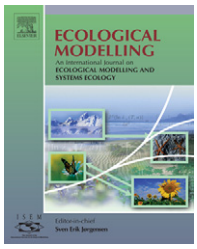
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*

Wouter Willems[a,*], *Peter Goethals*[b], *Dries Van den Eynde*[c], *Gert Van Hoey*[a], *Vera Van Lancker*[d], *Els Verfaillie*[d], *Magda Vincx*[a], *Steven Degraer*[a]

[a] *Marine Biology Section, Biology Department, Ghent University, Krijgslaan 281/S8 B-9000 Gent, Belgium*
[b] *Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Department of Applied Ecology and Environmental Biology, J. Plateaustraat 22 B-9000, Gent, Belgium*
[c] *Management Unit of the North Sea Mathematical Models, Gulledelle 100 B-1200 Brussels, Belgium*
[d] *Renard Centre of Marine Geology, Department of Geology and Soil Science, Ghent University, Krijgslaan 281/S8 B-9000 Gent, Belgium*

ARTICLE INFO

Article history:
Published on line 19 November 2007

Keywords:
Lanice conchilega
Polychaeta
Habitat preference
Generalized linear models (GLM)
Artificial neural networks (ANN)

ABSTRACT

Grab samples to monitor the distribution of marine macrobenthic species (animals >1 mm, living in the sand) are time consuming and give only point based information. If the habitat preference of a species can be modelled, the spatial distribution can be predicted on a full coverage scale from the environmental variables. The modelling techniques Generalized Linear Models (GLM) and Artificial Neural Networks (ANN) were compared in their ability to predict the occurrence of *Lanice conchilega*, a common tube-building polychaete along the North-western European coastline. Although several types of environmental variables were in the data set (granulometric, currents, nutrients) only three granulometric variables were used in the final models (median grain-size, % mud and % coarse fraction). ANN slightly outperformed GLM for a number of performance indicators (% correct predictions, specificity and sensitivity), but the GLM were more robust in the crossvalidation procedure.

1. Introduction

To scientifically underpin management decisions, there is a growing need to have detailed knowledge on the distribution of marine species. Predictive modelling is a time and cost effective method to produce detailed distribution maps. Predictive modelling objectively investigates the relation between the occurrence of a species and the abiotic habitat (Guisan and Zimmerman, 2000). This habitat is quantified by a number of environmental variables, directly measured at the time of sampling, through remote sensing or derived from other models (e.g. currents).

This research will focus on *Lanice conchilega*, a common tube-building polychaete along the North-western European coastline. This species was chosen because of its role as habitat engineer, increasing macrobenthic species diversity and abundance in rather low structured soft sediments through enhancement of the habitat complexity (Zühlke, 2001; Zühlke et al., 1998). *Lanice conchilega* is also an important food source for several demersal fish (Rijnsdorp and Vingerhoed, 2001) and, when occurring in high densities *Lanice* acts as a refugium against predation for many organisms (Woodin, 1978).

The aims of this research were: (1) to identify the environmental variables determining the distribution of *L. conchilega*, (2) to search for the most optimal model describing the habitat preferences of *L. conchilega* and (3) to compare the modelling performance of General Linear Models and Artificial Neural Networks when applied to a marine dataset.

* *Corresponding author*. Tel.: +32 92648527.
E-mail address: wouter.willems@ugent.be (W. Willems).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.10.017

2. Materials and methods

2.1. Data availability

All data used were collected in the near shore part of the Belgian continental shelf (Southern North Sea) within the framework of the HABITAT-project (Degraer et al., 2002, 2003) in October 1999, March 2000 and November 2000. The major part of the samples (265) were collected in the area of the Western Coastal Banks (WCB), a small complex of sandbanks and swales covering a wide range of soft sediment habitats (Degraer et al., 2002, 2003). Outside of the WCB, 38 additional samples were collected in November 2000, along four transects perpendicular to the coastline. The samples were collected with a Van Veen grab (sampling surface area: 0.1 m^2) and sieved over a 1 mm sieve. In each sample all adult *L. conchilega* individuals were counted. Since the goal was to predict presence or absence of the species, the densities were transformed to presence/absence.

