CONSTRUCTION OF DATA-DRIVEN MODELS TO PREDICT THE OCCURRENCE OF PLANKTONIC SPECIES IN THE NORTH SEA

G. EVERAERT[1], F. DE LAENDER[1], K. DENEUDT[2], P.L.M. GOETHALS[1] & C.R. JANSSEN[1]

[1]Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, B-9000 Ghent, Belgium
[2] Flanders Marine Institute, VLIZ – InnovOcean site, Wandelaarkaai 7, B-8400 Ostende, Belgium

INTRODUCTION

Marine habitat suitability models typically predict the potential distribution of organisms based on basic abiotic variables such as salinity, oxygen concentrations, temperature fluctuations (Gogina & Zettler, 2010) or sediment class information (Degraer et al., 2008; Willems et al., 2008). Recently, Dachs & Méjanelle (2010) claimed that the modification of biota composition due to marine pollution is a factor to be taken into account in marine habitat suitability models.
Although the anthropogenic pressure on the environment has been exponentially increasing during the last six decades (Dachs & Méjanelle, 2010), the global effect of human inputs on oceanic phytoplankton remains unknown (Echeveste et al., 2010). A limited number of studies have assessed the impact of anthropogenic stressors on phytoplankton in marine environments at a global level (Faust et al., 2003; Magnusson et al.,2008).

In order to fill this knowledge gap, this research tries to determine to what extent pollution data can be used to predict the occurrence of the phytoplanktonic organisms compared to basic abiotic variables. Here we explored this issue by developing classification trees relating physical-chemical variables with the occurrence of the potential harmful toxic algae *Odontella sinensis*.

MATERIAL AND METHODS

Study area and data collection

The study area included parts of the North Sea, Atlantic Ocean and the Baltic Sea bounded by the 70th and 30th parallel north and 45th west and 35th east meridian. This area was divided into compartments of 1 by 1 degree longitude. For each of these compartments we verified if *O. sinensis* was observed in the year 1990. The compartments received the label 1 if the species was present and the label 0 if it was absent in that year. Subsequently, the dataset was extended with the corresponding physical-chemical characteristics for each of the compartments. These physical-chemical characteristics included both the basic abiotic variables salinity, water temperature and secchi depth and pollution data like sediment copper, lead, mercury and iron concentrations (Table 1). The resulting dataset consisted of 59 cases with the occurrence of the *O. sinensis* and the corresponding physical-chemical measurements for each compartment.
Three different classification trees were constructed. The first tree predicts the occurrence of *O. sinensis* based on basic abiotic variables. The second

model relates the occurrence of *O. sinensis* with pollution data, whereas the final model makes similar predictions using both data types.

Model induction and evaluation

Classification trees are hierarchical structures, where the internal nodes contain tests on the input variables. Each branch of an internal test corresponds to an outcome of the test and the prediction of the occurrence of *O. sinensis* is stored in a leaf. By implementing independent physical-chemical input variables and following the hierarchical structure of the tree, these tests lead to the associated predicted occurrence of *O. sinensis*. For each internal node that is encountered on the path, the associated test in the node is applied. Depending on the outcome of the test, the path continues along the corresponding branch, goes to the left if the answer is 'yes' or goes to the right if the answer is 'no'. The resulting prediction of the tree is taken from the leaf at the end of the path (Everaert et al., in press).
Classification trees were built through applying the R package rpart (R Development Core Team, 2009).

Models were evaluated using mathematical criteria and ecological insight. The performances of the classification tree were assessed by the determination coefficient (R^2) and the percentage of Correctly Classified Instances (CCI). The determination coefficient is a measure of the goodness of fit of the regression model. Its value is always between 0 and 1, a value close to 1 indicates a better model prediction. In order to have a satisfactory model performance, the CCI should reach at least 70% (Gabriels et al., 2007).

Table 1. Observed characteristics in the study area in 1990 based on a 1 by 1 degree longitude grid.

Variable	Abbreviation	Unit	Matrix	Minimum	Maximum	Mean	Standard deviation
A) Basic abiotic variables							
Salinity	/	psu	Pelagic	21.1	35.8	33.9	2.4
Water Temperature	Temperature	°C	Pelagic	8.9	16.8	12.0	1.7
Secchi depth	Secchi	m	Pelagic	0.5	11.0	5.8	3.2
B) Pollution data							
Copper	Cu	mg/kg	Sediment	1.6	28.9	6.3	6.2
Lead	Pb	mg/kg	Sediment	4.5	68.7	17.4	14.6
Mercury	Hg	mg/kg	Sediment	0.01	0.33	0.04	0.05
Iron	Fe	mg/kg	Sediment	3100	44800	11740	9488

RESULTS AND DISCUSSION

The mathematical criteria evaluating the classification trees are summarized in detail in Table 2.

