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

Integrated river assessment by coupling water quality and ecological assessment models

Ine S. Pauwels[a], Gert Everaert[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: ine.pauwels@ugent.be)

Abstract: The European Water Framework Directive has a high impact on current water management in Belgium. Several restoration scenarios have been worked out by the government. To compare the effect of these scenarios, a combination of a water quality (PEGASE) and ecological assessment models (data driven models based on regression trees) has been developed and applied for the major part of the large river systems in Flanders (Belgium). The simulation results illustrate that the currently foreseen investments in water quality improvement will lead to drastic improvements of the ecological water quality, but are not enough to meet the target status in most of the streams in Flanders. Moreover, more data are needed to improve the quality of the ecological assessment models, in particular from stream systems characterised by a high ecological quality.

Keywords: ecological assessment models, water quality assessment models, restoration scenario's, regression trees, macroinvertebrates

1. INTRODUCTION

One of the most important guidelines for river managers in Europe is the European Water Framework Directive (EWFD) (EU, 2000). This guideline requires an improvement of the ecological water quality in all European member states to a good ecological quality by the year 2015. The directive covers both surface and ground water. One of the commitments in this framework is the preparation of river basin management plans. In Flanders two river basin management plans are made up, one for the River Scheldt and one for the River Meuse (CIW, 2009a, 2009b). Besides a lot of descriptive information, like a characterisation of the basin and the main anthropogenic pressure and impact on the environment, the basin management plans contain a summary of the postulated management program. This program contains all measures that are proposed to enhance the ecological quality of the Flemish surface water bodies. Some of these measures are already implemented or are planned to be in the near future (basic measures). Other measures (additional measures) are proposed for the plan-period of 2009-2015 and are necessary to achieve the goal of the EWFD (Peeters *et al.*, 2009).
The impact of the measures (proposed in the basin management plans) on the water quality is not straightforward, so it is unclear which combination of measures is most effective. Data-mining techniques can be used to induce models like regression trees (Breiman *et al.*, 1984) that relate physical-chemical river characteristics with the ecological water quality. The translation of the

effect of the measures on the physical-chemical water quality (predicted by a water quality model) towards their effect on the ecological water quality is consequently possible by implementing the induced regression trees on the simulation results of a water quality model. Using these regression trees one gains insight in the future ecological water quality based on physical-chemical data and quantifies the ecological results of proposed measures.

The water quality model Planification Et Gestion de l'ASsainissement des Eaux (PEGASE) has been trained for the prediction of the evolution of the major physical-chemical variables (mainly related to organic pollution and nutrient enrichment) in the Flemish watercourses. The PEGASE model has in this context been used to simulate different scenarios quantifying the effect of a combination of relevant measures on the physical-chemical water quality (Peeters *et al*., 2009).

In order to enable the ~~have a perfect~~ coupling between the regression tree and the water quality model~~scenario's~~, the regression trees was developed exclusively with the variables predicted by the PEGASE model. Different regression trees were developed, each for a different biological community (quality element-specific EQR score) that was modeled and the type (measured or predicted) and amount of data that were used as explanatory variables. All models were evaluated based on their reliability (expressed as *1-relative error*) and ecological relevance. Only the results of the most reliable and ecological relevant regression model are presented in this paper. This model predicts the Ecological Quality Ratio (EQR) for macroinvertebrates. The ratio is also named MMIF (Gabriels *et al*., 2006). The physical-chemical variables used to construct this regression tree were expressed as statistical derivatives equal to those used by Schneiders et al. (2009).

2. MATERIAL AND METHODS

The development of a regression tree is based on a set of independent, explanatory variables and the corresponding dependent, predicted variable (Breiman *et al*., 1984). In this case the predicted variable is the Ecological Quality Ratio (EQR) for macroinvertebrates (MMIF). This index gives an indication of the ecological water quality of a surface water body for macroinvertebrates (Gabriels *et al*., 2006). The index ranges from 0 to 1: the closer the score is to 1, the better the ecological quality of the water body. Based on this score five classes of ecological quality are defined, namely high, good, moderate, poor and bad ecological water quality. The class-boundaries (MMIF-scores) are indicated in the legend of figure 1. The variables used to predict the MMIF are those physical-chemical variables that are also modeled by the water quality model PEGASE and one hydromorphological variable, namely the slope.

