International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Development of data-driven models for the assessment of macroinvertebrates in rivers in Flanders

Gert Everaert[a], Ine S. Pauwels[a] and Peter L.M. Goethals[a]

Affiliation: [a]Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, B-9000 Ghent, Belgium
(Author for correspondence: Tel: +32(0)92643776; Fax; E-mail: gert.everaert@ugent.be)

Abstract: The Flemish Environment Agency (VMM) has been gathering water quality and biological data in more than 1000 sites per year since 1989. These data have been used to develop predictive models for macroinvertebrates based on data-driven methods (regression trees). These models relate the river status to the Multimetric Macroinvertebrate Index Flanders (MMIF), which is a score system developed to report in the context of the European Water Framework Directive. The trees have been developed in the R software, and several optimisations have been made by altering the dataset (variable and record selections). Models were evaluated based on mathematical criteria, ecological insight and user convenience (clarity, simplicity and coupling-potential with water quality models of the VMM). The study is a first attempt to construct a set of models that can be used by the VMM to evaluate the ecological benefits of their river management plans.

Keywords: biological water quality, ecological modelling, macroinvertebrates, regression trees

1. INTRODUCTION

The European Water Framework Directive (WFD) (EU, 2000) requires the European member states to achieve a good ecological and chemical surface water quality by the year 2015. The implementation of the WFD is based on a type-specific water quality assessment and integrates a set of biological quality elements. In Flanders, the biological quality elements 'macroinvertebrates', 'fish', 'macrophytes', 'phytobenthos' and 'phytoplankton' are used to assess the ecological water quality. The biological status of the watercourse for these elements is expressed through a scoring system between 0 and 1, known as the Ecological Quality Ratio (EQR). The EQR for each element is determined by the deviation exhibited from the expected type-specific reference condition. For the biological quality element 'macroinvertebrates', the good ecological water quality is reached when the EQR amounts to 0.7 (Gabriels *et al.*, 2009).
In order to achieve a good water quality, the European member states should take measures. The impact of these measures on the water quality in Flanders is not straightforward, because Flemish rivers are exposed to various external pressures (especially effluents from population, industry and agriculture). Managers lack predictive tools to help them decide how they can most effectively allocate the limited resources for ecological restoration. The possible ecological impact of measures can be predicted using numerical models relating the Multimetric Macroinvertebrate Index Flanders (MMIF) with physical-chemical and hydromorphological variables. Several techniques are available, but not all are equally suitable for supporting river managers in the short run. A structured method for understanding the relationships between the MMIF and the physical-chemical and hydromorphological variables are regression trees. Regression trees offer an advantage over traditional linear-regression

analysis techniques, because they introduce less prior assumptions about the relationships between the variables and have an inherent ability to discover patterns in the data that are not possible to detect using conventional models. Regression trees derive knowledge rules from the data that subsequently can be used to quantify the impact of the proposed measures (Džeroski & Drumm, 2003). Regression trees were already used by several authors in an ecological context. De'ath (2002) described the relationships between species and environmental characteristics by means of regression trees. Pesch & Schröder (2006) used this modelling techniques to relate the risk of metal bioaccumulation with site-specific and ecoregional characteristics. Kocev et al. (2009) used regression trees to model the quality of vegetation based on GIS-data. Also in other scientific branches like medical statistics, regression trees are used (Shi & Lyons-Weiler, 2007; Grubinger et al., 2010). The focus of our research is to evaluate which water quality measures will be most effective to reach the good ecological water quality. Therefore, regression trees are used to relate the MMIF to a selection of physical-chemical and hydromorphological variables.

2. MATERIALS AND METHODS

2.1 The data

The ultimate goal of this project was to evaluate to what extent the ecological water quality objectives, as stated in the WFD, will be met by environmental investment programmes and to decide what water quality measures are most effective in the Flemish context (Figure 1). Regression trees relating physical-chemical and hydromorphological variables with the biological water quality (MMIF) were developed with that intention.

