	17.3	3.3	10	4.3	57.4	0	5.8	1.8	0
Bathylaimus parafilicaudatus	4.8	0	2.5	0	56.8	0	35.8	0	0
Bolbolaimus riemanni	16.1	0	0	11.1	72.8	0	0	0	0
Calomicrolaimus parahonestus	0	0	13	4.3	0	32.7	0	50	0
Camacolaimus longicauda	0	0	22.4	9.6	13.9	50.3	3.8	0	0
Camacolaimus tardus	0	0	0	0	0	9.7	90.3	0	0
Ceramonema yunfengi	27.4	0	10.1	0	0	50.9	5.3	0	6.2
Daptonema fistulatum	0	0	3.5	0	5.3	0	0	91.2	0
Daptonema hirsutum	1.1	0	0	2	59.4	37.5	0	0	0
Daptonema kornoeense	0.1	90.1	7.6	0	0	2.2	0	0	0
Daptonema nanum	10.9	6.6	0	0	0	42.9	39.6	0	0
Daptonema normandicum	0	17.7	0	2	3.8	0	0	76.5	0
Daptonema proprium	0	3.3	0	0	0	0	96.7	0	0
Daptonema riemanni	0	91.9	0	0	0	8.1	0	0	0
Daptonema svalbardense	0	95.4	0	0	0	0	0	4.6	0
Daptonema tenuispiculum	0	0	1.8	0	0	21.3	76.9	0	0
Daptonema trichinus	0	100	0	0	0	0	0	0	0
Daptonema xyaliforme	0	0	0	0	10.8	0	89.2	0	0
Desmodora cephalata	11.2	0	18	14.4	17.2	39.2	0	0	0
Desmodora schulzi	9	15.3	17.1	30.5	0	28	0	0	0
Desmolaimus zeelandicus	0	0	0	11	0	0	86.9	0	2
Dichromadora hyalocheile	0	0	0	2.2	0	13.4	84.5	0	0

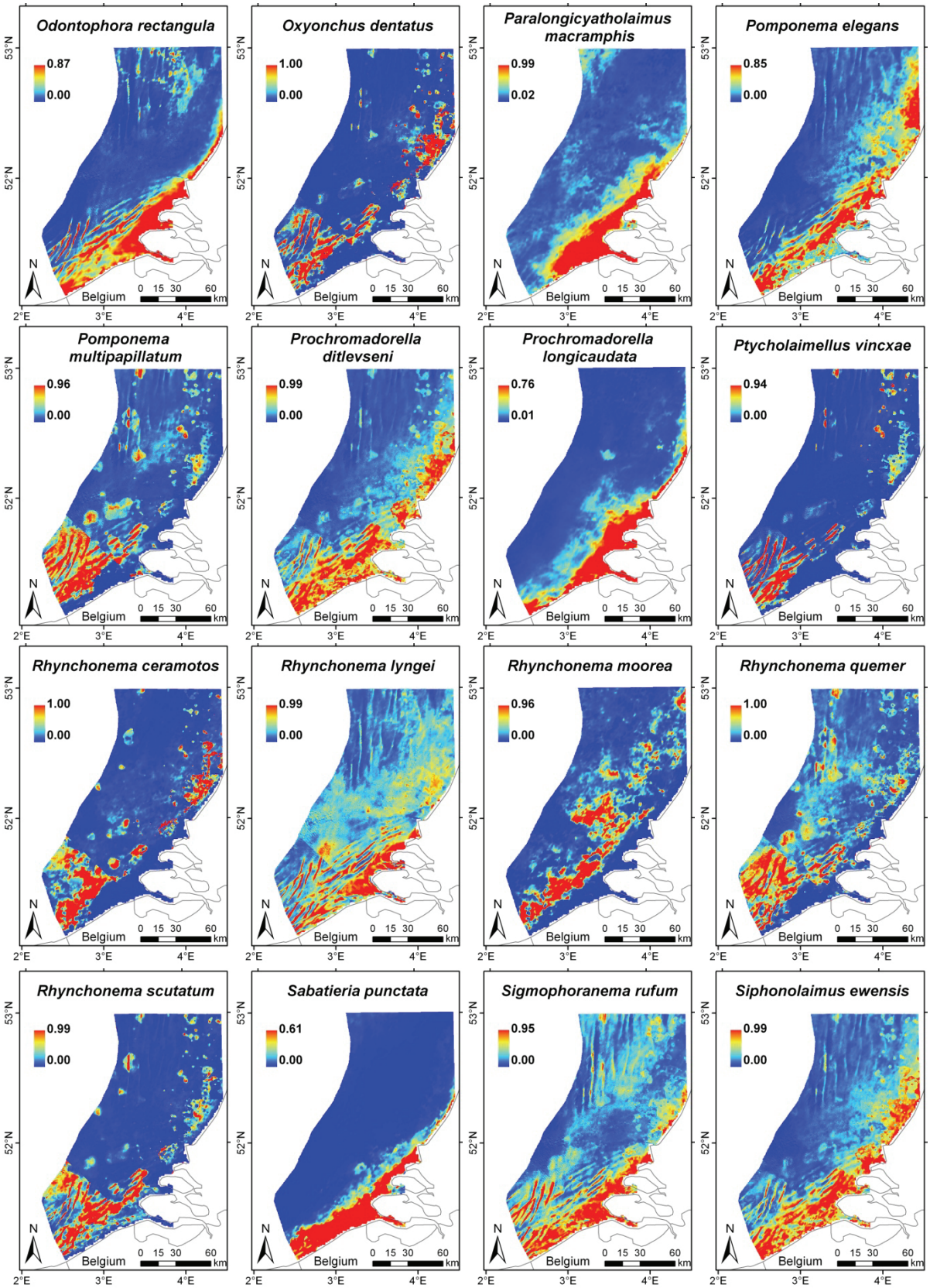
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Diplopeltula ostrita	26	0	0	7.2	66.8	0	0	0	0
Dracognomus tinae	23.7	0	6.2	0	11	46.8	4.9	1.7	5.6
Enoplolaimus conicollis	0	0	12.5	0	16.2	47.1	12.2	0	12
Enoplolaimus litoralis	12.6	1.7	0	0	14.6	0	71.2	0	0
Gammanema rapax	22.4	20.3	0	2.5	31.3	0	10.7	0	12.8
Gonionchus cumbraensis	19.2	0	0	0	18.8	61.9	0	0	0
Gonionchus inaequalis	19.4	0	0	12.1	68.5	0	0	0	0
Gonionchus longicaudatus	18.9	0	9.2	7.8	59.4	0	4.7	0	0
Halichoanolaimus robustus	23.1	14.6	25.4	0	22.1	0	3.5	1.4	9.8
Ixonema sordidum	15	11.5	7.7	6.9	16.5	38.4	3.9	0	0
Manunema annulatum	15.5	0	0	0.6	21.7	49	9.8	0	3.3
Mesacanthion diplechma	0	0	2.8	0	97.2	0	0	0	0
Metadesmolaimus aduncus	0	0	10.7	4.9	54.8	29.6	0	0	0
Metadesmolaimus gelana	29.7	5.4	0	2.7	62.3	0	0	0	0
Metepsilonema emersum	22.6	0	0	14.6	52.3	9.7	0.8	0	0
Metoncholaimus scanicus	0	0	0	0	0	0	0	0	100
Microlaimus annelisae	16.3	16.8	10	0	13.9	43	0	0	0
Microlaimus conothelis	0	0	0	0	0	0	100	0	0
Microlaimus conspicuus	0	0	0	0	83.1	0	16.9	0	0
Neochromadora angelica	0	7.6	26.4	3.9	10.3	45.7	6.1	0	0
Odontophora exharena	0	0	0	9.8	24.8	65.4	0	0	0
Odontophora rectangula	7.6	25.6	0	0	14.8	0	52	0	0
Oxyonchus dentatus	0	6.6	15.5	9.6	14.4	49.9	0	4	0
Paralongicyatholaimus macramphis	0	0	0	0	0	8.5	0	0	91.5
Pomponema elegans	0	0	40.7	0	59.3	0	0	0	0
Pomponema multipapillatum	14.3	0	0	16.2	14.3	47	0	8.3	0
Prochromadorella ditlevseni	0	0	11.1	5.4	32.7	50.9	0	0	0
Prochromadorella longicaudata	0	100	0	0	0	0	0	0	0

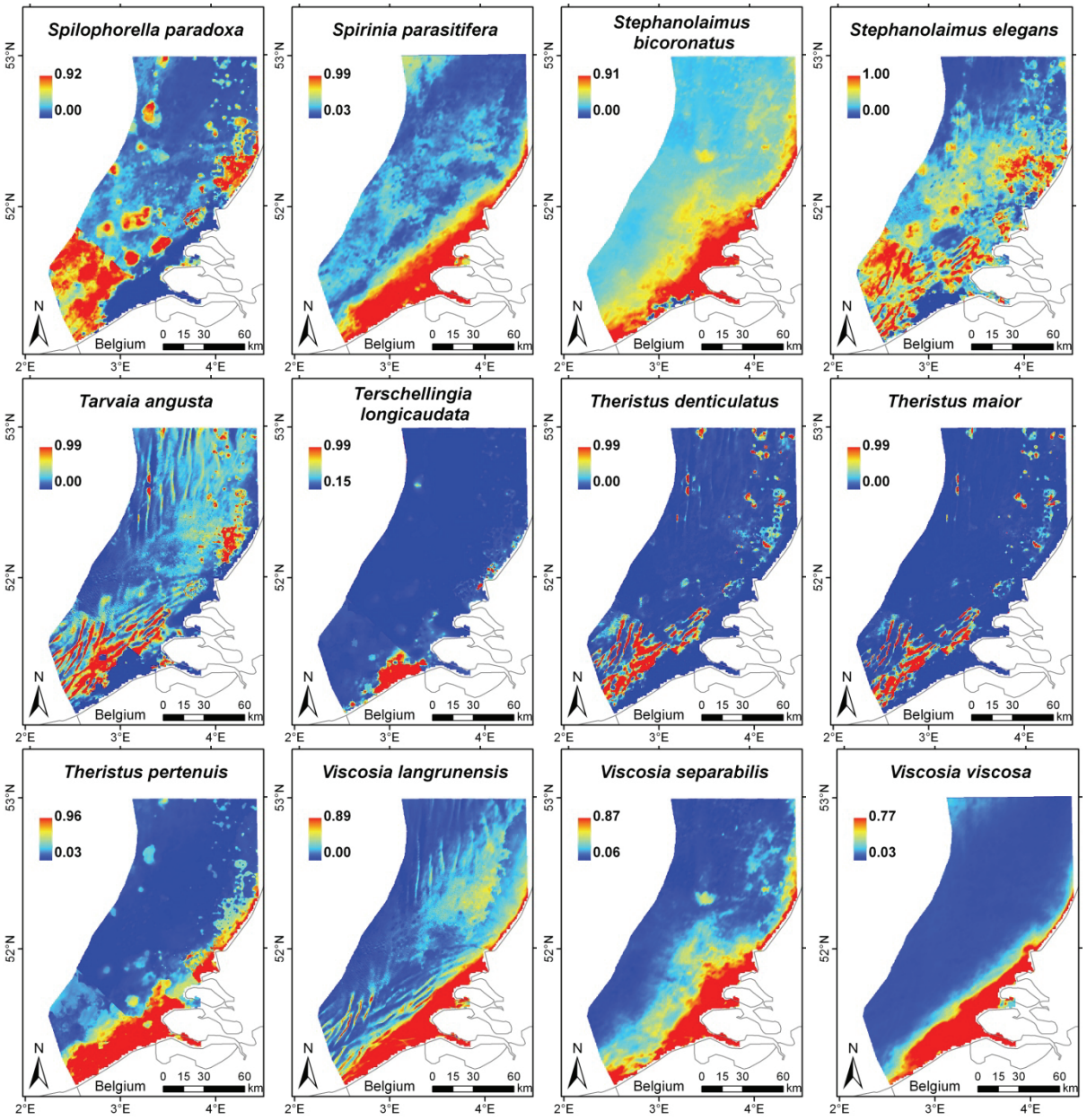
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Ptycholaimellus vincxae	14.1	0	0	3.7	27.4	41.7	9.2	0.4	3.6
Rhynchonema ceramotos	0	10.3	23.1	3.3	0	57.2	0	0	6.1
Rhynchonema lyngei	15.1	6.7	0	12.6	65.6	0	0	0	0
Rhynchonema moorea	9.3	34.1	56.6	0	0	0	0	0	0
Rhynchonema quemer	16.4	0	13.7	16.9	12.4	40.7	0	0	0
Rhynchonema scutatum	13.2	0	14.4	4.4	12.6	55.4	0	0	0
Sabatieria punctata	0	0	0	0	0	0	0	100	0
Sigmophoranema rufum	11.5	0	0	8	43.4	24.8	0	12.3	0
Siphonolaimus ewensis	0	5.3	13.7	1.1	42.9	36.9	0	0	0
Spilophorella paradoxa	0	0	0	18.3	0	64.3	0	0	17.4
Spirinia parasitifera	0	0	0	0	0	0	86.2	8.8	5
Stephanolaimus bicoronatus	0	92.4	0	0	0	7.6	0	0	0
Stephanolaimus elegans	0	0	10	25.1	17.8	37.6	0	9.5	0
Tarvaia angusta	21	0	0	22.1	36.9	20.1	0	0	0
Terschellingia longicaudata	0	0	0	0	0	100	0	0	0
Theristus denticulatus	20.9	0	15.5	5.1	16.8	41.6	0	0	0
Theristus maior	12.9	0	22.3	3.7	17.2	41.8	0	2.2	0
Theristus pertenuis	0	0	0	0	0	40.1	0	59.9	0
Viscosia langrunensis	0	0	0	0.6	20.5	0	75.3	3.5	0
Viscosia separabilis	0.1	99.9	0	0	0	0	0	0	0
Viscosia viscosa	0	0	0	0	0	0	93.2	0	6.8