2.1 Data collection

The regression tree is based on a dataset that contains biological data (MMIF) for different locations and sampling dates, and the corresponding data for the independent, abiotic variables. As independent variables the slope of the river (‰), biological oxygen demand or BOD (mg O_2/L), chemical oxygen demand or COD (mg O_2/L), dissolved oxygen or O_2 (mg O_2/L), nitrate or NO_3^- (mg N/L), Kjeldahl nitrogen or KjN (percentile over one year, expressed as mg N/L) and total phosphorus or Pt (mg P/L) were taken into account. The number of times a particular variable was measured at a specific location in the relevant year is unknown and can vary between years. Not all variables were measured at each site each year. The complete dataset contained data from 247 monitoring sites. The biological dataset encompassed 5655 MMIF

scores, also spread over different sampling locations and years (from 1989 to 2008). For minimally 400 sampling locations, a MMIF was calculated at least once during this period.
The data for the hydromorphological variable were delivered as a map containing the slope for all watercourses in Flanders. These map was made as part of the 'Natuurverkenning 2030' project (Peeters *et al.*, 2009; Schneiders *et al.*, 2009). 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).
All different sampling locations of the data described above are located in Flanders (Figure 1, location of Flanders in Europe). The sampling locations presented in Figure 1 correspond to the remaining locations (about 150) after the coupling of the datasets (see paragraph 2.2).

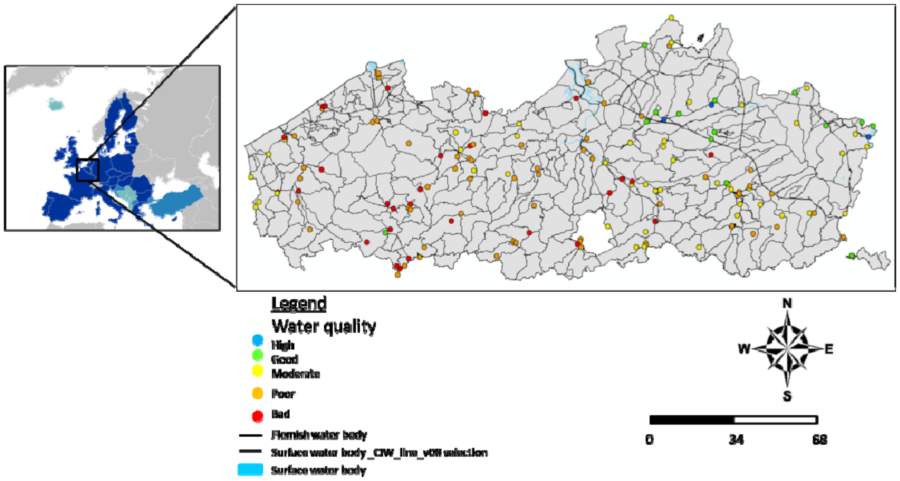


Figure 1. Location of Flanders in Europe and location of sampling points in Flanders of which data were used for the development of the regression tree. The colors of the sampling points indicate the ecological quality of the river stretch based on macroinvertebrates. (Scale map of Flanders: 1: 34000m)

2.2 Coupling of data

The data of the three different datasets where coupled according to sampling location and year of sampling. However, not all variables are measured as frequent (both in space and time) and for some variables a different monitoring network is used. Consequently, many data could not be coupled and only a small subset of all available data could be used for the development of the regression tree.

2.3 Model development

In case of the coupled dataset in this study, the chance of predicting a value for the MMIF within a certain quality-class was not equal for all classes because there was no equal distribution of the MMIF values over all classes. Most Flemish watercourses have a moderate or poor quality and only few have a good ecological quality, therefore a coupled dataset was stratified to a dataset with about 42 MMIF values of each quality class. Because there were only 14 examples of high ecological quality, the MMIF values of this class were taken together with the MMIF values indicating good ecological quality. Before stratifying the coupled dataset, records with extreme abiotic values were removed. This yielded a dataset with 171 records (MMIF value and corresponding values for the predicting, abiotic variables).