Biological, physical-chemical and hydromorphological data were delivered in three databases by the Flemish Environment Agency (VMM) and the Research Institute for Nature and Forest (INBO). The VMM collected and delivered all physical-chemical data, biological data and the sinuosity of the watercourses, whereas the INBO collected and delivered information about the slope of the streams, as an indicator of the stream velocity.

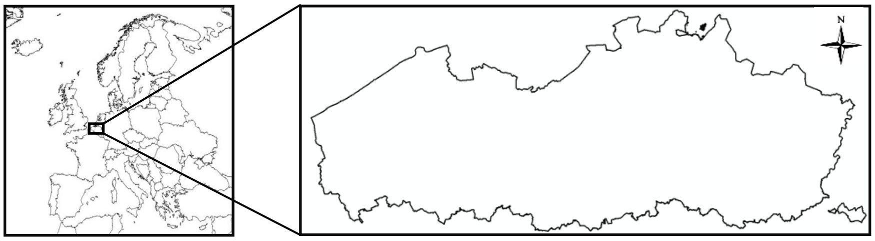

Figure 1 Location of Flanders in Europe.

2.1.1 Physical-chemical data

The physical-chemical variables shown in Table 1 cover various points and several years (from 1989 to 2009). These variables were available in the form of statistical derivatives instead of raw data. Mean, median, minimum and maximum and 5% - 10% - 90% - 95% percentiles were calculated for each variable over one year. The statistical derivative 'median' was used if all physical-chemical variables were used for model building. In case exclusively physical-chemical variables predicted by the water quality model Planification Et Gestion de l'ASsainissement des Eaux (PEGASE) were used to construct the models, the same statistical derivatives were used as Schneiders *et al.* (2009).

2.1.2 Hydromorphological data

Besides the physical-chemical data also hydromorphological data (sinuosity and the mean slope of the watercourse) were used to construct the predictive models. As part of the 'Natuurverkenning 2030' project, a map with the slope of the Flemish watercourses was made. This method assumed that the altitude of a watercourse, averaged over a certain distance, is a reasonable estimator of the slope of a watercourse and is related to the flow velocity (Dumortier *et al.*, 2009).

The sinuosity of the Flemish watercourses is calculated per segment of 100, 200 or 400m, depending on the type of surface water. The sinuosity is then calculated as the ratio between the sinuous distance from the beginning to the end of the segment and the straight distance between these points. Both slope and sinuosity of the Flemish watercourses were included in the final database.

2.1.3 Biological data

Concerning the biological assessment of the Flemish watercourses, this study focused on macroinvertebrates. The biological database encompassed 5655 MMIF-scores for different sampling locations ranging from the year 1989 to 2008. For more than 400 sampling locations the biological water quality was assessed at least once in this period. Only three sites were biologically assessed each year from 1989 to 2008.

Table 1 Observed characteristics in the Flemish watercourses, based on 1716 samples.

Variable	Abbreviation	Statistical derivative	Unit	Minimum	Maximum	Mean	Standard deviation
A) Input variables for models using all physical-chemical variables							
Biological oxygen demand	BOD_5	median	mg/L	0	235	4	11
Chemical oxygen demand	COD	median	mg/L	0	530	41	42
Chloride	Cl^-	median	mg/L	20	9003	202	613
Conductivity	-	median	µS/cm	184	23500	1121	1509
Dissolved oxygen	DO	median	mg/L	0.3	21.4	6.6	2.4
Kjeldahl nitrogen	KjN	median	mg N/L	0	29	4	4
Nitrate	NO_3^--N	median	mg N/L	0	20	4	3
Orthophosphate	oPO_4^{3-}-P	median	mg P/L	0.0	5.1	0.5	0.6
pH	-	median	-	6.2	8.8	7.6	0.3
Total phosphorus	Pt	median	mg P/L	0.0	9.6	0.8	0.9
Sinuosity	-	-	-	1000000	1979618	1055328	971772
Slope	-	mean	‰	-10.2	6.5	0.2	1.4
Suspended sediment	-	median	mg/L	0	592	25	25
B) Input variables for models using exclusively variables predicted by the PEGASE model							
Biological oxygen demand	BOD_5	maximum	mg/L	0	2270	22	105
Chemical oxygen demand	COD	maximum	mg/L	13	4920	107	240
Dissolved oxygen	DO	minimum	mg/L	0.0	21.4	3.6	2.2
Kjeldahl nitrogen	KjN	median	mg N/L	0	29	3	3
Nitrate	NO_3^--N	median	mg N/L	0	20	4	3
Orthophosphate	oPO_4^{3-}-P	mean	mg P/L	0.0	13.3	0.6	0.8
Total phosphorus	Pt	mean	mg P/L	0.0	19.4	0.9	1.1
Sinuosity	-	-	-	1000000	1979618	1055327	971772
Slope	-	mean	‰	-10.2	6.5	0.2	1.4