Amphimonhystera anechma
0.87
0.00
Anomonema deconincki
1.00
0.00
Anticoma acuminata
0.97
0.00
Ascolaimus elongatus
0.99
0.01
Axonolaimus orcombensis
0.97
0.00
Bathylaimus capacosus
0.99
0.00
Bathylaimus parafilicaudatus
0.94
0.00
Bolbolaimus riemanni
0.98
0.00
Calomicrolaimus parahonestus
0.98
0.02
Camacolaimus longicauda
0.99
0.00
Camacolaimus tardus
0.76
0.00
Ceramonema yunfengi
0.97
0.00
Daptonema fistulatum
0.86
0.00
Daptonema hirsutum
0.93
0.00
Daptonema kornoeense
0.97
0.00
Daptonema nanum
0.94
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Daptonema normandicum
0.90
0.00
Daptonema proprium
0.84
0.01
Daptonema riemanni
0.95
0.01
Daptonema svalbardense
0.87
0.00
Daptonema tenuispiculum
0.99
0.01
Daptonema trichinus
0.86
0.00
Daptonema xyaliforme
0.85
0.00
Desmodora cephalata
0.99
0.00
Desmodora schulzi
1.00
0.00
Desmolaimus zeelandicus
0.92
0.00
Dichromadora hyalocheile
0.97
0.00
Diplopeltula ostrita
0.98
0.00
Dracognomus tinae
0.97
0.00
Enoplolaimus conicollis
0.96
0.00
Enoplolaimus litoralis
1.00
0.00
Gammanema rapax
0.97
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Gonionchus cumbraensis
0.96
0.00
Gonionchus inaequalis
0.99
0.00
Gonionchus longicaudatus
0.99
0.00
Halichoanolaimus robustus
0.98
0.00
Ixonema sordidum
1.00
0.00
Manunema annulatum
0.92
0.00
Mesacanthion diplechma
0.88
0.01
Metadesmolaimus aduncus
0.98
0.01
Metadesmolaimus gelana
0.97
0.00
Metepsilonema emersum
0.99
0.00
Metoncholaimus scanicus
0.98
0.23
Microlaimus annelisae
1.00
0.00
Microlaimus conothelis
0.98
0.06
Microlaimus conspicuus
0.74
0.00
Neochromadora angelica
0.99
0.00
Odontophora exharena
0.96
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Odontophora rectangula
0.87
0.00
Oxyonchus dentatus
1.00
0.00
Paralongicyatholaimus macramphis
0.99
0.02
Pomponema elegans
0.85
0.00
Pomponema multipapillatum
0.96
0.00
Prochromadorella ditlevseni
0.99
0.00
Prochromadorella longicaudata
0.76
0.01
Ptycholaimellus vincxae
0.94
0.00
Rhynchonema ceramotos
1.00
0.00
Rhynchonema lyngei
0.99
0.00
Rhynchonema moorea
0.96
0.00
Rhynchonema quemer
1.00
0.00
Rhynchonema scutatum
0.99
0.00
Sabatieria punctata
0.61
0.00
Sigmophoranema rufum
0.95
0.00
Siphonolaimus ewensis
0.99
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km

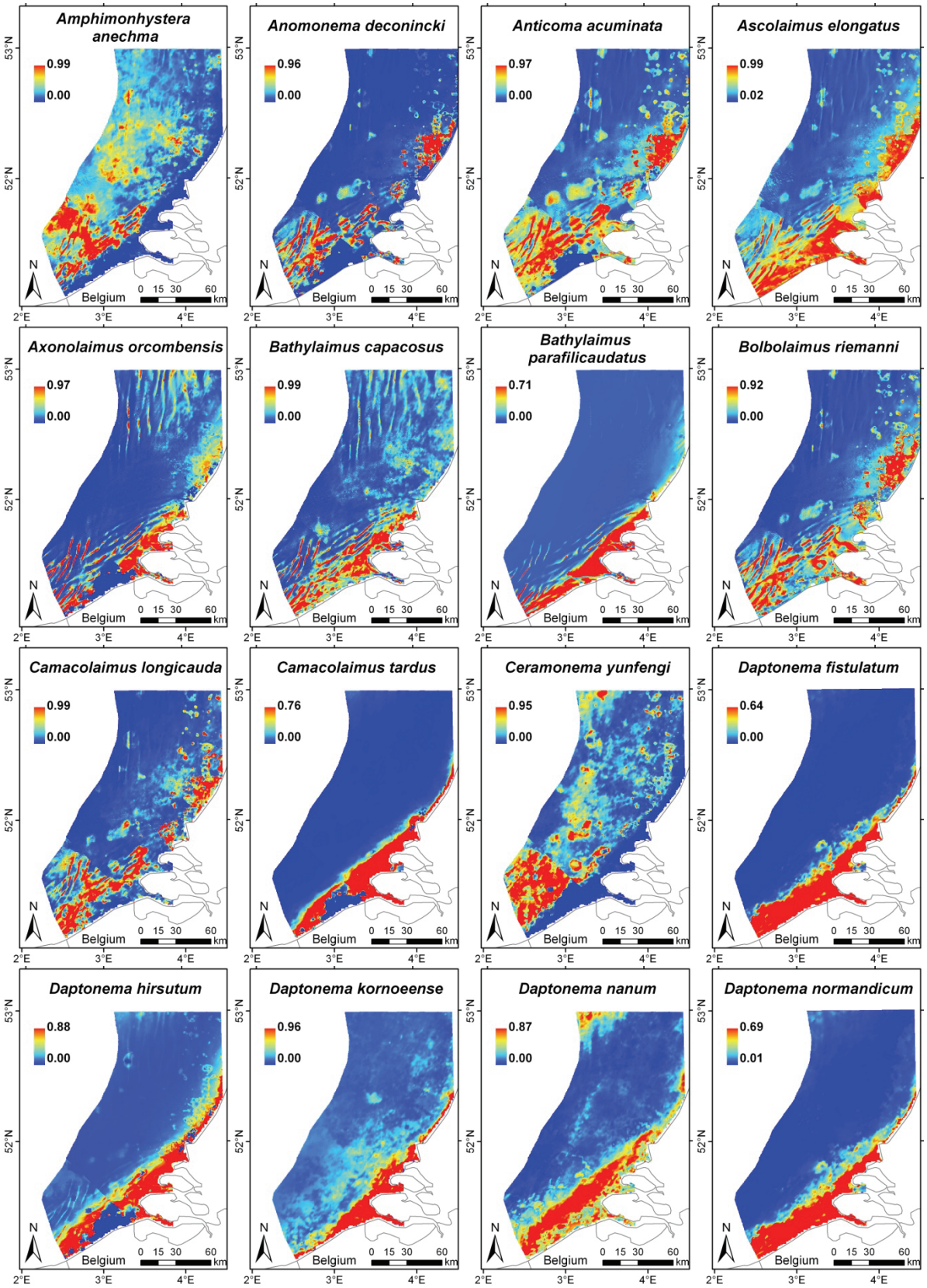

Spilophorella paradoxa
0.92
0.00
Spirinia parasitifera
0.99
0.03
Stephanolaimus bicoronatus
0.91
0.00
Stephanolaimus elegans
1.00
0.00
Tarvaia angusta
0.99
0.00
Terschellingia longicaudata
0.99
0.15
Theristus denticulatus
0.99
0.00
Theristus maior
0.99
0.00
Theristus pertenuis
0.96
0.03
Viscosia langrunensis
0.89
0.00
Viscosia separabilis
0.87
0.06
Viscosia viscosa
0.77
0.03
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km