Based on this dataset a regression tree was developed that relates the abiotic conditions with the EQR score for macroinvertebrates. The aim of regression tree analysis can be stated by predicting a continuous response variable Y based on a set of predictor variables $X = X_1, X_2,...,X_n$ (Grubinger *et al*., 2010). The relations that are found between the predictor variables and the dependent variable are expressed as rules following this example:

If explanatory variable X_i (e.g. the concentration of dissolved oxygen in the water) is bigger or smaller than a specific threshold, the score of the dependent variable Y (e.g. the ecological quality ratio for macroinvertebrates) will be x or can be a range of scores bigger/smaller than x (e.g. < 0,2).

The rules are estimated according to the principle that every rule splits the dataset in two sub-datasets in which the variability of the data is reduced. The regression tree is displayed as a real tree with a root, internal nodes, branches and leafs (Figure 2). The root of the tree contains the first rule that splits the dataset in two subsets. Two branches depart from the root and depending on the outcome of this rule, the right or left branch has to be followed (left if the rule is met, right if not). The first internal nodes are rules that divide the first two subsets in new subsets, so the variability in the subsets is further reduced. The splitting stops when no rule can significantly improve the homogeneity in the subsets. Depending on the software that was used to build the tree, the output of the splitting rules (shown in the leafs) is a linear equation or a constant that corresponds to a prediction of the dependent variable. In the software used in this study, a constant was returned.