2.1.4 Coupling of three databases

Regression trees linking physical-chemical and hydromorphological information to the biological water quality were constructed. These three data types were separated over different databases, consequently these were combined by means of sampling location and sampling year. However, often the associated biological, physical-chemical or hydromorphological information was not available on a particular sampling location at a specific time. Only those sampling locations and years for which biological data as well as hydromorphological data and physical-chemical data were measured, were selected for the dataset to induce the regression trees. After the aggregation of the different databases a relatively small proportion of all available data was retained for the development of regression trees: only for 1716 out of 5655 MMIF-scores corresponding hydromorphological and physical-chemical data were available.

2.2 Regression tree induction

The aim of a regression tree analysis can be stated by explaining a continuous response variable Y by a vector of n predictor variables $X = X_1, X_2,...,X_n$, which can be a mix of continuous, ordinal and nominal variables (Grubinger *et al.*, 2010). Regression trees are hierarchical structures, where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test and the prediction for the value of the target attribute 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 MMIF-score. 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 (to the corresponding subtree), go to the left if the answer is 'yes', go 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, which is a constant estimate of the response variable resulting from the sample mean of the response variable in that leave (piecewise-constant model) (Everaert *et al.*, submitted).

Regression trees were built through applying the R package rpart (R Development Core Team, 2009). Rules relating the MMIF with physical-chemical and hydromorphological conditions were created using the Classification and Regression Trees (CART) algorithm (Breiman *et al.*, 1984).

2.3 Evaluating the regression trees

The performances of the regression 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, but the closer the value to 1, the better the model predicts the training data. R^2 is calculated as 1 minus the ratio between the residual sum of squares (RSS) and the total sum of squares (TSS). In order to have a satisfactory model performance, the CCI should reach at least 70% (Gabriels *et al.*, 2007). The stability of the regression trees was tested by randomizing the records in the databases and making the models again based on these reshuffled databases.

The model training and evaluation was based on the 10-fold crossvalidation procedure (Witten & Frank, 2005). In 10-fold cross-validation, the original database is randomly partitioned into 10 subsamples. Of the 10 subsamples, a single subsample is retained as the validation dataset for testing the model, and the remaining 9 subsamples are used as training datasets. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used once as the validation dataset. The results from the 10 folds are averaged to produce a single prediction of the dependent variable. Crossvalidation is particularly useful when only a limited number of data are available for training and validating the model (Gabriels *et al.*, 2007).

2.4 Selection of input variables and record selection

Only 1716 out of 5655 MMIF-scores and related records were retained after the linkage of physical-chemical and hydromorphological information to the biological data (MMIF-scores for different sampling locations over several years). In this preliminary database some variables were not present for one or more records (incomplete measurement campaign). As a solution, a second database, following from the first one, was made. The only restriction for the composition of this second datafile was the evaluation of all variables for each record (complete measurement campaign). In this second database 365 out of 1716 MMIF-scores and related records were retained (Table 2).