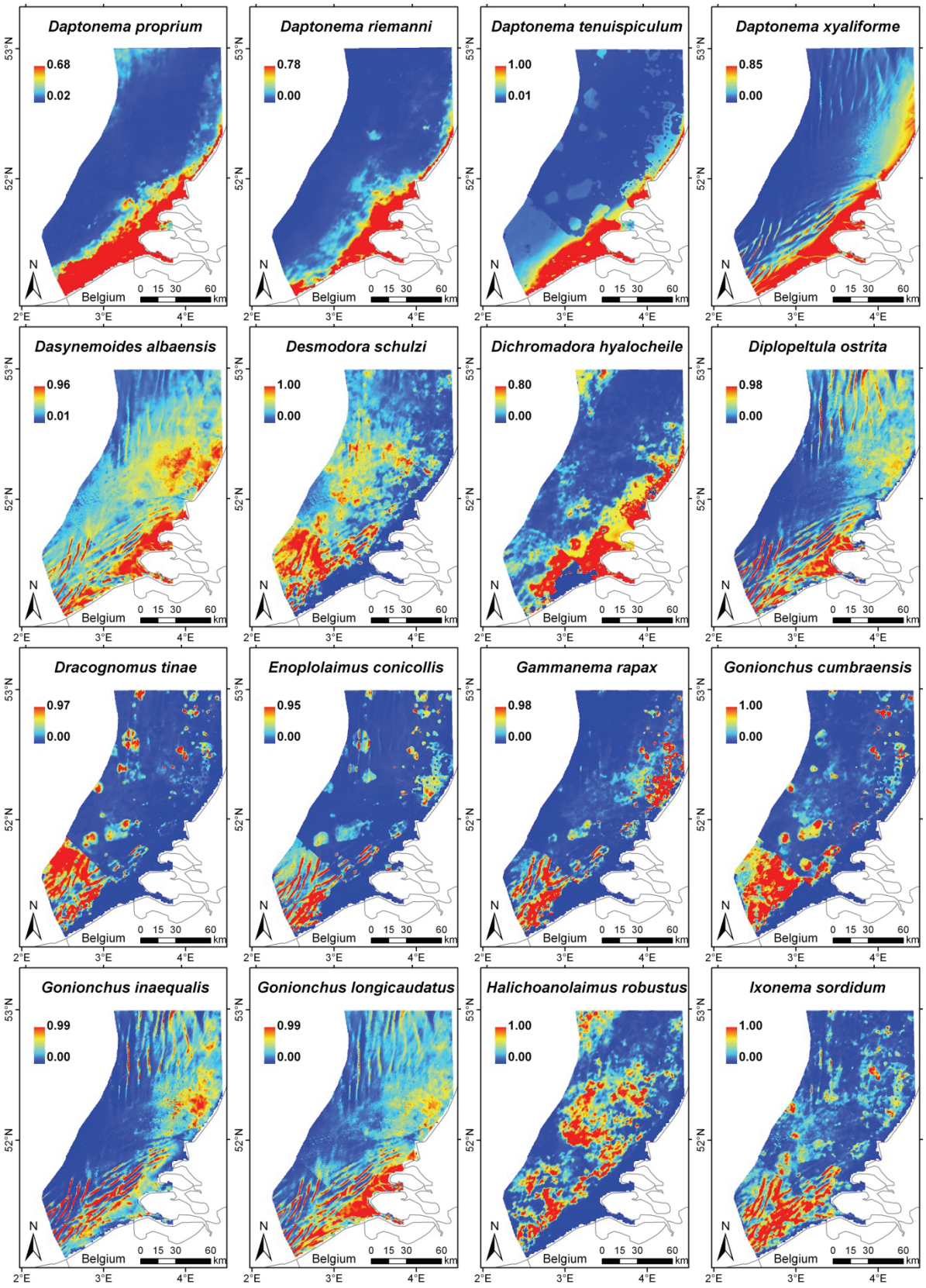
10 KM

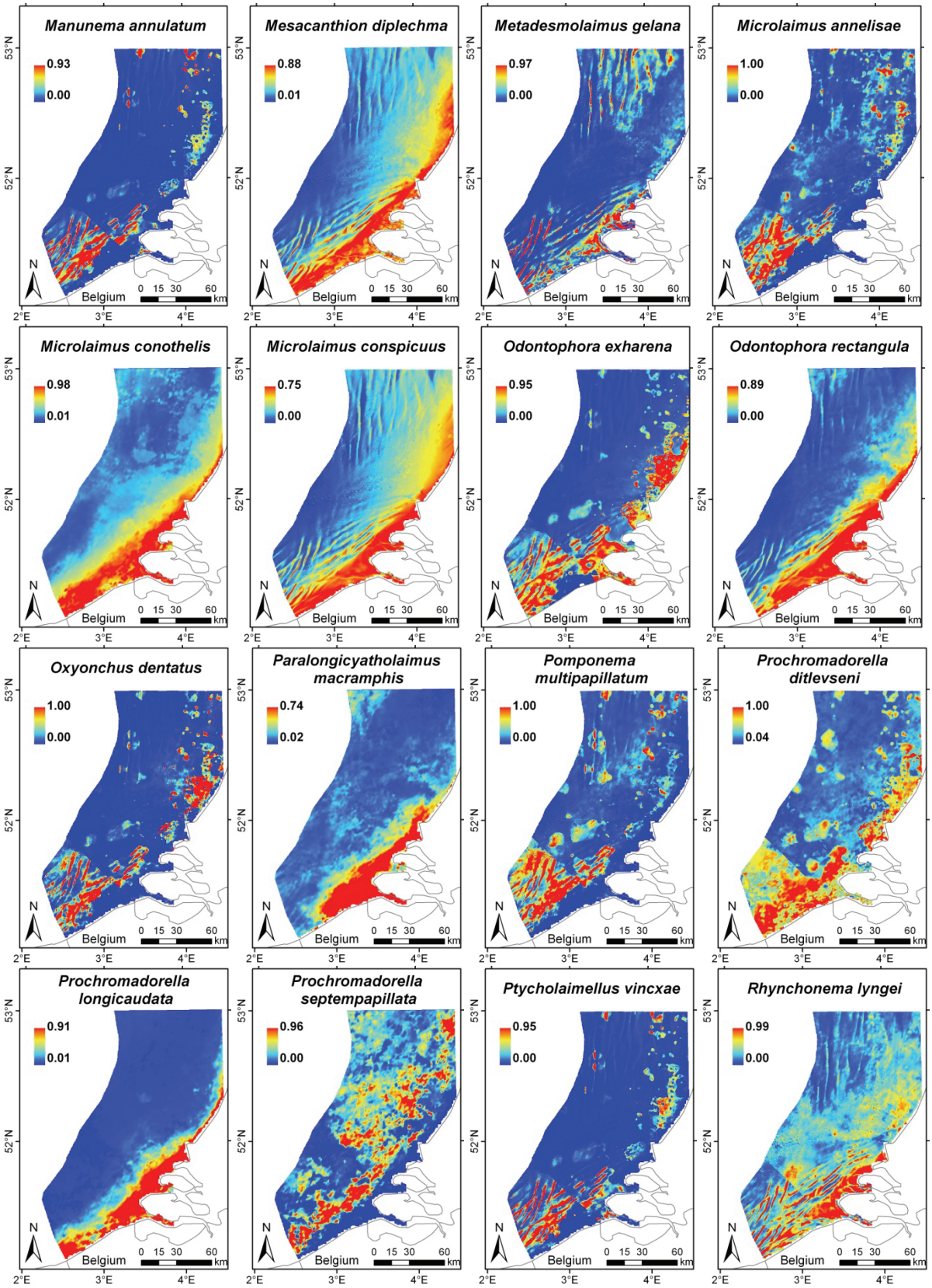
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Amphimonhystera anechma	18.7	0	12.6	33.3	8.5	26.9	0	0	0
Anomonema deconinckí	0	0	0	26.8	25.5	47.7	0	0	0
Anticoma acuminata	0	0	0	27.7	18	54.4	0	0	0
Ascolaimus elongatus	0	0	0	4.4	43.2	52.3	0	0	0
Axonolaimus orcombensis	15.7	0	0	4.4	61.2	18.7	0	0	0
Bathylaimus capacosus	15.5	0	10.8	5.2	65.3	0	0	3.1	0
Bathylaimus parafilicaudatus	3.5	0	0	0.8	95.7	0	0	0	0
Bolbolaimus riemanni	0	0	0	12.4	33.3	54.3	0	0	0
Camacolaimus longicauda	0	0	22.2	9.4	14.2	49.3	3.4	0	1.5
Camacolaimus tardus	0	0	0	0	0	9.7	90.3	0	0
Ceramonema yunfengi	28.3	0	9.2	0	0	53.2	0	3.4	5.9
Daptonema fistulatum	0	0	0	0	0	0	0	100	0
Daptonema hirsutum	0	0	0	0	6.6	20	69.9	0	3.5
Daptonema kornoeense	0	90.3	7.6	0	0	2.1	0	0	0
Daptonema nanum	11.1	0	10.4	0	0	0	78.5	0	0
Daptonema normandicum	0	0	0	0	0	0	0	100	0
Daptonema proprium	0	0	0	0	0	0	0	100	0
Daptonema riemanni	0	100	0	0	0	0	0	0	0
Daptonema tenuispiculum	0	0	0	1.6	0	19	79.4	0	0
Daptonema xyaliforme	0	0	0.2	0	99.8	0	0	0	0
Dasynemoides albaensis	0	0	0	18.9	75.6	0	0	5.6	0
Desmodora schulzi	18.1	0	13.7	34	10.3	23.9	0	0	0
Dichromadora hyalocheile	0	0	0	0	0	33.2	0	0	66.8
Diplopeltula ostrita	26	0	0	7.2	66.8	0	0	0	0
Dracognomus tinae	23.7	0	7.3	0	11.9	46.5	4.8	0	5.8
Enoplolaimus conicollis	13.4	0	0	7.9	23.2	38.1	9	0	8.5
Gammanema rapax	0	16.8	8.1	6.7	12.6	45.5	7	0	3.3
Gonionchus cumbraensis	10.6	12.4	13.8	2.1	0	61.1	0	0	0

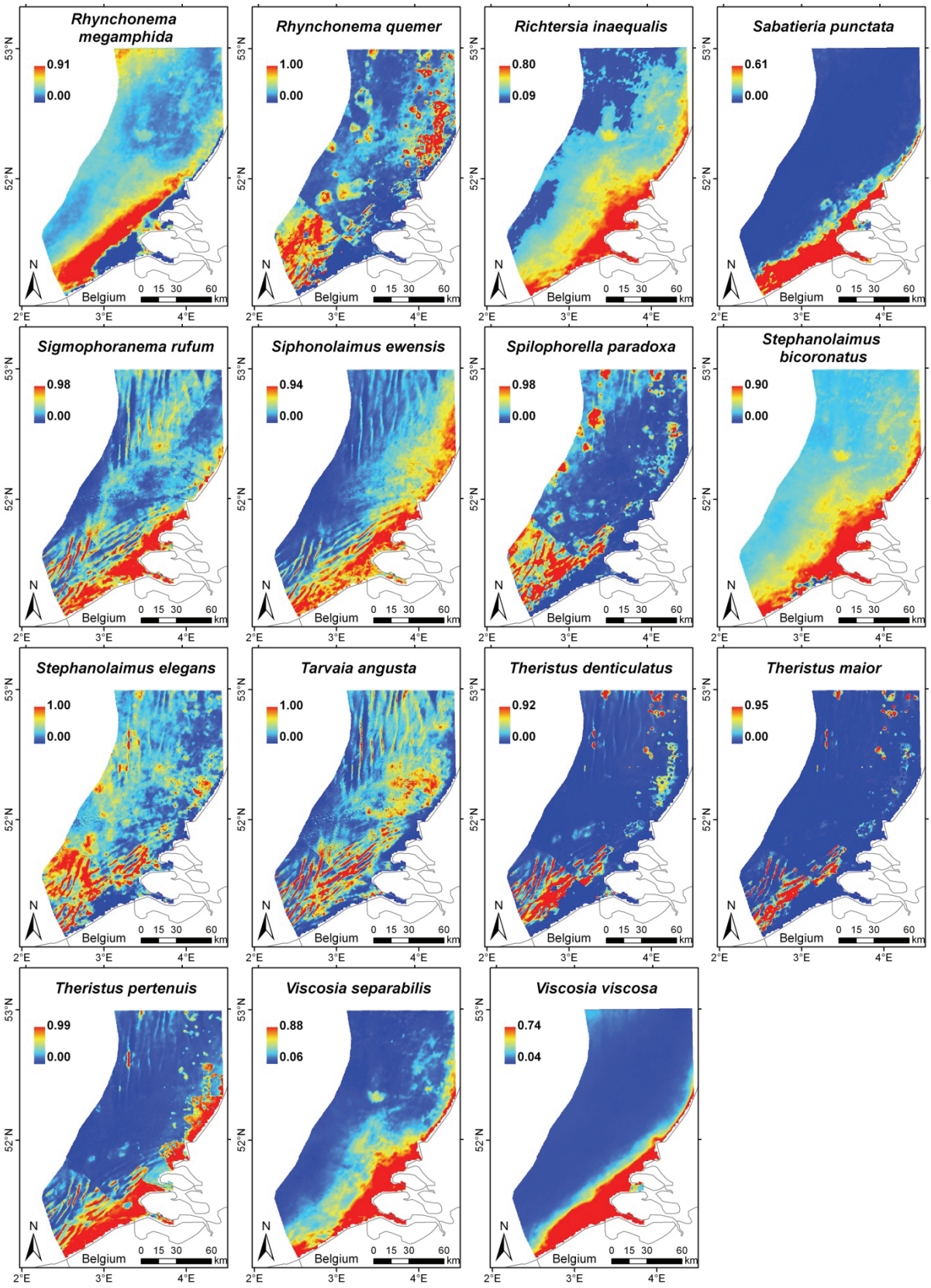
	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Gonionchus inaequalis	19.4	0	0	12.1	68.5	0	0	0	0
Gonionchus longicaudatus	15.3	0	0	10	74.7	0	0	0	0
Halichoanolaimus robustus	17.3	0	44.7	21.5	0	0	0	0	16.5
Ixonema sordidum	15.4	11.1	7.4	5.5	15.3	36.6	5.1	0	0
Manunema annulatum	17.3	0	0	0	23.4	49.3	10	0	0
Mesacanthion diplechma	0	0	2.8	0	97.2	0	0	0	0
Metadesmolaimus gelana	26.5	0	6.4	5.2	61.9	0	0	0	0
Microlaimus annelisae	16.9	13.5	8.1	0	13.3	40.6	7.6	0	0
Microlaimus conothelis	0	9.5	0	0	0	0	76.3	14.3	0
Microlaimus conspicuus	0	0	0	0	99.7	0	0	0	0.3
Odontophora exharena	0	0	0	9.8	24.8	65.4	0	0	0
Odontophora rectangula	0	0	0	0	62.2	0	0	0	37.8
Oxyonchus dentatus	0	6.7	15.2	10.6	14.9	49.3	3.4	0	0
Paralongicyatholaimus macramphis	0	0	0	0	0	0	0	0	100
Pomponema multipapillatum	15.8	0	8	14.2	11.8	42.6	7.6	0	0
Prochromadorella ditlevseni	0	0	24.5	0	0	75.5	0	0	0
Prochromadorella longicaudata	0.9	63.4	2.7	0	0	0	0	33	0
Prochromadorella septempapillata	15.4	0	84.6	0	0	0	0	0	0
Ptycholaimellus vincxae	16.1	0	0	5.4	30.3	47.1	0	1.2	0
Rhynchonema lyngei	15.1	6.7	0	12.6	65.6	0	0	0	0
Rhynchonema megamphida	5	23.7	0	0	0	0	57.8	0	13.5
Rhynchonema quemer	0	23.3	9.4	0	11.4	55.9	0	0	0
Richtersia inaequalis	0	100	0	0	0	0	0	0	0
Sabatieria punctata	0	0	0	0	0	0	0	100	0
Sigmophoranema rufum	16.2	0	0	12.4	58.3	0	0	13.1	0
Siphonolaimus ewensis	0	0	12.7	1.1	86.2	0	0	0	0
Spilophorella paradoxa	21.1	0	10	0	13.7	43.8	11.4	0	0
Stephanolaimus bicoronatus	0	93	0	0	0	7	0	0	0

	Average Chl *a*	Maximum Chl *a*	Minimum Chl *a*	Median grain size	Water depth	Silt-clay content	Average TSM	Maximum TSM	Minimum TSM
Stephanolaimus elegans	17.2	0	10.5	20.1	22.5	29.7	0	0	0
Tarvaia angusta	30.6	0	0	16.2	52.3	0	0.9	0	0
Theristus denticulatus	19.3	0	0	0	22.3	48.4	9.9	0	0
Theristus maior	16.1	0	20.8	0	16.7	40.4	6	0	0
Theristus pertenuis	1.2	0	0	0	34.9	60.6	0	0	3.2
Viscosia separabilis	0	100	0	0	0	0	0	0	0
Viscosia viscosa	0	0	0	0	0	0	100	0	0


Amphimonhystera anechma
0.99
0.00
Anomonema deconincki
0.96
0.00
Anticoma acuminata
0.97
0.00
Ascolaimus elongatus
0.99
0.02
Axonolaimus orcombensis
0.97
0.00
Bathylaimus capacosus
0.99
0.00
Bathylaimus parafilicaudatus
0.71
0.00
Bolbolaimus riemanni
0.92
0.00
Camacolaimus longicauda
0.99
0.00
Camacolaimus tardus
0.76
0.00
Ceramonema yunfengi
0.95
0.00
Daptonema fistulatum
0.64
0.00
Daptonema hirsutum
0.88
0.00
Daptonema kornoeense
0.96
0.00
Daptonema nanum
0.87
0.00
Daptonema normandicum
0.69
0.01
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Daptonema proprium
0.68
0.02
Daptonema riemanni
0.78
0.00
Daptonema tenuispiculum
1.00
0.01
Daptonema xyaliforme
0.85
0.00
Dasynemoides albaensis
0.96
0.01
Desmodora schulzi
1.00
0.00
Dichromadora hyalocheile
0.80
0.00
Diplopeltula ostrita
0.98
0.00
Dracognomus tinae
0.97
0.00
Enoplolaimus conicollis
0.95
0.00
Gammanema rapax
0.98
0.00
Gonionchus cumbraensis
1.00
0.00
Gonionchus inaequalis
0.99
0.00
Gonionchus longicaudatus
0.99
0.00
Halichoanolaimus robustus
1.00
0.00
Ixonema sordidum
1.00
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Manunema annulatum
0.93
0.00
Mesacanthion diplechma
0.88
0.01
Metadesmolaimus gelana
0.97
0.00
Microlaimus annelisae
1.00
0.00
Microlaimus conothelis
0.98
0.01
Microlaimus conspicuus
0.75
0.00
Odontophora exharena
0.95
0.00
Odontophora rectangula
0.89
0.00
Oxyonchus dentatus
1.00
0.00
Paralongicyatholaimus macramphis
0.74
0.02
Pomponema multipapillatum
1.00
0.00
Prochromadorella ditlevseni
1.00
0.04
Prochromadorella longicaudata
0.91
0.01
Prochromadorella septempapillata
0.96
0.00
Ptycholaimellus vincxae
0.95
0.00
Rhynchonema lyngei
0.99
0.00
53°N
52°N
2°E
3°E
4°E
N
Belgium
0 15 30 60
km


Rhynchonema megamphida
0.91
0.00
Rhynchonema quemer
1.00
0.00
Richtersia inaequalis
0.80
0.09
Sabatieria punctata
0.61
0.00
Sigmophoranema rufum
0.98
0.00
Siphonolaimus ewensis
0.94
0.00
Spilophorella paradoxa
0.98
0.00
Stephanolaimus bicoronatus
0.90
0.00
Stephanolaimus elegans
1.00
0.00
Tarvaia angusta
1.00
0.00
Theristus denticulatus
0.92
0.00
Theristus maior
0.95
0.00
Theristus pertenuis
0.99
0.00
Viscosia separabilis
0.88
0.06
Viscosia viscosa
0.74
0.04
Belgium
0 15 30 60 km
N
53°N
52°N
2°E
3°E
4°E

ADDENDUM 4

MAXENT MODELS

OF

CHAPTER 6

ADDENDUM 4

MAXENT MODELS OF CHAPTER 6

Maxent models can be written as an equation based on the λ-values and values of k_1 and k_2 associated with the λ-values. The value of k_1 and k_2 have a different meaning depending on the feature (see table A4.1). With these values the logistic output of the Maxent model can be calculated for each grid cell.