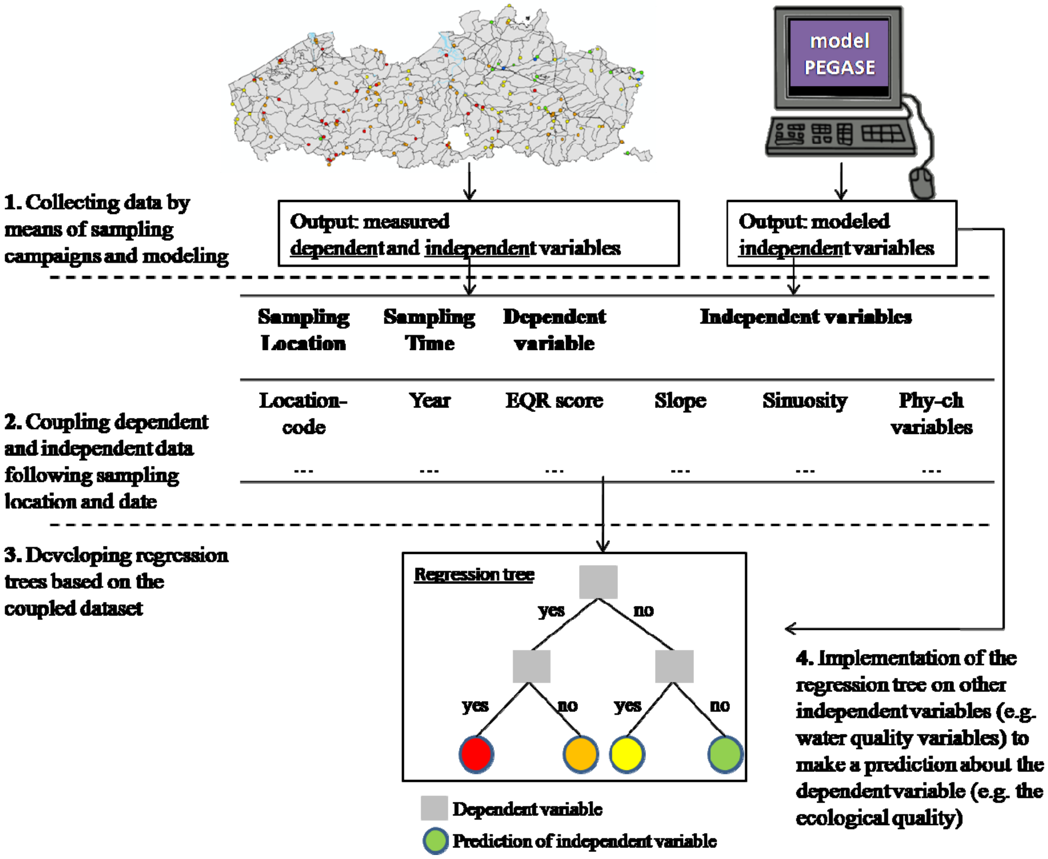

model
PEGASE
1. Collecting data by means of sampling campaigns and modeling
Output: measured dependent and independent variables
Output: modeled independent variables
Sampling Location
Sampling Time
Dependent variable
Independent variables
Location-code
Year
EQR score
Slope
Sinuosity
Phy-ch variables
---
2. Coupling dependent and independent data following sampling location and date
3. Developing regression trees based on the coupled dataset
Regression tree
yes
no
4. Implementation of the regression tree on other independent variables (e.g. water quality variables) to make a prediction about the dependent variable (e.g. the ecological quality)
Dependent variable
Prediction of independent variable

Figure 2. Schematic overview of the different steps in coupling the output of the water quality model PEGASE with an ecological assessment model (regression tree).

The performance of the regression tree was 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 regression trees in this study were developed in the statistical software program 'R', written by Ross Ihaka and Robert Gentleman (Ihaka & Gentleman, 1996). This program is more transparent than other known software to build regression trees (e.g. Weka), which allows to trace faults when inexpected results occur. It is also freely available on the internet and clear handouts make it possible to learn the program yourself. The algorithm used in this program to build regression trees is the CART algorithm (Breiman *et al*., 1984).

2.4 Implementation of the model

Ones the regression tree is developed it can be used to make predictions about the dependent variable (the MMIF in this study) based on other independent values than the values that were used to build the tree (modeled values from PEGASE for slope, $O_{2,}$ BOD, COD, KjN, NO_3^- and Pt in this study) (Figure 2).

To get an idea about the effect of the management plans on the ecological water quality of the Flemish water courses we applied the developed regression model to the data generated by simulations with the water quality model PEGASE. In PEGASE different water quality scenario's could be simulated for 2006, 2015 and 2027. The regression tree was applied on the resulting chemical data of the water quality scenario's. The incorporation of the year 2006 allows to evaluate the predictions of the model by comparing them with the measured values for the MMIF. The datasets for the scenario's contain 32751 (scenario's 2006 and 2015) and 26611 (scenario 2027) records.

3. RESULTS AND DISCUSSION

3.1 Coupling of data

Linking the three different datasets yielded a dataset in which for 964 MMIF values data were available for all abiotic variables that are also predicted by the water quality model PEGASE. The MMIF values covered all ecological classes ranging from high to bad ecological quality. The spread of the MMIF scores over the ecological quality classes is shown in Table 1.

Table 1. Spread of the EQR scores for macroinvertebrates (MMIF) over the quality classes (EQR classes), expressed as absolute number (percentage (%)).

	EQR classes				
Total dataset	high	good	moderate	poor	bad
964 MMIF records	14 (1%)	81(8%)	216 (22%)	349 (36%)	304 (32%)

The 964 MMIF records are spread over about 150 sampling locations in Flanders (Figure 1).

Because of the necessity to couple different datasets, many data became useless for the development of the models. A better coordination between the different monitoring institutions regarding their monitoring networks and the encoding of their sampling locations could result in a more comprehensive dataset and consequently more reliable models

3.2 Model development

The resulting regression tree shows that the total phosphorus and oxygen concentrations are essential to reach a good ecological quality (Figure 3). Specifically the concentration of total phosphorus should be less than 0.305 mg P/L and the concentration of dissolved oxygen should be between 4.190 and 6.865 mg O_2/L. If the concentration of total phosphorus amounts more or equals to 0.545 mg P/L, a poor ecological quality can only be reached when the biological oxygen demand is less than 12.6 mg O_2/L, the concentration of Kjeldahl nitrogen is less than 3.4 mg N/L and the concentration of total phosphorus is less than 0.6 mg P/L (left hand side of the tree). The ecological quality of the river stretch will be bad when these abiotic variables do not meet these conditions. When the concentration of total phosphorus meets the first condition (concentration of Pt less than 0.545 mg P/L), but not the second (concentration of Pt less than 0.305 mg P/L) the ecological quality can only reach a moderate quality when the concentration of oxygen is 3.46 mg O_2/L or higher. Otherwise the ecological quality will be poor.