The water quality model PEGASE predicts the evolution of some fundamental physical-chemical variables in the Flemish watercourses. The PEGASE model simulates different scenarios quantifying the effect of several measures on the physical-chemical water quality (Peeters *et al.*, 2009). The translation of the effect of the measures on the physical-chemical water quality towards their effect on the ecological water quality is possible by implementing the knowledge rules derived with the regression trees on the water quality scenario's. However, some physical-chemical variables like chloride concentration, conductivity, pH, temperature and suspended sediments concentration are not included in the PEGASE model. In order to have a perfect coupling between the knowledge rules and the water quality scenario's, a third database was made including exclusively variables modelled by the PEGASE model (Table 2). Using this dataset, a regression tree was made exclusively including the variables predicted in the PEGASE model. Additionally, the physical-chemical variables used to construct the third regression tree were not expressed as median values over one year, but their statistical derivatives were equal to those used by Schneiders *et al.* (2009). This tree was compared to regression trees based on all available physical-chemical variables.

3. RESULTS AND DISCUSSION

Regression trees were built to understand the relationship between the biological water quality (expressed as the MMIF) and the physical-chemical and hydromorphological variables. These statistical models can predict the ecological effect of water quality measures and help decision makers to select those measures that are best implemented to reach these objectives.
Three different databases encompassing physical-chemical, hydrological and biological data were aggregated into one general database. Following from the aggregated database several options have been tested. The results of the regression trees are drawn in detail in Table 2.

Table 2 Performance evaluators of regression trees relating physical-chemical and hydromorphological data with the biological water quality evaluated through the macroinvertebrate community composition.

Regression tree	Number of records	Variables	Measurement campaign	R²	CCI (%)
Figure 2	1716	All	Incomplete	0.56	58
Figure 3	365	All	Complete	0.73	54
Figure 4	964	PEGASE	Complete	0.57	54

The first regression model has a R^2 of 0.56 and the CCI is 58% (Table 2). This model results from a database that includes all physical-chemical variables provided by the VMM (Table 1, part A). In total 1716 records were used, but the measurement campaign was incomplete, which means that for some records one or more variables were missing. The median (med) value of thirteen basic environmental variables (slope, sinuosity, BOD_5, COD, Cl^-, conductivity, DO, KjN, NO_3^--N, oPO_4^{3-}-P, pH, Pt and suspended sediment) calculated over one year at the particular sampling place was used (Table 1, part A).

At the root, the median total phosphor concentration has a major influence on the ecological water quality. In case the median phosphorus concentration exceeds 0.57 mg P/L, a watercourse can reach maximally a poor ecological water quality (Figure 2). The regression tree, shown in Figure 2, does not succeed to predict a good or high ecological water quality.

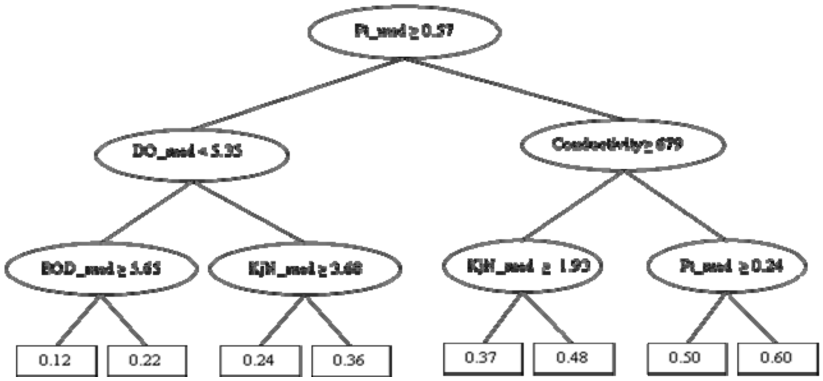

Figure 2 Regression tree relating the ecological water quality, assessed through the EQR for macroinvertebrates (MMIF), with basic physical-chemical and hydromorphological variables. The database contained 1716 records, but for some records one or more variables were missing.

The second regression model can, contradictory to the first model, predict a good ecological water quality (Figure 3). Basically, the only difference between the first two models originates from a small change in the database. Unlike the first model, the second regression tree results from a database with a complete measurement campaign, all variables were present for all records. This adjustment results in a regression tree with better performance ($R^2 = 0.73$; CCI = 54%) (Table 2).

The first node of the second regression tree shows that the median conductivity is important for the ecological water quality (Figure 3). If the median conductivity exceeds 677 µS/cm, a watercourse can maximally reach a moderate water quality. A good ecological quality can be obtained only if the sinuosity of the watercourse is sufficiently high. In the context of the WFD this tree is promising due to his ability to differentiate between the moderate and good ecological water quality.