Models based on basic abiotic variables

The first model has a R^2 of 0.43 and a CCI of 67% (Table 2). This model predicts the occurrence of the species exclusively based on the basic abiotic variables. All three predictors are used to predict the occurrence of *O. sinen-*

sis (Figure 1). At the root, the salinity of the marine environment has a major influence. In case the salinity exceeds 35.53 psu, *O. sinensis* will be predicted as absent, below this salinity the occurrence depends on the temperature and secchi depth.

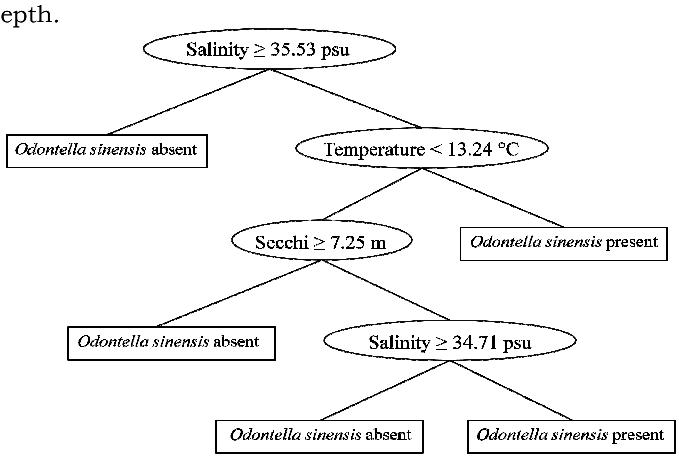
Salinity ≥ 35.53 psu
Odontella sinensis absent
Temperature < 13.24 °C
Secchi ≥ 7.25 m
Odontella sinensis present
Odontella sinensis absent
Salinity ≥ 34.71 psu
Odontella sinensis absent
Odontella sinensis present

Figure 1. Classification tree relating the occurrence of *Odontella sinensis* with a selection of basic abiotic variables.

Models based on pollution data

The second model uses pollution data to predict the occurrence of *O. sinensis*. Compared to the first model the performance of the model decreased based on R^2 but increased according to the CCI (Table 2).

The first node of the second classification model shows that the sediment copper concentration is the most important predictor (Figure 2). If the concentration of this metal exceeds 8.74 mg/kg *O. sinensis* will be predicted as absent, below this threshold it presence depends on the lead concentration. Based on this exploratory study an explanation for copper being a predictor cannot be given at this stage.

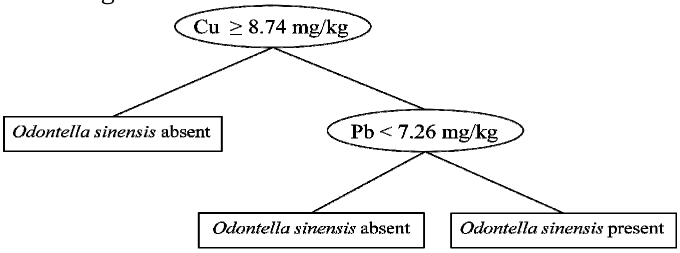
Cu ≥ 8.74 mg/kg
Odontella sinensis absent
Pb < 7.26 mg/kg
Odontella sinensis absent
Odontella sinensis present

Figure 2. Classification tree relating the occurrence of *Odontella sinensis* with a selection of pollution data.

Table 2. Performance evaluators of the classification trees

Classification tree	Predictors	R^2	CCI
Figure 1	Basic abiotic variables	0.43	67%
Figure 2	Pollution data	0.37	73%
Figure 3	Basic abiotic variables and Pollution data	0.59	73%

Models based on basic abiotic variables and pollution data

The third model makes predictions based on both basic abiotic variables and pollution data. The first knowledge rule says that if the salinity exceeds 35.53 psu, *O. sinensis* will be predicted as absent, in all other cases the occurrence depends on the copper and lead concentrations (Figure 3).