A sediment subsample was taken with a 3.6 cm diameter core to measure nutrient concentrations ($NO_3 + NO_2$, NH_4, PO_4 and Si) in the interstitial water. Sediment granulometry was determined: the sediment fraction <1 mm was analysed with a LS Coulter laser counter (vol.%), while the sediment fraction >1 mm was weighted (mass%). The following variables were calculated: median grain-size, mean grain-size, mean/median grain-size ratio, mode, variance, skewness, kurtosis, the volume percentages of the 0–63 μm (hereafter: % mud), 63–125 μm, 125–250 μm, 250–500 μm and 500–800 μm fractions, as well as the mass percentage of the >1 mm fraction (hereafter % coarse fraction).

Bottom current speed and bottom shear stress were obtained from the 3D baroclinic hydrodynamic COHERENS model (Luyten et al., 1999). This model has a horizontal resolution of about 250 × 250 m and a vertical resolution of ten layers. U and U_{max} are the maximum and median bottom current, and BSRTM and BSRTX are the median and maximum bottom shear stress. Median and maximum chlorophyll-*a* concentration in the surface water were obtained from MERIS satellite images of 2003 from the REVAMP-project (Peters et al., 2005).

3. Modelling techniques

3.1. Variable selection

Since related variables (i.e. granulometry, nutrients) were expected to be highly correlated and thus redundant, Principal Component Analysis (PCA) was used to analyse the relationships between the variables for inclusion the models. A varimax rotation was performed to maximise the independence of the Principal Components (PCs). The non-parametric correlation coefficient Kendall's τ was used to explore the correlation between the potential environmental variables for the modelling, because it can deal better with outliers and extreme distributions of the variables (Arndt et al., 1999). Based on the PCA and the correlation analysis different sets of variables were offered to the forward selection algorithm of the GLM (see further). It was avoided to enter highly correlated variables in such a set or too much variables which were highly associated with one PC.

3.2. GLM: logistic regression

To predict the absence or presence of *L. conchilega* multiple logistic regression (Trexler and Travis, 1993), a type of GLM, was used. Logistic regression has been widely used in ecology (Paruelo and Tomasel, 1997; Ysebaert et al., 2002) and predicts the probability (between 0 and 1) that a species will occur, based on the environmental variables. Since the sample distribution was binary (absent or present), the logit link was used. The forward stepwise likelihood ratio method was used to select the best set of variables. Interaction terms and non-linear terms (i.e. quadratic) of each variable were also included in the set of variables. A cut-off value for species presence was based on the percentage of the samples in which *L. conchilega* was present in the data set (26% of the samples, cut-off of 0.26) (Ysebaert et al., 2002). The analysis was performed with SPSS version 12.0 (SPSS, Inc., Chicago, IL).

Threefold crossvalidation was used to test the robustness of the models. The data set was randomly split in three parts and two parts were iteratively used to construct a model, and the third part to test the model. If the predictive performance of the model for each fold was similar, a final model was constructed with all the data.

3.3. Artificial neural networks

Artificial Neural Networks (ANN) are a technique from the field of artificial intelligence (Lek and Guégan, 1999). They have a similar structure as the human brain: a network of connected neurons. The neurons are the building blocks of the ANN. Data enters a neuron from several other neurons, is summed and then fed into an activation function, which generates the output of the neuron. Neurons can pass on information because they are connected. The importance of a connection is expressed as an interconnection weight. The adjustment of these weights will influence the model output (Lek and Guégan, 1999). Through a learning algorithm, the weights will be adjusted iteratively, increasing the agreement between the observed and predicted presence of the species (Lek and Guégan, 1999). The Ann's in this research have their neurons organised in three layers: environmental variables are presented at the input layer, are passed on to the hidden layer which processes the information and finally the output layer generates the prediction of the probability of presence of *L. conchilega*. As with GLM, threefold crossvalidation and output visualisation were used (Fig. 1b and d) was used to test the robustness of the models. For the ANN the species was predicted to be present if the model output was larger than 0.5. The ANN's were constructed in MATLAB 6.1 using the neural networks toolbox.

3.4. Model performance and variable importance

In order to assess and compare the predictive power of GLM and ANN several performance indicators were cal-