The regression tree has a good performance. The determination coefficient (R^2) is 0.75 and for 49% of all records the predicted quality class is the same as was measured , leading to a percentage of Correctly Classified Instances (CCI) of 0.49. The regression tree is also ecologically relevant. The last rule that states that the concentration of oxygen should be less than 6.865 mg O_2/L, is probably due to indirect reasons like an excessive oxygen production by algae during the day in eutrophic water conditions.

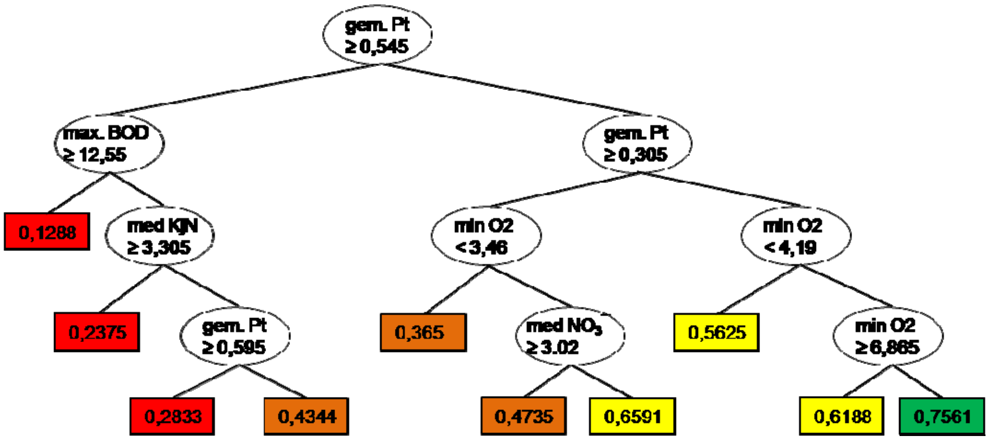
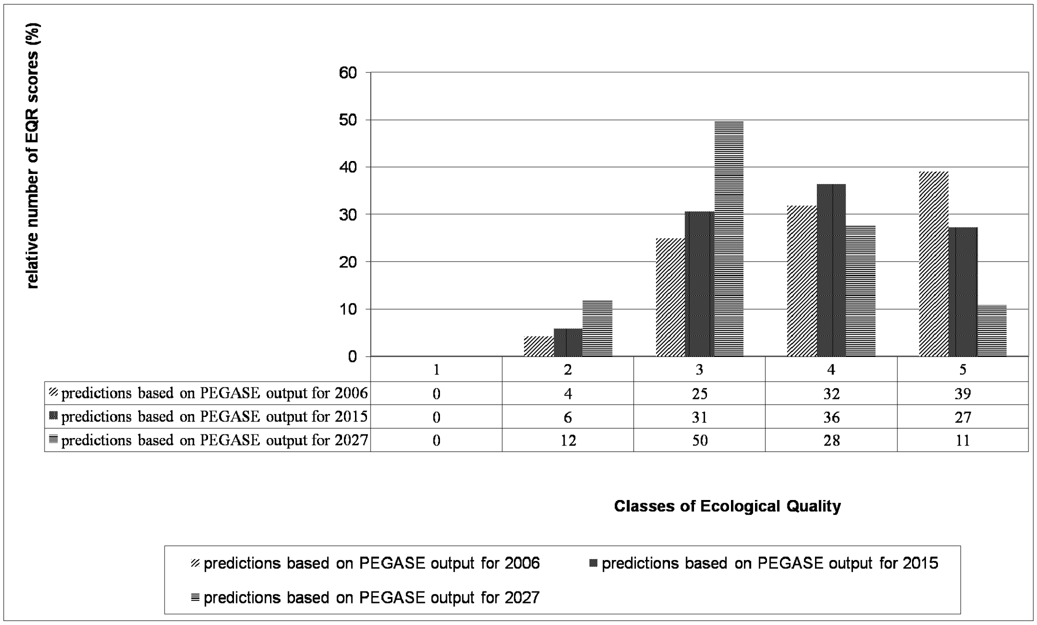


Figure 3. Result of the model development. This regression tree relates the abiotic variables to the ecological water quality expressed as the MMIF (ranging from 0 to 1). Abiotic variables: biological oxygen demand or BOD (mg O_2/L), chemical oxygen demand or COD (mg O_2/L), dissolved oxygen (mg O_2/L), nitrate or NO_3^- (mg N/L), Kjeldahl nitrogen or KjN (mg N/L) and total phosphor or Pt (mg P/L). gem.: mean concentration over one year, med.: median concentration over one year, min.: minimum concentration over one year, max.: maximum concentration over one year.