In Table A4.2 the λ-values and values of k_1 and k_2 are given for the final threshold models of Chapter 6.

For each grid cell the value of the environmental variables is known and the feature function $f(x)$ can be calculated (Table A4.1 and A4.2). When all the feature functions are calculated for a single grid cell, these functions are summed:

$S = \sum_{i=1}^{n} f_n(x) - L$

with n the number of features, and L being the Linear Prediction Normalizer (LPN in Table A4.2).

$q_\lambda(x) = \frac{e^S}{Z_\lambda}$

with Z_λ the Density Normalizer (DN in Table A4.2).

This is the raw data output which is then logistic-scaled to give the final output of the model.

$L_{output} = \frac{q_\lambda(x).e^{entropy}}{1 + q_\lambda(x).e^{entropy}}$

The value of $entropy$ can be found in Table A4.2.

To compute the complete output map, this should be repeated for each grid cell.

Linear	$f(x) = \lambda.\frac{x - k_1}{k_{2-}k_1}$	$k_1$and k_2 are the minimum and maximum value of the variable
Quadratic	$f(x) = \lambda.\frac{x^2 - k_1}{k_2 - k_1}$	$k_1$and k_2 are the minimum and maximum value of the squared variable
Product feature	$f(x,y) = \lambda.\frac{x.y - k_1}{k_2 - k_1}$	$k_1$and k_2 are the minimum and maximum value of the product of the two variables
Forward hinge	if x < k_1 then $f(x) = 0$ else $f(x) = \lambda.\frac{x-k_1}{k_2-k_1}$	k_1= hinge k_2 is the maximum value of the variable
Reverse hinge	if x < k_2 then $f(x) = \lambda.\frac{k_2-x}{k_2-k_1}$ else $f(x) = 0$	k_2= hinge k_1 is the minimum
Threshold	if x < threshold then $f(x) = 0$ else $f(x) = \lambda$	k_1 and k_2 represent the outcome of the threshold test: 0 if true and 1 if false.

Table A4.1. Application of λ, $k_1$and k_2 for each feature.

	variable	feature	λ	k_1	k_2
Daptonema tenuispiculum	Tavg	linear	4.85	0.98	24.06
		LPN	4.85		
		DN	337.34		
		entropy	8.02		
Dichromadora cucullata	Tavg^2	quadratic	-70.62	2.85	579.12
	Cavg*Dept	product	17.99	-15.32	353.94
	Cavg*Tavg	product	-100.46	4.20	337.08
	Cmin*D50x	product	5.98	0.89	2039.59
		LPN	7.27		
		DN	779.14		
		entropy	8.48		
Enoploides spiculohamatus	Cmax	linear	3.44	4.89	38.83
		LPN	3.44		
		DN	1133.01		
		entropy	8.80		
Onyx perfectus	Cavg	linear	-6.08	1.98	17.84
	Cavg^2	quadratic	-25.53	3.93	318.16
	D50x^2	quadratic	1.76	16.15	363174.99
	Dept^2	quadratic	-14.97	0.00	2484.22
	`Cavg	reverse hinge	2.25	1.98	4.35
	`Cmax	reverse hinge	-2.90	4.89	10.15
	`Dept	reverse hinge	0.14	-1.00	9.24
	`Dept	reverse hinge	0.17	-1.00	9.29
	`Dept	reverse hinge	0.11	-1.00	9.52
	`Tavg	reverse hinge	-2.93	1.69	7.53
	`Tmin	reverse hinge	-0.38	0.32	0.86
	`Tmin	reverse hinge	-4.47	0.32	1.06
	`Tmin	reverse hinge	1.07	0.32	1.90
	`Tmin	reverse hinge	0.55	0.32	1.97
	`Tmin	reverse hinge	0.42	0.32	2.82
	`Tmin	reverse hinge	0.61	0.32	2.91
	'Cmax	forward hinge	0.74	28.02	38.83
	'Cmax	forward hinge	0.30	28.15	38.83
		LPN	-0.64		
		DN	167.44		
		entropy	7.61		

	variable	feature	λ	k_1	k_2
Sabatieria celtica	Cmin	linear	7.12	0.04	13.69
	Tavg	linear	4.03	0.98	24.07
	Tmax	linear	1.28	2.42	65.74
	Cavg^2	quadratic	-1.85	1.72	383.22
	Cmin^2	quadratic	-4.02	0.00	187.50
	Tavg^2	quadratic	-2.98	0.96	579.17
	Tmax^2	quadratic	1.67	5.87	4321.48
		LPN	4.18		
		DN	610.23		
		entropy	8.90		
Sabatieria punctata	Cmax	linear	1.35	4.89	38.83
	Tmax	linear	3.00	3.03	55.95
	D50x^2	quadratic	-5.34	16.15	363174.99
	D50x*Tmax	product	1.06	123.91	15593.29
	(18.02<Tmax)	threshold	1.11	0.00	1.00
	(28.06<Tmax)	threshold	0.11	0.00	1.00
	(19.55<Cmax)	threshold	0.31	0.00	1.00
	(304.28<D50x)	threshold	-0.32	0.00	1.00
	(36.48<Tmax)	threshold	-0.83	0.00	1.00
	(2.02<Cmin)	threshold	-0.42	0.00	1.00
	(0.11<Cmin)	threshold	-0.30	0.00	1.00
	(28.56<Tmax)	threshold	0.04	0.00	1.00
	(35.13<Cmax)	threshold	-0.22	0.00	1.00
		LPN	4.59		
		DN	390.95		
		entropy	7.50		

Table A4.2. Model parameters of the threshold models of the six nematode species modelled in Chapter 6.

ADDENDUM 5

MATLAB AND R-CODE

ADDENDUM 5

MATLAB AND R-CODE

This addendum includes the most relevant Matlab and R-scripts developed during the course of this research and has 5 subdivisions with code related to chapters 2 to 6.

Text comments are in italics and gray font and are preceded by % for the Matlab code and by # for the R-code. The script is followed by '…' when the Matlab code continues on the next line.

CODE RELATED TO CHAPTER 2: SPECIES ASSEMBLY RULES

This paragraph summarises the Matlab code for the two swapping algorithms for presence/absence data. The scripts for calculating the co-occurrence indices for the original data matrix are not shown here, since they can be easily derived from the code below.

The code concerns only the feeding type 1A. Adapting the code for the other feeding types or all the data can be easily achieved by replacing '1A' by another feeding type or by 'AllData'.

Swap 1 for presence/absence data

```
%% data needed
% replicodes_1A = file with three columns about data of feeding type 1A
   % name of the replicate// number of the sample// number of replicates in the sample
% matrix_1A = presence/absence matrix with
   % replicates in rows and species in columns

for aantal=1:10              % aantal = number of files with 'aantalRand' null models
    aantalRand=100;              % number of null models in one loop ('aantal')
    % load data
    load('replicodes_1A'); load('matrix_1A');
    replicodes_nieuw=replicodes_1A;
    matrix_nieuw=matrix_1A;
    clear matrix_1A; clear replicodes_1A;
    aant=int2str(aantal);
    [X,Y]=size(matrix_nieuw);
    % Calculate row number in p/a matrix where a new sample starts
    N=1; x=1;
    while x<X
            aantalReps=replicodes_nieuw{1,3}(x);
            rijen(N)=x;
            N=N+1;
            x=x+aantalReps;
    end
```

```
samples=N-1; aantalRep=X; aantalSpec=Y; probleem=0;
 resultaat_null=zeros(aantalRand,8);
 rij=1; rij1=1;
for rand=1:aantalRand
      for n=1:samples % repeat swapping algorithm for all samples
            aantalReps=replicodes_nieuw{1,3}(rijen(n));
            % submatrix = data of 1 sample with 'aantalReps' replicates
                  submatrix=[];
            submatrix=matrix_nieuw(rijen(n):(rijen(n)+aantalReps-1),:);
            MM=1;
            % delete species which do not appear in the sample
             kolomleeg=zeros(Y,1);
             for y=Y:-1:1
                        if submatrix(1:aantalReps,y)==zeros(aantalReps,1)
                        submatrix(:,y)=[];
                        kolomleeg(y,1)=1; % remember the empty species columns
                   end
            end
            [A,B]=size(submatrix);
            swaps=500;
            sw=0;
            aantalSwaps=0;
            if A>1 && B>1     % try swaps if the remaining matrix contains more
                                    % than 1 replicate and more than 1 species
                  while sw<swaps % try 1000 swaps
                        % choose randomly 2 replicates and 2 species
                        randomNr1 = randint(2,1,[1,A]);
                        randomNr2 = randint(2,1,[1,B]);
                        % check if 2 different species and 2 different replicates are chosen
                        if (randomNr1(1)~= randomNr1(2)) && …
                              (randomNr2(1)~= randomNr2(2))
                              sw = sw+1;
                              % check if matrix is like [0 1; 1 0] or [1 0; 0 1]
                              if ((submatrix(randomNr1(1),randomNr2(1))== ...
                                          submatrix(randomNr1(2),randomNr2(2))) && ...
                                          (submatrix(randomNr1(1),randomNr2(2))== ...
                                          submatrix(randomNr1(2),randomNr2(1))) && ...
                                          (submatrix(randomNr1(1),randomNr2(1))~= ...
                                          submatrix(randomNr1(1),randomNr2(2))))
                                    % perform swap
                                    submatrix(randomNr1(1),randomNr2(1)) = ...
                                          abs(submatrix(randomNr1(1),randomNr2(1))-1);
                                    submatrix(randomNr1(1),randomNr2(2)) = ...
                                          abs(submatrix(randomNr1(1),randomNr2(2))-1);
                                    submatrix(randomNr1(2),randomNr2(1)) = ...
                                          abs(submatrix(randomNr1(2),randomNr2(1))-1);
                                    submatrix(randomNr1(2),randomNr2(2)) = ...
```

```
                    abs(submatrix(randomNr1(2),randomNr2(2))-1);
                aantalSwaps=aantalSwaps+1;
            end
        end         %if
    end         %1000 swap attempts
end
% repair the submatrix by adding absent species in submatrix
Mr=1; NN=1;
submatrix2=[];
for y=1:Y
    if kolomleeg(y,1)==1
        submatrix2(1:aantalReps,y)=zeros(aantalReps,1);
    else submatrix2(1:aantalReps,y)=submatrix(1:aantalReps,NN);
        NN=NN+1;
    end
end
clear kolomleeg
if NN-1~=B
    probleem=1;
 end
% put the swapped replicates back in the overall matrix
matrixNull(rijen(n):(rijen(n)+aantalReps-1),:)=submatrix2;
clear submatrix2; clear submatrix;
end

%% calculate the indices on the randomised data matrix
    [X,Y]=size(matrixNull); N=1; x=1;
    Comb=Y*(Y-1)/2; % number of species pairs
% C-score
    NCU=1; CU=0; Tsc=0; Scr=0; checker=0;
    for y1=1:(Y-1)              % species y1
    for y2=(y1+1):Y   % species y2 (different from species y1)
        % C
        ri=sum(matrixNull(:,y1));
        rj=sum(matrixNull(:,y2));
            % calculate how many times species y1 and y2 appear together in a
            % replicate (S)
            som=matrixNull(:,y1)+matrixNull(:,y2);
            verschil=abs(matrixNull(:,y1)-matrixNull(:,y2));
            S=sum(som-verschil)/2;
            CU=CU+(ri-S)*(rj-S);
            Tsc=Tsc + S*(aantalRep+S-ri-rj);        %T
            Scr=Scr+S;                              %S
             if max((som))==1                       %checker
            checker = checker+1;
            end
            NCU=NCU+1;
```

```
            end
            end
            Cscore=CU/(Y*(Y-1)/2);
        Tscore=Tsc/(Y*(Y-1)/2);
        Sscore=Scr/(Y*(Y-1)/2);
        % V-score
        ST2=0;
        Tgem=sum(sum(matrixNull))/aantalRep;
            for j=1:aantalRep
            Tj=sum(matrixNull(j,:));
            ST2=ST2+(Tj-Tgem)^2;
            end
        ST2=ST2/aantalRep;
        sigmai2=0;
        for i=1:aantalSpec
            ni=sum(matrixNull(:,i));
            sigmai2=sigmai2+(ni/aantalRep)*(1-(ni/aantalRep));
        end
        Vscore=ST2/sigmai2;
        clear matrixNull
        resultaat_null(rand,1)=Cscore;
        resultaat_null(rand,2)=Tscore;
        resultaat_null(rand,3)=Sscore;
        resultaat_null(rand,4)=Vscore;
        resultaat_null(rand,5)=checker;
    end
    fileN1=['resultaat_STCV_null_Swap1_1A_' aant];
    save(fileN1,'resultaat_null')
    clear
end
```

Swap 2 for presence/absence data

In this paragraph only the code about the swapping algorithm is shown. The rest of the code is equal to what is shown above.

```
[A,B]=size(submatrix);
swaps=1000;
sw=0;
aantalSwaps=0;
if A>1 && B>1
    while sw<swaps
        % choose randomly 2 replicates and 1 species
        randomNr1 = randint(2,1,[1,A]);
        randomNr2 = randi(B);
        % check if it concerns 2 different replicates
```

```
        if (randomNr1(1)~= randomNr1(2))
            sw = sw+1;
            w1=submatrix(randomNr1(1),randomNr2);
            w2=submatrix(randomNr1(2),randomNr2);
            % perform swap
             submatrix(randomNr1(1),randomNr2) = w2;
            submatrix(randomNr1(2),randomNr2) = w1;
            % check if no empty replicates are present
            % if one of the replicates contains no species, restore the submatrix
            if sum(submatrix(randomNr1(1),:))==0 || ...
                sum(submatrix(randomNr1(2),:))==0
                submatrix(randomNr1(1),randomNr2) = w1;
                submatrix(randomNr1(2),randomNr2) = w2;
            end
            clear w1; clear w2;
        end             % if
    end                 % 1000 swap attempts
end
```