The best performing models are based on complete datasets. In case incomplete measurement campaigns were used, the performance evaluators declined drastically. In order to have the ability to make more reliable models, the VMM should focus on the completeness of their measurements, rather than increasing the number of monitored sites. It is better to monitor all basic physical-chemical variables in each location, instead of measuring some special variables randomly without measuring other crucial physical-chemical variables.

The PEGASE model predicts a limited number of physical-chemical variables. The coupling between the different scenario's and the regression models is possible if the same variables are included in both databases. Therefore, only those variables included in the PEGASE model were selected to make the third model. The minimum (min), average (avg). median (med) or maximum (max) value of eight basic environmental variables (slope, sinuosity, BOD_5, COD, DO, KjN, NO_3^--N and oPO_4^{3-}-P) calculated over one year at the particular sampling place was used (Table 1, part B).

The resulting model cannot predict the good and high ecological water quality (Figure 4). Although the database was complete, the performances declined compared to the second model (R^2 = 0.57 and CCI = 54%) (Table 2). The exclusion of some physical-chemical variables, having a major impact on the ecological water quality, probably caused this decrease. Variables like conductivity giving an integrated insight in the overall water quality were selected in the first two regression models. However, the PEGASE model does not predict the conductivity of the watercourses. Therefore, the inclusion of the conductivity in the regression trees was not convenient from a model integration point of view. Including such variables in the PEGASE models will further optimize the models produced.

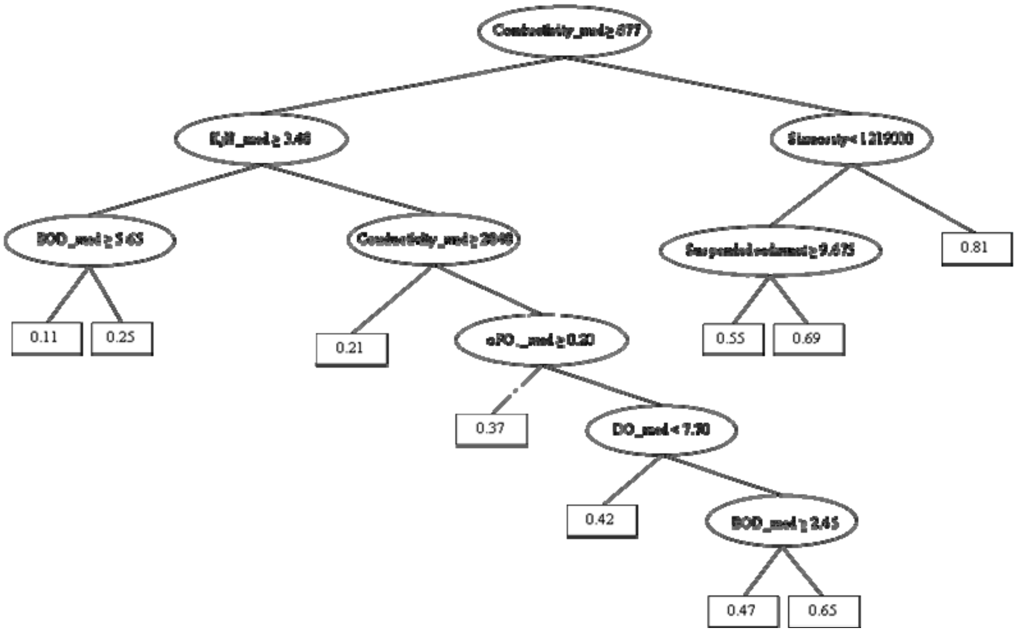

Figure 3 Regression tree relating the ecological water quality, assessed through the EQR for macroinvertebrates (MMIF), with basic physical-chemical and hydromorphological variables. The database contained 365 records, all variables were present for all records.