The resulting classification tree integrates the best of both previous trees (Figure 1; Figure 2) as the model consists of the most selective rule of the first model and the second model. This adjustment results in a classification tree with better performance than the previous models (R^2 = 0.59; CCI = 73%).

Maximizing R^2 and CCI does not always result in the most optimal model. In further research additional indicators like the Akaike's Information Criterion (AIC) will be used as it helps to decide if the improved fit caused by an additional predictor is worth the decreased degrees of freedom and increased complexity (Akaike, 1973).

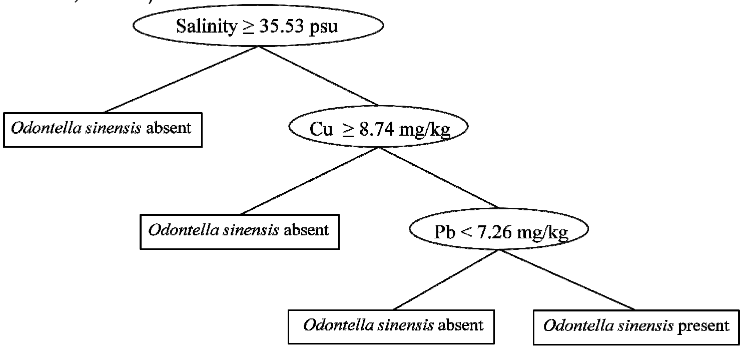
Salinity ≥ 35.53 psu
Odontella sinensis absent
Cu ≥ 8.74 mg/kg
Odontella sinensis absent
Pb < 7.26 mg/kg
Odontella sinensis absent
Odontella sinensis present

Figure 3. Classification tree relating the occurrence of *Odontella sinensis* with a selection of basic abiotic variables and pollution data.

CONCLUSION

Pollution data can be used to predict the presence of phytoplanktonic species (in this case *Odontella sinensis*). The classification trees with the highest performance were those combining basic abiotic variables and pollution data. We concluded that pollution data may have a beneficial effect on the modelling performances, compared to the use of basic abiotic variables or pollution data separately.

ACKNOWLEDGEMENTS

We would like to thank the data center of the Flanders Marine Institute (VLIZ) for their collaboration.

REFERENCES

AKAIKE H. (1973). Information theory and an extension of the maximum likelihood principle. In: PETROV B, CSAKI F (Eds.) Second International Symposium on Information Theory: 267–281. Budapest: Akademiai Kiado.

DACHS J., MEJANELLE L. (2010). Organic pollutants in coastal waters, sediments and biota: a relevant driver for ecosystems during the anthropocene? Estuaries Coasts, **33**:1-14.

DEGRAER S., VERFAILLIE E., WILLEMS W., ADRIAENS E., VAN LANCKER V., VINCX M. (2008). Habitat suitability as a mapping tool for macrobenthic communities: an example from the Belgian part of the North Sea. Cont. Shelf. Res., **28**:369-379.

ECHEVESTE P., DACHS J., BERROJALBIZ N., AGUSTI S. (2010). Decrease in the abundance and viability of oceanic phytoplankton due to trace levels of complex mixtures of organic pollutants. Chemosphere, **81**:161–168.

EVERAERT G., BOETS P., LOCK K., DZEROSKI S., GOETHALS P.L.M. (2010). Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium. Ecol. Model., doi:10.1016/j.ecolmodel.2010.08.013.

FAUST M., ALTENBURGER R., BACKHAUS T., BLANCK H., BOEDEKER W.,
GRAMATICA P., HAMER V., SCHOLZE M., VIGHI M., GRIMME L.H. (2003). Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action. Aquat. Toxicol., **63**:43–63.

GABRIELS W., GOETHALS P.L.M., DEDECKER A.P., LEK S., DE PAUW N. (2007). Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. Aquat. Ecol., **41:**427–441.

GOGINA M., & ZETTLER M.L. (2010). Diversity and distribution of benthic macrofauna in the Baltic Sea: Data inventory and its use for species distribution modelling and prediction. J. Sea Res., **64**:313-321.

MAGNUSSON M., HEIMANN K., NEGRI A.P. (2008). Comparative effects of herbicides on photosynthesis and growth of tropical estuarine microalgae. Mar. Pollut. Bull., **56**:1545–1552.

R DEVELOPMENT CORE TEAM (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

WILLEMS W., GOETHALS P., VAN DEN EYNDE D., 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.* Ecol. Model., **212**:74-79.