Histograms

```
groep='1A'                          % define feeding type
Nrandom=100;                        % number of null models in one file
NrandomTot=1000;                    % total number of null models
fileN1=['resultaat_ori_' groep '.mat'];     % values of original data matrix
load(fileN1)

% load all values of null models
for aantal=1:10
    aant=int2str(aantal)
    fileN1=['result_random_' aant '_' groep];
    load(fileN1)
    resultaat((aantal-1)*Nrandom+1:aantal*Nrandom,:)=resultaat_null;
    clear resultaat_null
end

fid = fopen('labels_hor_as.txt');
labelslijst = textscan(fid,'%s');       % names of labels of x-axis
fclose(fid);

% calculate the two-sided 95% confidence interval of the 1000 null models
quant97=ceil(NrandomTot*97.5/100);
[X,Y]=size(resultaat);
resultaat(1000,:)=[];
for kolom=1:Y
    B=sort(resultaat(:,kolom));
    grens(kolom,1)=B(quant97);
```

```
    clear B
end
for kolom=1:Y
    B=sort(resultaat(:,kolom),'descend');
    grens2(kolom,1)=B(quant97);
    clear B
end

for kolom=1:Y
    % define minimum and maximum value of x-axis
    maximum=max(resultaat(:,kolom));
    minimum=min(resultaat(:,kolom));
    if resultaat_ori(kolom)>maximum
            maximum=resultaat_ori(kolom);
    end
    if resultaat_ori(kolom)<minimum
            minimum=resultaat_ori(kolom);
    end
    maximum=(maximum-minimum)/20 + maximum;
    minimum= -(maximum-minimum)/20 + minimum;
    hist(resultaat(:,kolom));          % plot histogram
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.7,0.7,0.7],'EdgeColor',[0.3,0.3,0.3])
    h2 = findobj(gca,'Type','axes');
    set(h2,'FontSize',30)
    set(gca,'xlim',[minimum maximum]);
    hold on
    x=resultaat_ori(kolom)*ones(1,11);
    y=0:20:200
    x2=grens(kolom,1)*ones(1,11);
    y2=0:20:200
    x3=grens2(kolom,1)*ones(1,11);
    y3=0:20:200
    % plot original value on histogram
    plot(x,y,'color','k','LineWidth',3)
    hold on
    % plot upper border of 95% CI on histogram
    plot(x2,y2,'color','k','LineWidth',3,'LineStyle','--')
    hold on
    % plot lower border of 95% CI on histogram
    plot(x3,y3,'color','k','LineWidth',3,'LineStyle','--')
    xlabel(labelslijst{1,1}{kolom,1},'FontSize',38)
    ylabel('Frequency','FontSize',36)
    fname=['hist_dens_' int2str(kolom) '_' groep '_190411.jpg' ];
    print('-r100','-djpeg',fname)
    hold off
end
```

CODE RELATED TO CHAPTER 3: ARTIFICIAL NEURAL NETWORKS

Find optimal number of neurons

The optimal network transfer functions and learning methods where at this point already defined. Finding the optimal total network can be realised by adding two more loops (one for the transfer function and one for the learning method) within the neuron loop.

```
%% calculate neural networks for different neurons
% samples with geographical coordinates (replicate ID, latitude, longitude)
stations = load('stations_011007_4.csv');
% data matrix with environmental variables in columns,replicates in rows
Pfile = load('omgeving011007_4.csv');
% target: target values (biodiversity indices) in columns, replicates in rows
Tfile = load('target_compl_sampleIndependent_011007_4.csv');
% file with stratified data: column 8 assigns fold for replicate
fanny=load('fanny_clustering_011007_4.csv');
[Pr,Pk]=size(Pfile);
coeff = 0.5;        % minimum correlation coefficient to retain neural network
neuronStart=1; neuronInterval=1; neuronEind=6;
datum=date;             % date is used in file name
NZdata=Pfile;
[Ta1,Ta2]=size(Tfile);

% Preprocess environmental variables: mean = 0 and standard deviation = 1
P1 = NZdata.';
[P2,P2info] = mapstd(P1);
% reduce number of environmental variables by principal component analysis
% keep only those variables which contribute more than 1% of the variation % in the data
[P2,P2infoPCA] = processpca(P2,0.01);
[P2R,P2K]=size(P2);
 Tst=Tfile.';
folds=10;
for index=1:Ta2                     % find neural networks for every target (diversity index)
    indx=int2str(index);
    N=1;
    result=zeros(1000,26);
    T1=Tst(index,:);
    ind=int2str(index);
    [T2,T2info] = mapminmax(T1);            % transform target to interval [-1 1]
    for neuron = neuronStart:neuronInterval:neuronEind      % find # neurons
            for fold=1:folds        % tenfold cross-validation
            fld=int2str(fold);
            clear P; clear M; clear val; clear T; clear rij; clear test;
            P=P2; T=T2; M=1; Mt=1;
            % split up data in three sets: test, validation & training set
```

```
for rij=Pr:-1:1
      if fanny(rij,9)==fold
            val.P(:,M)=P2(:,rij);
            P(:,rij)=[];
             val.T(1,M)=T(1,rij);
            T(:,rij)=[];
            M=M+1;
            elseif fanny(rij,9)==fold+1 || (fold==10 && fanny(rij,9)==1)
            test.P(:,Mt)=P2(:,rij);
            P(:,rij)=[];
            test.T(1,Mt)=T(1,rij);
            T(:,rij)=[];
            Mt=Mt+1;
            end
end
for keer=1:20                    % construct 20 neural networks
            kr=int2str(keer);
            % NET = NEWFF creates a new network: first transfer function = tansig
            % transfer function of ultimate layer = purelin
            net1=newff(minmax(P), [neuron, 1],{'tansig', 'purelin'},'trainlm');
            net1.trainParam.goal=0.0005;
            net1.trainParam.max_fail=10;
            % train network
            [net, tr] = train(net1, P, T,[],[],val,test);
            % simulate data with network
            aL=sim(net,P);
            % backtransform output of network to original range
            aL=mapminmax('reverse',aL,T2info);
            TL=mapminmax('reverse',T,T2info);
            % calculate performance parameters for training data
            RL= corrcoef(aL,TL);
            rSpearmanL = corr(aL.',TL.','type','spearman');
            nL = length(aL);
            RMSEL= sqrt((1/nL)*sum((aL-TL).^2));
            RMSELgem=sqrt((1/nL)*sum((aL-TL).^2))/mean(T1);
            RSEL= sqrt(sum((aL-TL).^2)/sum((TL-mean(T1)).^2));
            EL= (1/nL)*sum((aL-TL).^2);

      % calculate performance parameters for independent test data
            aT=sim(net,test.P);
            aT=mapminmax('reverse',aT,T2info);
            testT=mapminmax('reverse',test.T,T2info);
            RT= corrcoef(aT,testT);
            rSpearmanT = corr(aT.',testT.','type','spearman');
            nT = length(aT);
            RMSET= sqrt((1/nT)*sum((aT-testT).^2));
            RMSETgem=sqrt((1/nT)*sum((aT-testT).^2))/mean(T1);
```

```
                RSET= sqrt(sum((aT-testT).^2)/sum((aT-mean(T1)).^2));
                ET= (1/nT)*sum((aT-testT).^2);

                % calculate performance parameters for all data
                aAlles=sim(net,P2);
                aAlles=mapminmax('reverse',aAlles,T2info);
                Ralles= corrcoef(aAlles,T1);
                rSpearmanalles = corr(aAlles.',T1.','type','spearman');
                nalles = length(aAlles);
                RMSEalles= sqrt((1/nalles)*sum((aAlles-T1).^2));
                RMSEallesgem=sqrt((1/nalles)*sum((aAlles-T1).^2))/mean(T1);
                RSEalles= sqrt(sum((aAlles-T1).^2)/sum((aAlles-mean(T1)).^2));
                Ealles= (1/nalles)*sum((aAlles-T1).^2);

                % test normality of residuals
                residuelen=aAlles-T1;
                [hL pL lL cL]=lillietest(residuelen);
                if lL>cL
                NormResi=0;
                else NormResi=1;
                end
                result(N,1)=index;          result(N,2)=neuron;
                result(N,3)=fold;               result(N,4)=keer;
                result(N,5)=RL(2,1);        result(N,6)=rSpearmanL;
                result(N,7)=RMSEL;          result(N,8)=RMSELgem;
                result(N,9)=RSEL;           result(N,10)=EL;
                result(N,11)=RT(2,1);       result(N,12)=rSpearmanT;
                result(N,13)=RMSET;         result(N,14)=RMSETgem;
                result(N,15)=RSET;          result(N,16)=ET;
                result(N,17)=Ralles(2,1);       result(N,18)=rSpearmanalles;
                result(N,19)=RMSEalles;         result(N,20)=RMSEallesgem;
                result(N,21)=RSEalles;      result(N,22)=Ealles;
                result(N,23)=NormResi;          result(N,24)=cL;
                result(N,25)=lL;                result(N,26)=pL;
                N=N+1;
                % save the network for later use
                nm=['Res\netT_' datum '_K' kr '_F' fld '_T' indx '.mat'];
                save(nm,'net');
                clear net, clear tr, clear net1, clear av,
            clear Rv, clear nv
        end             % keer
        end                 % fold
end                     % neuron
name = [Res\result' indx '_' datum];
save(name,'result')
clear result; clear T1; clear ind; clear T1tr; clear transformatie;
clear indexTr; clear T; clear T2info;
```

```
end       % target
```

Perturb

```
stations = load('stations.csv');  % load station data
Pfile = load('omgeving.csv');         % load environmental data
Tfile = load('target.csv');         % load target data
neuron=2; folds=10;               % best model has 2 neurons & 10-fold cross-validation was
applied
intervallen=20;                   % number of intervals considered in variable contribution
Bestes=10;                        % Find the ten best models
NZdata=Pfile;
[NZ1,NZ2]=size(NZdata);
P1 = NZdata.';
 [P2,P2info] = mapstd(P1);            % preprocess environmental variables
[P2,P2infoPCA] = processpca(P2,0.01);
targets=16;
for target=1:targets
      target
      ind=int2str(target);
      trgt=int2str(target);
      datumF='24-Oct-2007';
      name=['Res\result_T' trgt '_N2_' datumF];
      load(name);
      clear name;
      temp=result;
      temp(:,36)=temp(:,13);
      % find the ten best neural networks based on RMSE of test set
      [Te1,Te2]=size(temp);
      for fold=1:10
                  min=10000;
                  for te1=1:Te1
                  if temp(te1,3)==fold && temp(te1,36)<min
                              min=temp(te1,36);
                              besteNN(fold,:)= temp(te1,:);
                  end
                  end
      end

      % add increasing levels of noise to the variable of interest and monitor
      % the effect on the output
      for fold=1:10
            beste=1;
            bst=int2str(beste);
            keer=besteNN(fold,4);
            fld=int2str(fold);
            krOK=int2str(keer);
            naamN=['Res\netT_' datumF '_K' krOK '_F' fld '_T' trgt '_N2'];
```

```
        load(naamN)
        clear naamN;
        Tst=Tfile.';
        T1=Tst(target,:);
        [T2,T2info] = mapminmax(T1);
        for nz2=1:NZ2                % do this loop for every environmental variable
            NZdata=Pfile;
            variabele=NZdata(:,nz2);
            MIMA=minmax(variabele.');
            % add different levels of noise to the environmental variable under
            % consideration
            for noise=1:(intervallen+1)
                    ruis=randn(NZ1,1).';
                    [ruis2,ruisInfo]=mapminmax(ruis);
                    ruis2T=ruis2.';
                var4=variabele+ruis2T*(MIMA(2)-MIMA(1))*(noise-1)/(2*intervallen);
                    var4=mapminmax(var4.',MIMA(1),MIMA(2)).';
                    NZdata(:,nz2)=var4;
                    P1 = NZdata.';
                    P2 = mapstd('apply',P1,P2info);
                    P2=processpca('apply',P2,P2infoPCA);
                    % simulate output with altered variables data
                    a=sim(net,P2);
                    % backtransform
                    a=mapminmax('reverse',a,T2info);
                    Rt= corrcoef(a,T1);
                    rSpearman = corr(a.',T1.','type','spearman');
                    nt = length(a);
                    % calculate root mean squared error on output
                    RMSEt= sqrt((1/nt)*sum((a-T1).^2));
                    resultNoise(target,nz2,noise,fold)=RMSEt;
            end
            end
    end
end
datum=date;
naamR=['Res\resultNoise_somVLT_10folds_N2' datum];
naamB=['Res\besteNN_somVLT_10folds_N2' datum];
save(naamR, 'resultNoise')
save(naamB, 'besteNN')
datum='06-Nov-2007';
naamR=[Res\resultNoise_somVLT_10folds_N2' datum];
load(naamR)

[P1,P2]=size(Pfile);
[T1,T2]=size(Tfile);
[N1,N2]=size(resultNoise);
```

```
 % calculate the relative increase in RMSE by adding noise to each
% environmental variable
for target=1:targets                    % loop for targets (biodiversity indices)
    for noise=1:(intervallen+1)             % loop for level of noise
            for nz2=1:NZ2                       % loop for each environmental variable
            for fold=1:10               % loop for each fold
                matrix4(target,nz2,noise,fold)= (resultNoise(target,nz2,noise,fold)- ...
                resultNoise(target,nz2,1,fold))*100/resultNoise(target,nz2,1,fold);
            end
            end
    end
end

% calculate the average increase for each target and each variable
for target=1:targets
    for var=1:NZ2
        kolom=1
        for fold=1:10
                temp(1:5,kolom)=matrix4(target,var,17:21,fold);
                kolom=kolom+1;
        end
        [Te1,Te2]=size(temp);
        temp2 = reshape(temp,Te1*Te2,1);
        gemiddelde(var,target)=mean(temp2);
        standdev(var,target)=std(temp2);
        clear temp; clear temp2;
    end
end
ondergrens=gemiddelde-standdev;
for target=1:targets
    somT=sum(gemiddelde(:,target));
    gemiddeldeP(:,target)=gemiddelde(:,target)*100/somT;
    standdevP(:,target)=standdev(:,target)*100/somT;
    clear temp; clear temp2;
end
ondergrensP=gemiddeldeP-standdevP;
save('Res\SENSperturb_LVT_B10F_gemP','gemiddeldeP')
save('Res\SENSperturb_LVT_B10F_stdP','standdevP')
```