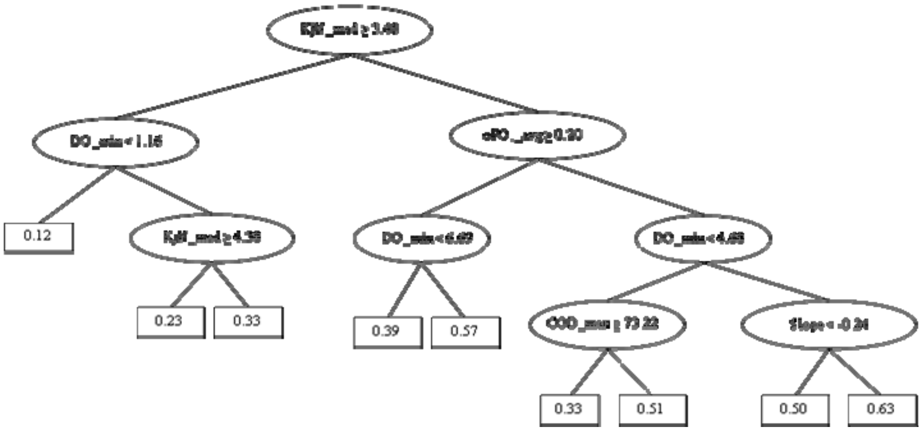

Figure 4 Regression tree relating the ecological water quality, assessed through the EQR for macroinvertebrates (MMIF), with basic physical-chemical and hydromorphological variables. The database contained 964 records, all variables were present for all records.

A sensitivity analysis can be used to see how different values of an independent variable impact a particular dependent variable under a given set of assumptions. The regression model shown in Figure 4 was selected to illustrate the effect of changing DO and KjN concentrations on the ecological water quality expressed as the MMIF. In order to see the real impact of both variables on the MMIF, it was assumed that the other independent variables included in the regression tree (COD, slope and oPO_4^{3-}-P) caused no restrictions. Therefore, in the dataset used to perform the sensitivity analysis, a constant value was given to these variables so that the highest possible MMIF could be obtained. In the dataset the KjN resp. DO concentration varied and the effect of these changes on the MMIF was evaluated (Figure 5).

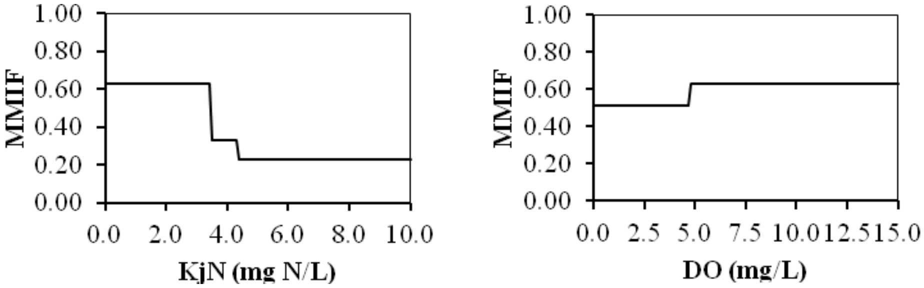

Figure 5 Sensitivity analysis illustrating the effect of changing Kjeldahl nitrogen (KjN, left) and Dissolved oxygen (DO, right) concentrations on the ecological water quality (MMIF).

The impact of a changing KjN concentration on the MMIF is shown in the left part of Figure 5. A gradual increase of the KjN concentration (from 0.0 to 10.0 mg N/L) results in a decreasing MMIF (from 0.63 to 0.23) and thus in a lower ecological water quality. One could expect the MMIF being higher at KjN concentrations ranging from 0 to 2 mg N/L. However, in the original database only a limited number of sites had a high or good ecological water quality. Therefore, the model was not sufficiently trained to predict these water quality classes. Watercourses with a KjN concentration lower than 3.48 mg N/L are predicted having a moderate ecological water quality, whereas those with a KjN concentration higher or equal to 4.38 mg N/L are predicted as bad. Watercourses with a KjN concentration ranging from 3.48 to 4.37 mg N/l are evaluated having a poor ecological water quality. One can conclude that the higher the KjN concentration, the lower the ecological water quality.
The impact of a changing DO concentration on the MMIF is shown in the right part of Figure 5. Whereas the impact of the KjN concentration on the MMIF was substantial, the impact of DO seems smaller. Increasing the DO concentration from 0.0 to 15.0 mg/L results in an increase of the MMIF from 0.51 to 0.63, the only tipping point found was a DO concentration of 4.68 mg/L. Studying Figure 4, one can see that DO is also included in other decision rules (eg. Left branch of the regression model), but due to the assumption that other independent variables (COD, slope and oPO_4^{3-}-P) had no limiting effect in determining the MMIF, the branches including these rules were neglected during the sensitivity analysis.