3.3 Implementation of the model

This first analysis of the potential impacts of the foreseen water quality management plans on the ecological quality in the Flemish watercourses shows an improvement from a bad and poor quality to a moderate and good ecological water quality. The best results are achieved based on the scenario 2027. The model predicts a potential improvement to the good ecological water quality for 8% of all studied watercourses, and a moderate ecological quality for 50% of all watercourses (Figure 4).

The presented model can still be optimised in some aspects. The model was only based on data from 171 records. A better coordination of the monitoring networks and encodings can yield a more comprehensive dataset and more reliable and ecological relevant models. Besides this, the model can also be improved technically.



	1	2	3	4	5
▨ predictions based on PEGASE output for 2006	0	4	25	32	39
■ predictions based on PEGASE output for 2015	0	6	31	36	27
≡ predictions based on PEGASE output for 2027	0	12	50	28	11

Figure 4. Results of the application of the regression tree on the simulations of the water quality model PEGASE for the dependent variables (slope, BOD, COD, $O_{2,}$ NO_3^-, KjN and Pt). The first bar of every set of three indicates the results for the scenario 2006, the second bar for the scenario 2015 and the third for the scenario 2027. Quality classes (high, good, moderate, poor and bad) are indicated on the x-as.

4. CONCLUSION

Integrated models like the one presented here give an added value to water policy by coupling ecological quality to a set of water quality measures based on the water quality model PEGASE. To optimise the models, more data should be collected in surface waters characterized by a good ecological quality and more variables need to be monitored (in particular hydromorphological characteristics). The actual models are interesting to get insight in the critical characteristics that determine the ecological quality of a water body, what measures could be relevant and what will be their effectiveness. Moreover the models give insights in the relevance of monitoring programmes and how to improve these (locations, characteristics, frequency, standardization, ...).

5. REFERENCES

Coördinatiecommissie Integraal Waterbeleid (CIW), Ontwerp stroomgebiedbeheerplan Schelde, *CIW*, *D/2008/6871/041,* 283p, 2009a.

Coördinatiecommissie Integraal Waterbeleid (CIW), Ontwerp stroomgebiedbeheerplan Maes, *CIW*, *D/2008/6871/042*, 213p, 2009b.

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

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

EU, EC Water Framework Directive, Official Journal of the European Communities, *European Commission, Brussels, Belgium,* 2000.

Gabriels, W., Goethals, P.L.M., De Pauw, N., Development of a multimetric assessment system based on macroinvertebrates for rivers in Flanders (Belgium) according to the European Waterframework Directive, *International association of theoretical and applied limnology*, 29, 2279-2282, 2006.

Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks, *Aquatic Ecology,* 41, 427-441, 2007.

Grubinger, T., Kobel, C., Pfeiffer, K.P., 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, 2010.

Peeters, B., D'heygere, T., Huysmans, T., Ronse, Y., Dieltjens, I. staf algemeen directeur – Team stroomgebiedbeheer, Vlaamse Milieumaatschappij, Toekomstverkenning Stroomgebiedbeheerplan/Milieuverkenning 2030: Modellering Waterkwaliteitsscenario's. *Wetenschappelijk rapport thema 'kwaliteit van het oppervlaktewater'. MIRA-T jaarrapport,* 86 p, 2009.

Ihaka, R. and Gentleman, R., R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5, 299-314, 1996.

Schneiders, A.; De Bruyn, L., Vismodellering: NARA 2009 - Wetenschappelijk rapport: Aquatisch luik – deel 4. *Rapporten van het Instituut voor Natuur- en Bosonderzoek,* 27, 74p., 2009.

6. ACKNOWLEDGMENTS

This research was funded by the Flemish Environment Agency. The first author is grant holder of the Institute for the promotion of Innovation through Science and Technology in Flanders (IWT).