Profile

The first part of the code is identical to the first part of the previous paragraph and will not be repeated here.

```
% load data
% preprocess environmental variables
 [V1,W1]= size(P1);
```

```
perc = 100*(0:0.2:1);       % define percentile values
[PE1,PE2]=size(perc);
for v1=1:V1                 % Calculate percentile values for each environmental variable
    percentiles = prctile(P1(v1,:),perc);
    waarden(v1,1:PE2)=percentiles;
end
profile=zeros(NZ2,abN+1,targets,Bestes);
for target=1:targets
    target
    ind=int2str(target);
    trgt=int2str(target);
    datumF='24-Oct-2007';
    % load result file from previous analysis (first paragraph)
    name=['Res\result_T' trgt '_N2_' datumF];
    load(name);
    clear name;
    [Res1,Res2]=size(result);
    temp=result;
    temp(:,36)=temp(:,13)
     % find the ten best neural networks based on RMSE of test set
    [Te1,Te2]=size(temp);
    for fold=1:10
            min=10000;
            for te1=1:Te1
            if temp(te1,3)==fold && temp(te1,36)<min
                min=temp(te1,36);
                    besteNN(fold,:)= temp(te1,:);
            end
            end
    end
    for fold=1:Folds
        [RE1,RE2] =size(temp);
        keer=besteNN(fold,4);
        fld=int2str(fold);
        krOK=int2str(keer);
        % load the network corresponding with the best result
        naamN=['Res\netT_' datumF '_K' krOK '_F' fld '_T' trgt '_N2'];
        load(naamN)
        clear naamN;
        Tst=Tfile.';
        T1=Tst(target,:);
         [T2,T2info] = mapminmax(T1);
        [V,W]= size(P1);
        % create dataset where 1 environmental variable changes
            for vari=1:V        % do this for all environmental variables
            m=waarden(vari,1);
            M=waarden(vari,PE2);
```

```
            abI=(M-m)/abN;
            for interv=1:(abN+1)  % do this for 'abN' intervals
                  NZab=waarden;
                  Interval=m+abI*(interv-1);
                  % keep percentile values of all environmental variables
                  % replace data of 1 variable by constant value
                  NZab(vari,:)=Interval*ones(1,PE2);
                  NZab = mapstd('apply',NZab,P2info);
                  NZab = processpca('apply',NZab,P2infoPCA);
                  % calculate output for altered environmental variables matrix
                  aNZab=sim(net,NZab);
                  aNZ=mapminmax('reverse',aNZab,T2info);
                  profile(vari,interv,target,fold)=mean(aNZ);
            end         %intervals
            end               % variables
      end
end                     %target
save('Res\SENSprofile_10BF','profile')
load('Res\SENSprofile_10BF')
% calculate influence of environmental variables
for target=1:targets
      for beste=1:Bestes
            prof=profile(:,:,target,beste);
            extremen=minmax(prof);
            verschil=extremen(:,2)-extremen(:,1);
            resultaat(:,target,beste)=verschil;
            clear verschil;
      end
end
save('Res\SENSprofile_10BF_res','resultaat')
load('Res\SENSprofile_10BF_res')

% convert result to percentages
somK=sum(resultaat,1)
for target=1:targets
      for beste=1:10
            resultaatP(:,target,beste)=resultaat(:,target,beste)*100./somK(1,target,beste);
      end
end
gemiddeld=mean(resultaatP,3);
standdev=std(resultaatP, 0, 3);
ondergrens=gemiddeld-standdev;
[PR1,PR2,PR3,PR4]=size(profile);

% check if influence of environmental variable on diversity index is positive or negative
for target=1:targets
      for beste=1:10
```

```
            prof=profile(:,:,target,beste);
        extremen=minmax(prof);
            for vari=1:PR1
            MAX=NaN; MIN=NaN;
            for interv=1:(abN+1)
                    if profile(vari,interv,target,beste)==extremen(vari,1)
                    MIN=interv;
                    elseif profile(vari,interv,target,beste)==extremen(vari,2)
                    MAX=interv;
                    end
            end
            if MAX - MIN<0
                posneg(vari,target,beste)=-1;
            elseif MAX - MIN>0
                posneg(vari,target,beste)=1 ;
            else posneg(vari,target,beste)=NaN;
            end
            end
    end
end
save('Res\SENSprofile_10BF_posneg','posneg')

for vari=1:V1
    for target=1:targets
            result10B(vari, target,1)=mean(resultaat(vari,target,:));
            result10B(vari, target,2)=std(resultaat(vari,target,:));
    end
end
save('Res\SENSprofile_10BF_gemiddelde','result10B')

for vari=1:V1
    for target=1:targets
            resultPN10B(vari, target,1)=mean(posneg(vari,target,:));
            resultPN10B(vari, target,2)=std(posneg(vari,target,:));
    end
end
save('Res\SENSprofile_10BF_posnegGem','resultPN10B')
```

Modified Profile

In this paragraph only the part where the code differs from the original 'Profile method' is shown.

```
% check if influence of environmental variable on diversity index
% is positive or negative
% create environmental dataset where 1 environmental variable changes
% but the other variables are kept at their original value
```

```
for vari=1:V        % do this for all environmental variables
    mM=minmax(P1(vari,:));
    abI=(mM(2)-mM(1))/abN;
    for interv=1:(abN+1)                    % do this for 'abN' intervals
            NZab=P1;
            Interval=mM(1)+abI*(interv-1);
            NZab(vari,:)=Interval*ones(W,1);        % change the values of 1
                                                    % environmental variable
            NZab = mapstd('apply',NZab,P2info);
            NZab = processpca('apply',NZab,P2infoPCA);
            aNZab=sim(net,NZab);
            aNZ=mapminmax('reverse',aNZab,T2info);
            waarde(vari,interv,fold,target)=mean(aNZ);
    end         %intervals
    end             % variables
```

CODE RELATED TO CHAPTER 4: GEOSTATISTICS

In this paragraph only the R-code for the generalised least squares model is shown. More Matlab code is developed to transform output data of the geostatististical models to arcgis maps, but this is not shown in this chapter.

Generalised least squares model

This paragraph contains R-code instead of Matlab Code!

```
rm(list = ls())
getwd()
setwd("D:\\Bmerckx\\MyMatlab\\kaart_MM_SB\\meio_art")
library(nlme)
meio=read.table("rekenR2.txt", header=TRUE)          # load training data
validatie=read.table("valR.txt",header=TRUE)         # load validation data
dimensies=dim(meio)
meio2=scale(meio[,1:(dimensies[2])])                 # rescale both matrices
dimensies2=dim(validatie)
center2=mat.or.vec(dimensies2[1], dimensies2[2])
scale2=mat.or.vec(dimensies2[1], dimensies2[2])
for (rij in 1:dimensies2[1]){
    center2[rij,]=attr(meio2,"scaled:center")[1:dimensies2[2]]
    scale2[rij,]=attr(meio2,"scaled:scale")[1:dimensies2[2]]
    validatie2=(validatie - center2)/scale2
}

meio2=as.data.frame(meio2)
attach(meio2)
meio.train=meio2
```

```
# choose which column to model
index = 2
divIndex= meio.train[,11+index]

## before variable selection
d502 <- d50^2
mud2 <- mud^2
TSM_max2 <- TSM_max^2
TSM_min2 <- TSM_min^2
TSM_mean2 <- TSM_mean^2
chl_max2 <- chl_max^2
chl_min2 <- chl_min^2
chl_mean2 <- chl_mean^2
diepte2 <- diepte^2
meio.gls=gls(divIndex~ E+N+d50+mud+TSM_mean+TSM_max+TSM_min+chl_mean
      + chl_max+chl_min+diepte+year+d502+mud2+TSM_mean2
      + TSM_max2+TSM_min2+chl_mean2+chl_max2+chl_min2+diepte2
      + d50:mud + d50:TSM_mean + d50:TSM_min + d50:chl_mean
      + d50:chl_min + d50:diepte + mud:TSM_mean
      + mud:TSM_min + mud:chl_mean + mud:chl_min + mud:diepte,
      cor=corSpher(form= ~EL_UTMuniek+NB_UTMuniek),data=meio.train)
summary(meio.gls)

# after variable selection
mud2=mud^2
meio.gls=gls(divIndex~ mud + TSM_mean + mud2,
      cor=corSpher(form= ~EL_UTMuniek+NB_UTMuniek),data=meio.train)
summary(meio.gls)

# residual analysis
er=resid(meio.gls)
er[abs(er)>3*sd(er)]
e1=er[abs(er)<3*sd(er)]
shapiro.test(e1)

## validation
attach(meio2)
meio2$mud2=meio2$mud^2
meio.lm.predR=predict(meio.gls,newdata=meio2)
meio.lm.residR=residuals(meio.gls,newdata=meio2)
attach(validatie2)
validatie2$mud2=validatie2$mud^2
meio.lm.predV=predict(meio.gls,newdata=validatie2)
meio2=scale(meio[,1:(dimensies[2]-1)])
predictionV=
```

```
    meio.lm.predV*attr(meio2,"scaled:scale")[index+11]+attr(meio2,"scaled:center")[index
+11]
predictionR=
    meio.lm.predR*attr(meio2,"scaled:scale")[index+11]+attr(meio2,"scaled:center")[index
+11]
write.csv(predictionV, "meioGlsV_ES25.csv")
write.csv(predictionR, "meioGlsR_ES25.csv")

# calculate quality parameters of prediction
meio.lm.MSPE=mean((meio.lm.pred-validatie2[,11+index])^2)
meio.lm.MSPE
meio.lm.spear=cor(meio.lm.pred,validatie2[,11+index],method="spearman")
meio.lm.pearson= cor(meio.lm.pred,validatie2[,11+index],method="pearson")
meio.lm.spear
meio.lm.pearson
MSE[subst,index] = meio.lm.MSPE
spear[subst,index] = meio.lm.spear
pear[subst,index] = meio.lm.pearson
```

CODE RELATED TO CHAPTER 5: NULL MODELS MAXENT

In this paragraph only the code to create random subsets with a minimum distance between them will be given. The code concerning model selection in Maxent can be found in the next paragraph.