Two main problems were revealed during the research. In order to make more performant models in future, additional information should be gathered in sites with a high or good ecological water quality. Also the inclusion of other physical-chemical and hydromorphological variables in the dataset would be beneficial for the predictive power of the models. Therefore, the water quality model PEGASE should predict some additional variables so that these can be included in the models.

4. CONCLUSIONS

Regression trees relating the biological water quality with physical-chemical and hydrological variables can predict the ecological effect of water quality measures taken by the European member states. These data-driven models can support decision makers to select measures that are most effective and cost-efficient to reach the objectives stated by the WFD. In order to make reliable models, the VMM better changes its data collection strategy towards databases where all variables are gathered during each sampling event. Additionally, some integrative variables (like conductivity) should be included in the PEGASE model.

ACKNOWLEGDEMENTS

This research project was funded by the Flemish Environment Agency (VMM).

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, USA, 358 pp.

De'ath, G., 2002. Multivariate regression trees: a new technique for modelling species-environment relationships. *Ecology, 83*, 1105-1117.

Dumortier, M., De Bruyn, L., Hens, M., Peymen, J., Schneiders, A., Van Daele, T., Van Reeth, W., 2009. Natuurverkenning 2030. Natuurrapport Vlaanderen, NARA 2009. Mededeling van het Instituut voor Natuur- en Bosonderzoek, INBO.M.2009.7, Brussel, 224 pp.

Džeroski, S., Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. *Ecological Modelling, 170*, 219-226.

EU, 2000. EC Water Framework Directive (2000/60/EC). Official Journal of the European Communities. European Commission, Brussels, Belgium.

Everaert, G., Boets, P., Lock, K., Džeroski, S., Goethals, P.L.M., Submitted. Using classification trees to analyze the ecological impact of invasive species in polder lakes in Flanders, Belgium. *Ecological Modelling*.

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. *Aquatic Ecology, 41*, 427-441.

Gabriels, W., Lock, K., De Pauw, N., Goethals, P.L.M., in press. Multimetric Macroinvertebrate Index Flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium). *Limnologica,* doi:10.1016/j.limno.2009.10.001.

Grubinger, T., Kobel, C., Pfeiffer, K.P., 2010. Regression tree construction by bootstrap: Model search for DRG-systems applied to Austrian health-data. *BMC Medical Informatics and Decision Making*, doi:10.1186/1472-6947-10-9.

Kocev, D., Džeroski, S., White, M.D., Newell, G.R, Griffoens, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modelling, 220*, 1159-1168.

Peeters, B., D'heygere, T., Huysmans, T., Ronse, Y., Dieltjens, I., 2009. Toekomstverkenning SGBP/MIRA-S 2009: Modellering waterkwaliteitsscenario's. Wetenschappelijk rapport thema 'Kwaliteit oppervlaktewater', VMM, 83 pp.

Pesch, R., Schröder, W., 2006. Integrative exposure assessment through classification and regression trees on bioaccumulation of metals, related sampling site characteristics and ecoregions. *Ecological informatics, 1*, 55-65

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Schneiders, A., Simoens, I., Belpaire, C., 2009. Waterkwaliteitscriteria opstellen voor vissen in Vlaanderen. Wetenschappelijk rapport, NARA 2009. INBO.R.2009.22, Brussel, 94 pp.

Shi, H., Lyons-Weiler, J., 2007. Clinical decision modeling system. *BMC Medical Informatics and Decision Making, 7*, 23.

Witten, I.H., Frank, E., 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, USA, 560 pp.