Create five subsets with a minimum distance between these subsets

```
% read station data with three columns:
   % speciesnumber  // Easting (UTM_coordinates) // Northing (UTM_coordinates)
lijst=csvread('maxent_lijst.csv');
fid = fopen('specieslist2.txt');        % read text file with 1 column: species names
soortenlijst = textscan(fid,'%s'); fclose(fid);
distances=[5000, 10000];            % choose for which distances you want to create subsets
NDist=length(distances); [X,Y]=size(lijst);
[Sp1,Sp2]=size(soortenlijst{1,1});
soortenlijst=soortenlijst{1,1};
for dist=1:NDist                    % create cross-validation files for different distances
    mindist=distances(dist);
    mdint=round(mindist/1000);
    md=int2str(mdint);
    for species=1:Sp1                   % create different  files for each species
            minATOK=0;
            N=1;
            tempSp=[];
        for rij=1:X                 % find data for each species
            if lijst(rij,1)==species
                    tempSp(N,:)=lijst(rij,:); N=N+1;
```

```
    end
end
stations=tempSp; [S1,S2]=size(stations); [T1,T2]=size(tempSp);
pogingen=0; geslaagd=0;
% create subsets with 1 element, which are at least 5 or 10 km separated from each
% other
temp1OK=[]; temp2OK=[]; temp3OK=[]; temp4OK=[]; temp5OK=[];
    while pogingen<1000 && geslaagd==0  % if necessary try 1000 times to make
                                        % subsets
    coord=tempSp(:,2:3);
    random2=randperm(T1);
    aantalSt=T1;
    temp1=[]; temp2=[]; temp3=[]; temp4=[]; temp5=[];
    aantal=1; reserve=0; aantalT=0;
    % create 5 subsets, each with 1 station and at least 'mindist' distance apart
    while aantal<6 && aantalT<T1
        aantalT=aantal+reserve;
        duo=coord(random2(aantalT),1:2);
            afstand1nieuw=0; afstand2nieuw=0;
        afstand3nieuw=0; afstand4nieuw=0; afstand5nieuw=0;
        success=0;
        if isempty(temp1)
            temp1(1,:)=duo; aantal=aantal+1; success=1;
        end
        afstand1nieuw = sqrt((temp1(1,1)-duo(1))^2+(temp1(1,2)-duo(2))^2);
        if isempty(temp2)&& ~isempty(temp1) && afstand1nieuw > mindist
            temp2(1,:)=duo; aantal=aantal+1; success=1;
        end
        if ~isempty(temp2)
            afstand2nieuw = sqrt((temp2(1,1)-duo(1))^2+(temp2(1,2)-duo(2))^2);
        end
        if isempty(temp3)&& ~isempty(temp2) && afstand2nieuw > …
            mindist && afstand1nieuw > mindist
            temp3(1,:)=duo;  aantal=aantal+1; success=1;
        end
        if ~isempty(temp3)
            afstand3nieuw = sqrt((temp3(1,1)-duo(1))^2+(temp3(1,2)-duo(2))^2);
        end
        if isempty(temp4)&& ~isempty(temp2) && ~isempty(temp3) && …
            afstand3nieuw > mindist && afstand2nieuw > mindist && ...
            afstand1nieuw > mindist
            temp4(1,:)=duo;  aantal=aantal+1; success=1;
        end
        if ~isempty(temp4)
            afstand4nieuw = sqrt((temp4(1,1)-duo(1))^2+(temp4(1,2)-duo(2))^2);
        end
        if isempty(temp5) && ~isempty(temp2) && ~isempty(temp3) && ...
```

```
            ~isempty(temp4) && afstand4nieuw > mindist && afstand3nieuw > ...
            mindist && afstand2nieuw > mindist && afstand1nieuw > mindist
            temp5(1,:)=duo;  aantal=aantal+1; success=1;
      end
      if success==0
            reserve=reserve+1;
      end
end
for t1=T1:-1:1          % delete the assigned stations from list
            duo=coord(t1,1:2);
            if isequal(temp1,duo) || isequal(temp2,duo) || …
            isequal(temp3,duo) || isequal(temp4,duo) || isequal(temp5,duo)
            coord(t1,:)=[];
            end
end
aantal=5; aantal2=1; reserve=0; aantalT=0; pogingsets=0;
r1=2; r2=2; r3=2; r4=2; r5=2;
random3=randperm(T1-5);
RG1=1;
reservegroep=[];
% add randomly stations to the 5 subsets while considering the minimum
% distance constraints
while aantal<=T1 && aantalT<T1-5
      aantalT=aantal2+reserve;
      duo=coord(random3(aantalT),1:2);  % choose a random pair from
                                        % coordinate list
      mdeerste(1:5)=ones(1,5)*10^20;
      success=0;
      [d11,d12]=size(temp1);
      % calculate distance
      for r=1:d11
            afstand=sqrt((temp1(r,1)-duo(1))^2+(temp1(r,2)-duo(2))^2);
            if afstand<mdeerste(1)
                  mdeerste(1)=afstand;
            end
      end
            [d21,d22]=size(temp2);
      for r=1:d21
            afstand=sqrt((temp2(r,1)-duo(1))^2+(temp2(r,2)-duo(2))^2);
            if afstand<mdeerste(2)
                  mdeerste(2)=afstand;
            end
      end
       [d31,d32]=size(temp3);
      for r=1:d31
            afstand=sqrt((temp3(r,1)-duo(1))^2+(temp3(r,2)-duo(2))^2);
            if afstand<mdeerste(3)
```

```
            mdeerste(3)=afstand;
        end
    end
    [d41,d42]=size(temp4);
    for r=1:d41
        afstand=sqrt((temp4(r,1)-duo(1))^2+(temp4(r,2)-duo(2))^2);
        if afstand<mdeerste(4)
                mdeerste(4)=afstand;
        end
    end
    [d51,d52]=size(temp5);
    for r=1:d51
        afstand=sqrt((temp5(r,1)-duo(1))^2+(temp5(r,2)-duo(2))^2);
        if afstand<mdeerste(5)
            mdeerste(5)=afstand;
        end
    end
    %for which subset is the minimum distance found
        Min1=min(mdeerste);
        Min2=10^10;
        for k=1:5
        if mdeerste(k)>Min1 && mdeerste(k)<Min2
                Min2=mdeerste(k);
                m2=k;
        elseif mdeerste(k)==Min1
                m1=k;
        end
    end
    tr=zeros(1,5);
    if Min1==0
        reserve=reserve+1;
            %  if the pair is distant enough from all sets,
            % assign to a back-up group
        elseif Min1>mindist
        reservegroep(RG1,1:2)=duo;
        RG1=RG1+1;
        aantal2=aantal2+1;
    % if the distance of newly chosen pair falls within 'mindist' range
    % from one subset, but is not too close to any other of the four sets
    %  then assign the pair
    elseif Min1<=mindist && Min1>0 && Min2>mindist
        if m1==1 && r1<(ceil(aantalSt/5) + 1)
            temp1(r1,:)=duo; r1=r1+1; success=1; aantal2=aantal2+1;
            elseif m1==2 && r2<(ceil(aantalSt/5) + 1)
            temp2(r2,:)=duo; r2=r2+1; success=1; aantal2=aantal2+1;
            elseif m1==3 && r3<(ceil(aantalSt/5) + 1)
            temp3(r3,:)=duo; r3=r3+1; success=1; aantal2=aantal2+1;
```

```
                elseif m1==4 && r4<(ceil(aantalSt/5) + 1)
                temp4(r4,:)=duo; r4=r4+1; success=1; aantal2=aantal2+1;
                elseif m1==5 && r5<(ceil(aantalSt/5) + 1)
                temp5(r5,:)=duo; r5=r5+1; success=1; aantal2=aantal2+1;
                else success=0;
                reserve=reserve+1;
                end
            elseif success==0;
            reserve=reserve+1;
            end
        end
        % assign the pair in the back-up group to the subset with the least data
        % and check if the 'mindist' condition still holds
        if ~isempty(reservegroep)
            [RGR,RGK]=size(reservegroep);
            for rijRG=1:RGR
                duo=reservegroep(rijRG,:);
                % calculate distance
                mdeerste(1:5)=ones(1,5)*10^20;
                [d11,d12]=size(temp1);
                for r=1:d11
                afstand=sqrt((temp1(r,1)-duo(1))^2+(temp1(r,2)-duo(2))^2);
                if afstand<mdeerste(1)
                    mdeerste(1)=afstand;
                end
                end
            [d21,d22]=size(temp2);
            for r=1:d21
                afstand=sqrt((temp2(r,1)-duo(1))^2+(temp2(r,2)-duo(2))^2);
                if afstand<mdeerste(2)
                    mdeerste(2)=afstand;
                end
            end
            [d31,d32]=size(temp3);
            for r=1:d31
                afstand=sqrt((temp3(r,1)-duo(1))^2+(temp3(r,2)-duo(2))^2);
                if afstand<mdeerste(3)
                    mdeerste(3)=afstand;
                end
            end
            [d41,d42]=size(temp4);
            for r=1:d41
                afstand=sqrt((temp4(r,1)-duo(1))^2+(temp4(r,2)-duo(2))^2);
                if afstand<mdeerste(4)
                        mdeerste(4)=afstand;
                end
            end
```

```
[d51,d52]=size(temp5);
for r=1:d51
    afstand=sqrt((temp5(r,1)-duo(1))^2+(temp5(r,2)-duo(2))^2);
    if afstand<mdeerste(5)
        mdeerste(5)=afstand;
    end
end
% find subset where the minimum distance is found
    Min1=min(mdeerste);
    Min2=10^10;
    for k=1:5
    if mdeerste(k)>Min1 && mdeerste(k)<Min2
            Min2=mdeerste(k);
            m2=k;
    elseif mdeerste(k)==Min1
            m1=k;
    end
    end

    % check the number of data in each subset
    aantalEl(1,1)=numel(temp1)/2; aantalEl(2,1)=numel(temp2)/2;
aantalEl(3,1)=numel(temp3)/2; aantalEl(4,1)=numel(temp4)/2;
aantalEl(5,1)=numel(temp5)/2;
    aantalEl(1:5,2)=1:5;
    aantalEl=sortrows(aantalEl,1);
% if the pair is distant enough from all sets, assign to the subset
% with the least data
if (Min1>mindist && Min1>0)
        nummer=aantalEl(1,2);
        switch nummer
        case 1
                temp1(d11+1,:)=duo;
        case 2
                temp2(d21+1,:)=duo;
        case 3
                temp3(d31+1,:)=duo;
        case 4
                temp4(d41+1,:)=duo;
        case 5
                temp5(d51+1,:)=duo;
        end
% if the distance of newly chosen pair falls within 'mindist' range
% from one subset, but is not too close to any other of the
% four sets then assign the pair
elseif (Min1<=mindist && Min1>0 && Min2>mindist)
        switch m1
        case 1
```

```
                        temp1(d11+1,:)=duo;
                    case 2
                        temp2(d21+1,:)=duo;
                    case 3
                        temp3(d31+1,:)=duo;
                    case 4
                        temp4(d41+1,:)=duo;
                    case 5
                        temp5(d51+1,:)=duo;
                    end
                end
            end
        end
        % check how many pairs have been assigned in total
        AT(1)=numel(temp1)/2; AT(2)=numel(temp2)/2;   AT(3)=numel(temp3)/2;
        AT(4)=numel(temp4)/2; AT(5)=numel(temp5)/2;
        minAT=min(AT);
        % if more data have been assigned to the 5 subsets then during the previous

        % attempt, then remember the new subsets
        if minAT>minATOK
            temp1OK=temp1; temp2OK=temp2; temp3OK=temp3;
                temp4OK=temp4; temp5OK=temp5; minATOK=minAT;
        end
        pogingen=pogingen+1;
    end

        clear temp1; clear temp2; clear temp3; clear temp4; clear temp5;
        temp1=temp1OK;      temp2=temp2OK; temp3=temp3OK; temp4=temp4OK;
    temp5=temp5OK;
        AT(1)=numel(temp1)/2; AT(2)=numel(temp2)/2; AT(3)=numel(temp3)/2;
    AT(4)=numel(temp4)/2; AT(5)=numel(temp5)/2;
        minAT=min(AT);
    geselecteerd=(numel(temp1)+numel(temp2)+numel(temp3)+numel(temp4)+...
        numel(temp5))/2;
    if pogingen==1000 && minATOK>1
        geslaagd=1;
        else geslaagd=0;
        end
        MiMa=minmax(AT);
        % subsets should have the same number of data ±1
        if MiMa(2)-MiMa(1)>1
        while numel(temp1)/2> MiMa(1)+1;
            RP=randperm(numel(temp1)/2);
            temp1(RP(1),:)=[];
        end
        while numel(temp2)/2> MiMa(1)+1;
```

```
        RP=randperm(numel(temp2)/2);
        temp2(RP(1),:)=[];
    end
    while numel(temp3)/2> MiMa(1)+1;
        RP=randperm(numel(temp3)/2);
        temp3(RP(1),:)=[];
    end
    while numel(temp4)/2> MiMa(1)+1;
        RP=randperm(numel(temp4)/2);
        temp4(RP(1),:)=[];
    end
    while numel(temp5)/2> MiMa(1)+1;
        RP=randperm(numel(temp5)/2);
        temp5(RP(1),:)=[];
    end
end
% save the 5 cross-validation files in .csv-format
    if geslaagd==1
    randSt=randperm(ceil(aantalSt/5));
    rand5=randperm(5); tllr=1;
    VR(1)=numel(temp1)/2; VR(2)=numel(temp2)/2; VR(3)=numel(temp3)/2;
    VR(4)=numel(temp4)/2; VR(5)=numel(temp5)/2;
    ST=0; VRS(1)=0;
    for su=1:5
        ST=VR(su)+ST;
        VRS(su+1)=ST;
    end
    temp=[];
    temp((VRS(1)+1):VRS(2),:)=temp1; temp((VRS(2)+1):VRS(3),:)=temp2;
    temp((VRS(3)+1):VRS(4),:)=temp3; temp((VRS(4)+1):VRS(5),:)=temp4;
    temp((VRS(5)+1):VRS(6),:)=temp5;
    AS=int2str(aantalSt);
    [T1,T2]=size(temp);
    for sub=1:5
        subs=int2str(sub);
        spec_name=[soortenlijst{species,1} '_' subs];
        spec_nameF=[soortenlijst{species,1}];
        testset(1:VR(sub),:)=temp((VRS(sub)+1):VRS(sub+1),:);
        tempx=temp;
        for rijx=VRS(sub+1):-1:(VRS(sub)+1)
            tempx(rijx,:)=[];
        end
        rekenset=tempx;
        file_name_test=['crossfilesAC_km' md '\' spec_name '_test' '.csv'];
        lijn1='species, dd long, dd lat';
        dlmwrite(file_name_test, lijn1, 'delimiter','');
        [X1,X2]=size(testset);
```

```
                    for rijT=1:X1
                        long=num2str(testset(rijT,1));
                        lat=num2str(testset(rijT,2));
                        lijn2=[spec_name ',' long ',' lat];
                        dlmwrite(file_name_test, lijn2, 'delimiter','','newline', 'pc', '-append');
                        clear lijn2; clear long; clear lat;
                    end
                    clear file_name_test
                    file_name_reken=['crossfilesAC_km' md '\' spec_name '_reken' '.csv'];
                    dlmwrite(file_name_reken, lijn1, 'delimiter','');
                    clear lijn1
                    [Y1,Y2]=size(rekenset);
                    for rijR=1:Y1
                        long=num2str(rekenset(rijR,1));
                        lat=num2str(rekenset(rijR,2));
                        lijn2=[spec_name ',' long ',' lat];
                        dlmwrite(file_name_reken, lijn2, 'delimiter','', 'newline', …
                        'pc', '-append');
                        clear lijn2; clear long; clear lat;
                    end
                    clear file_name_reken; clear spec_name; clear spec_nameF; clear subs;
                    clear testset; clear rekenset;
                end
            end
                lijstac(species,1)=species;
                lijstac(species,2)=geslaagd;
                if geslaagd
                lijstac(species,3)=sum(VR); lijstac(species,4)=VR(1); lijstac(species,5)=VR(2);
                lijstac(species,6)=VR(3); lijstac(species,7)=VR(4); lijstac(species,8)=VR(5);
                else
                lijstac(species,3)=0; lijstac(species,4)=0; lijstac(species,5)=0; lijstac(species,6)=0;
                lijstac(species,7)=0; lijstac(species,8)=0;
            end
        end
    end
```

CODE RELATED TO CHAPTER 6: FIND OPTIMAL HSM FOR DIFFERENT DENSITIES

In this paragraph the code for the forward selection of features of the Maxent models in shown.

```
% The only parameter which needs adjustment after one loop of variable selection is
completed is
% 'Reeks'.
```

```
Reeks=1;
varselectie=[int2str(Reeks) 'varFW'];
Nspecies=6; Nfeatures=5; TotaalAantalVar=9;
NrVar=Reeks+1;
fid = fopen('omgeving_SB.txt');
maxent_omgeving_asc = textscan(fid,'%s');
fclose(fid);

% open text file with one column: the species names
fid = fopen('specieslist.txt');
soortenlijst = textscan(fid,'%s');
fclose(fid);
teller=1;

% open text file with one column: the features names
fid = fopen('features.txt');
featuresF = textscan(fid,'%s');
fclose(fid);
Nfolds=4;

% name the frequencies for which you want to make models
% (100=code for RA)
ondergrenzen=[0,1,5,10,100];
Nthresholds=length(ondergrenzen);

% read the html output files of Maxent
auc_lijst=zeros(Nthresholds *Nfeatures*Nspecies*TotaalAantalVar*Nfolds,8);
for PA=1:Nthresholds                    % loop for models with different frequencies
    pa=int2str(ondergrenzen(PA));
    for features=1:Nfeatures        % loop for models with different features
                                    % (i.e. linear, quadratic)
        ftrs=int2str(features);
            for species=1: Nspecies         % loop for different species
                for var=1:TotaalAantalVar       % loop for different variables
                for fold=1:Nfolds           % loop for different folds in cross-validation
                    fld=int2str(fold);
                    folder=[int2str(NrVar-1) 'varFW\' pa 'PA\' ftrs  'Features\'
                    maxent_omgeving_asc{1,1}{var,1}(1:4)];
                    N=1;
                    spec_name=[soortenlijst{1,1}{species,1}];
                    spec_nameF=[folder '\' soortenlijst{1,1}{species,1}  '_' fld];
                     file_name_html=[folder '\' spec_name '_' fld '.html'];
                    % read the .html file
                    fid = fopen(file_name_html);
                    if fid>0
                            tekst = textscan(fid,'%s');
                            [T1,T2]=size(tekst{1,1});
```

```
                                trshld=tekst{1,1}{335,1};
                                [s1,s2]=size(trshld);
                                aanpassing=[];
                                for a=1:(100-s2)
                                aanpassing=[' ' aanpassing];
                                end
                                treshold(teller,1:100)=[tekst{1,1}{335,1}  aanpassing];
                                spec(teller)=species;
                                t1=(T1-200);
                                verderdoen=1;
                                % find the line concerning the AUC of test and training data
                                % in the file (the line is different for every file)
                                while verderdoen
                                if sum(size(tekst{1,1}{t1,1}))==9 && …
                                      sum(size(tekst{1,1}{t1+1,1}))==4 && ...
                                      strcmp('training', tekst{1,1}{t1,1}(1:8)) && ...
                                      strcmp('AUC', tekst{1,1}{t1+1,1}(1:3))
                                           verderdoen=0;
                                else t1=t1+1;
                                end
                                end
                                auc_test_temp=tekst{1,1}{t1+17,1};
                                auc_training_temp=tekst{1,1}{t1+3,1};
                                test_temp=str2double(auc_test_temp(1:5));
                                train_temp=str2double(auc_training_temp(1:5));
                                auc_lijst(teller,1)=teller;          auc_lijst(teller,2)=species;
                                auc_lijst(teller,3)=var;                  auc_lijst(teller,4)=fold;
                                auc_lijst(teller,5)=train_temp; auc_lijst(teller,6)=test_temp;
                                auc_lijst(teller,7)=PA;              auc_lijst(teller,8)=features;
                                teller=teller+1;
                                fclose(fid);
                        else
                                auc_lijst(teller,1)=teller;          auc_lijst(teller,2)=species;
                           auc_lijst(teller,3)=var;                  auc_lijst(teller,4)=fold;
                           auc_lijst(teller,5)=-1;              auc_lijst(teller,6)=-1;
                           auc_lijst(teller,7)=PA;              auc_lijst(teller,8)=features;
                                teller =teller+1;
                        end
                        clear trshld; clear file_name_html;
                        end
                end
            end
        end
end
csvwrite(['Nthresholds_auc_lijst_6spec_' varselectie '.csv'],auc_lijst)

%% calculate the average value for the fivefold cross-validation
```

```
data=auc_lijst;
[R,K]=size(data);
rij=1;
spec=data(rij,2);
resultaat=zeros(Nspecies*TotaalAantalVar*Nthresholds*Nfeatures,8);
rijr=1;
for PA=1:Nthresholds
    pa=int2str(ondergrenzen(PA));
    for features=1:Nfeatures
        ftrs=int2str(features);
        for species=1:Nspecies
            for var=1:TotaalAantalVar
            N=1;
            temp=[];
            for rijtje=1:R
                if auc_lijst(rijtje,7)==PA && auc_lijst(rijtje,8)==features && ...
                auc_lijst(rijtje,3)==var && auc_lijst(rijtje,2)==species
                temp(N,:)=data(rijtje,5:6)
                N=N+1;
                end
            end
            reken2=[]; test2=[];
            N1=1; N2=1; N3=1; N4=1;
            for t1=1:Nfolds
                if temp(t1,1)>0
                reken2(N1)=temp(t1,1);
                N1=N1+1;
                end
                if temp(t1,1)>0
                test2(N2)=temp(t1,2);
                N2=N2+1;
                end
            end
            resultaat(rijr,1)=species;          resultaat(rijr,2)=var;
            resultaat(rijr,3)=mean(reken2);     resultaat(rijr,4)=N1-1;
            resultaat(rijr,5)=mean(test2);      resultaat(rijr,6)=N2-1;
            resultaat(rijr,7)=PA;               resultaat(rijr,8)=features;
            rijr=rijr+1 ;
            end
        end
    end
end
csvwrite(['Nthresholds_auc_CV_6spec_' varselectie '.csv'],resultaat)

%% Find the variable which entails the highest average AUC-valuefor each threshold, each
% species and each feature
N=1;
```

```
[BB1,BB2]=size(resultaat);
for PA=1:Nthresholds
    pa=int2str(ondergrenzen(PA));
    for features=1:Nfeatures
            ftrs=int2str(features);
            for species=1:Nspecies
            NN=1
            for rijtje=1:BB1
                    if resultaat(rijtje,7)==PA && resultaat(rijtje,8)==features && ...
                    resultaat(rijtje,1)==species
                    temp(NN,1:8)=resultaat(rijtje,1:8);
                    NN=NN+1;
                    end
            end
            maxi=max(temp);
            M=1;
            per=randperm(TotaalAantalVar);
            for rijp=1:TotaalAantalVar
                    rij=per(rijp)
                    if temp(rij,5)==maxi(5) && M==1
                    selectie(N,1:2)=temp(rij,1:2);
                    selectie(N,3:4)=temp(rij,7:8);
                    selectie(N,5)=temp(rij,5);
                    N=N+1;
                    M=M+1;
                    elseif isnan(maxi(5)) && M==1
                    selectie(N,1:2)=temp(rij,1:2);
                    selectie(N,3:4)=temp(rij,7:8);
                    selectie(N,5)=NaN;
                    M=M+1; N=N+1;
                    end
            end
            end
    end
end
 csvwrite(['Nthresholds_auc_besteTestWaarde_6spec_' varselectie '.csv'],selectie)

%% read the results from the previous forward selection of the variables
for reeks=1:Reeks
    rks=int2str(reeks);
    selectie=csvread(['Nthresholds_auc_besteTestWaarde_6spec_' rks 'varFW.csv'])
    [Se1,Se2]=size(selectie);
    volledigeLijst(1:Se1,reeks+3)=selectie(1:Se1,2);
    volledigeLijstAUC(1:Se1,reeks+3)=selectie(1:Se1,5);
end
volledigeLijst(1:Se1,1)=selectie(1:Se1,3);
volledigeLijst(1:Se1,2)=selectie(1:Se1,4);
```

```
volledigeLijst(1:Se1,3)=selectie(1:Se1,1);
volledigeLijstAUC(1:Se1,1)=selectie(1:Se1,3);
volledigeLijstAUC(1:Se1,2)=selectie(1:Se1,4);
volledigeLijstAUC(1:Se1,3)=selectie(1:Se1,1);
[VL1,VL2]=size(volledigeLijst);

%% make a new list of variables to use in the model, thus with all the
% previously selected variables and one extra variable
N=1;
for PA=1:Nthresholds
    pa=int2str((PA));
    for features=1:Nfeatures
            ftrs=int2str(features);
            for species=1:Nspecies
            temp=[];
            for rij=1:VL1
                    if volledigeLijst(rij,1)==PA && volledigeLijst(rij,2)==features && ...
                    volledigeLijst(rij,3)==species
                    temp=volledigeLijst(rij,1:VL2);
                    temp2=volledigeLijstAUC(rij,1:VL2);
                    end
            end
            if ~isempty(temp)
                    for vars=1:TotaalAantalVar
                    OK=true;
                    for kolom=4:VL2
                            if vars==temp(1,kolom) || isnan(temp2(1,kolom))
                            OK=false;
                            end
                    end
                    if OK==true
                            lijst(N,1:VL2)=temp;
                            lijst(N,VL2+1)=vars;
                            N=N+1;
                    end
                end
            end
        end
    end
end

%% write the batch file with Maxent command lines, to automatically run the
% different models
 all_var=maxent_omgeving_asc;
[AV1,AV2]=size(all_var{1,1})
file_name_bat=['Nthresholds_maxent_6spec_' int2str(NrVar) 'FW.bat'] ;
[Sp1,Sp2]=size(soortenlijst{1,1});
```

```
N=1;
[Lij1,Lij2]=size(lijst);
for PA=1:Nthresholds
    pa=int2str(ondergrenzen(PA));
    mapCF=['crossfilesOndergrens\' pa '\'];
    for features=1:Nfeatures
            ftrs=int2str(features);
            featureStr=[];
            if features==1
            featureStr=[];
            else
            for fe=1:features
                featureStr=[featureStr ' ' featuresF{1,1}(fe,:)  ];
            end
            end
            for species=1:Nspecies
                LLL=1;
                tlijst=[];
                for rijl=1:Lij1
                if lijst(rijl,1)==PA && lijst(rijl,2)==features && lijst(rijl,3)==species
                        tlijst(LLL,:)=lijst(rijl,:);
                        LLL=LLL+1;
                end
            end
            if ~isempty(tlijst)
                    all_var=maxent_omgeving_asc;
                    specname=soortenlijst{1,1}(species,:)
                    lijstvar=1:TotaalAantalVar;
                    NrModel=TotaalAantalVar-NrVar+1
                    temp=[];
                    temp=tlijst(:,4:Lij2);
                    for rij=1:(TotaalAantalVar-NrVar+1)
                    eruit=1:TotaalAantalVar;
                    for vars=1:(Lij2-3)
                            eruit(temp(rij,vars))=0;
                    end
                    nieuwErin=temp(rij,Lij2-3);
                    omgevingStr=[];
                    for av2=1:AV1
                        if eruit(av2)>0
                            omgevingStr=['togglelayerselected='  …
                            all_var{1,1}{av2,1}(1:4) ' ' omgevingStr];
                        end
                    end
                    map=[int2str(NrVar) 'varFW\' pa 'PA\' ftrs 'Features\' ...
                    maxent_omgeving_asc{1,1}{nieuwErin,1}(1:4)];
                    mkdir(map)
```

```
                    spec=int2str(species);
                    for sub=1:Nfolds
                            sb=int2str(sub);
                            lijn1=['java -mx512m -jar maxent.jar ...
                            environmentallayers=C:\Bmerckx\nemspec4\kaarten_OK ' ...
                            omgevingStr 'samplesfile=C:\Bmerckx\nemspec4\' mapCF ...
                            specname{1,1} '_' sb '_reken.csv ...
                            testsamplesfile=C:\Bmerckx\nemspec4\' mapCF ...
                            specname{1,1} '_' sb '_test.csv ...
                            outputdirectory=C:\Bmerckx\nemspec4\' map  ' ' ...
                            featureStr ' noplots invisible nowarnings ...
                            nooutputgrids autorun'];
                        dlmwrite(file_name_bat, lijn1, 'delimiter', ...
                            '','newline', 'pc', '-append');
                        clear lijn1;
                    end
                    clear spec;
                    lijst1(N,1:NrVar) = temp(1:NrVar);
                    N=N+1;
                    end
                end
            end
    end
end
```

PUBLICATION LIST - A1-PEER REVIEWED ARTICLES

Escaravage, V., Herman, P.M.J., Merckx, B., Wlodarska-Kowalczuk, M., Amouroux, J.M., Degraer, S., Grémare, A., Heip, C.H.R., Hummel, H., Karakassis, I., Labrune, C., Willems, W., 2009. Distribution patterns of macrofaunal species diversity in subtidal soft sediments: biodiversity-productivity relationships from the MacroBen database Marine Ecology Progress Series 382, 253-264.

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

Merckx, B., Van Meirvenne, M., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2010. Mapping nematode diversity in the Southern Bight of the North Sea. Marine Ecology Progress Series 406, 135-145.

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

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J. (subm.). Habitat suitability modelling of common species. Journal of Sea Research.

Vanaverbeke, J., Merckx, B., Degraer, S., Vincx, M., 2010. Sediment-related distribution patterns of nematodes and macrofauna: Two sides of the benthic coin? Marine Environmental Research 71, 31-40.

Vanreusel, A., Fonseca, G., Danovaro, R., da Silva, M.C., Esteves, A.M., Ferrero, T., Gad, G., Galtsova, V., Gambi, M.C., da Fonsêca-Genevois, V., Ingels, J., Ingole, B., Lampadariou, N., Merckx, B., Miljutin, D., Miljutina, M.A., Muthumbi, A., Netto, S.A., Portnova, D., Radziejewska, T., Raes, M., Tchesunov, A.V., Vanaverbeke, J., Van Gaever, S., Venekey, V., Bezerra, T.N., Flint, H.C., Copley, J., Pape, E., Zeppilli, D., Martinez Arbizu, P., Galeron, J., 2010. The contribution of deep-sea macrohabitat heterogeneity to global nematode diversity. Marine Ecology 31, 6-20.

Cover picture of nematode by Sofie Derycke
Printed by *Druk in de weer*, Ghent, Belgium
on 100% recycled paper and with vegetal ink