# Use of Genetic Algorithms to select Input Variables in Artificial Neural Network Models for the Prediction of Benthic Macroinvertebrates

**T. D'heygere, P.L.M. Goethals and N. De Pauw**

*Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Gent, Belgium (tom.dheygere@rug.ac.be)*

**Abstract:** The first consideration in predictive ecological modelling is the selection of appropriate input variables. Numerous variables can however be involved and most of them cannot be omitted without a significant loss of information. The collection of field data on the other hand is both time-consuming and expensive. Therefore, rigorous methods are needed to detect which variables are essential and those which are not. Appropriate selection of input variables is not only important for modelling objectives as such, but also to ensure reliable decision-support in river management and policy-making. In this paper, the use of genetic algorithms is explored to automatically select the relevant input variables for artificial neural networks (ANNs), predicting the presence or absence of benthic macroinvertebrate taxa. The applied database consisted of measurements from 360 sites in unnavigable watercourses in Flanders (Belgium). The measured variables are a combination of physical-chemical, eco-toxicological and structural ones. The predictive power of the ANNs was assessed on the basis of the number of Correctly Classified Instances (CCI). The selected genetic algorithm introduced different sets of input variable to the ANN models and compared their predictive power to select the optimal combination of input variables. With this technique, the number of input variables could be reduced from 17 to 5-11. In addition, the prediction success increased with maximum 5 percent. By means of this technique, the key variables that determine the presence or absence of benthic macroinvertebrate taxa in Flanders could also be identified.

***Keywords:*** ANN; Benthic Macroinvertebrates; Genetic Algorithms; Predictive Models

## 1. INTRODUCTION

In river quality assessment, for long, analyses of aquatic biota have been used to identify the structural or functional integrity of ecosystems. Empirical evidence from studies of river ecosystems under stress suggests that the assessment of these systems can be based on a limited number of biological indicators. However, physical and chemical features of the (natural) environment affect these indicators, the structure and function of which may be changed by human activities (Norris & Thoms, 1999). Therefore could modelling allow for the integration of physical, chemical and biological characteristics into measures, rather than just observations of causes and effects. It was shown that machine learning techniques such as ANNs and genetic algorithms basically mimic aspects of biological information processing for data modelling and could be useful in ecology (Recknagel, 2001). Pre-processing of data could reduce the number of input variables and therefore help to understand the processes behind the models. Altering the input variables results in the induction of models with different prediction reliability. An exhaustive search for appropriate sets of model input variables by trying out all possible combinations is very labour intensive because the number of possible variable subsets increases exponentially with the number of variables. A manual selection of the most convenient set of model input variables based on trial and error is also labour intensive and does not guarantee a good result. An alternative could be to consult an expert and exploit his or her knowledge. The disadvantage is that knowledge may vary from expert to expert. This is mainly because the knowledge of species-habitat interrelations is still insufficient. For this reason, it was aimed for in this paper to develop an automated method for selecting appropriate variables for modelling applications by means of a genetic algorithm.

## 2. MATERIALS AND METHODS

### 2. 1 Database

The data used in this study were gathered for the development and improvement of the TRIAD methodology used for the assessment of the sediments in the navigable and unnavigable watercourses in Flanders (De Cooman *et al.*, 1999). The TRIAD assessment approach is based on a combination of biological (analysis of benthic macroinvertebrate communities), ecotoxicological (bioassays with several test organisms) and physical-chemical (measurements and analyses) data. For this modelling study, only the data of unnavigable lowland watercourses were used. This database consisted of 360 different sites in 13 river basins (Kleine Nete, Grote Nete, Meuse, Voeren, Dender, Demer, Upper-Scheldt, Leie, Ijzer, Dijle, Zenne, Ghent canal zone) that were sampled between 1996 and 1998 (Figure 1).

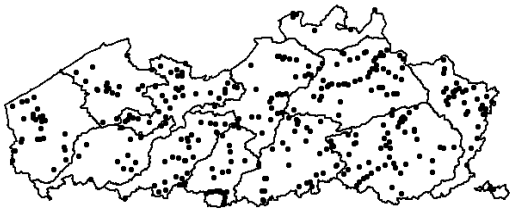

**Figure 1.** Sampling sites (360) in the unnavigable watercourses of Flanders (Belgium).

Seven physical-chemical variables were integrated into our models. Temperature, pH, dissolved oxygen concentration (% saturation) and conductivity (µS/cm) were measured in the water column. Total organic carbon (mg C/kg DS), Kjeldahl nitrogen (mg kJ-N/kg DS), total phosphorus concentrations (mg P/kg DS) and the organic matter were measured in the sediment. Two ecotoxicologial variables were determined: a 24h growth-inhibition-test with *Raphidocelis subcapitata* and a 72h growth-inhibition-test with *Tamnocephalus platyurus*, both on pore water of the sediments. Also some structural variables which were measured during sampling were available. The variables are width, depth and flow velocity of the river together with a granulometric distribution of the river sediment consisting of the percentage of clay (0-2 µm), loam (2-50 µm) and sand (50-2000 µm). Because sampling was performed during the whole year, a variable 'day' was added to account for this variation which also affects temperature, dissolved oxygen concentrations,… All variables were continuous, except flow velocity (6 classes, ranging from stagnant (class 1) to very fast (class 6)) and the ecotoxicological variables. These variables were categorised into the classes TRUE and FALSE, which represent whether or not growth inhibition is expected based on the laboratory tests.

In the sampled rivers, 92 macroinvertebrate taxa were found. By means of TWINSPAN-analysis and the calculation of a 'Species Screening Level Concentration' (SSLC), indicator scores were assigned to the taxa in order to develop a Biotic Index for sediment quality assessment of Flemish watercourses. This research resulted in the following scores for 5 selected groups of indicator taxa (rank 1 = very low tolerance,…, rank 5 = very high tolerance) (De Pauw *et al.*, 2002):

1. Trichoptera
2. Gammaridae, Bivalvia and *Sialis*
3. Hirudinea, Gastropoda (excl. *Physa*) and Asellidae
4. Chironomidae group *non thummi-plumosus*
5. Oligochaeta and Chironomidae group *non thummi-plumosus*

The results also showed that only in very rare cases common indicator taxa like Plecoptera, Ephemeroptera and Odonata were observed in or on the sediments. As a consequence, these taxa were excluded from the list. This is why this study will concentrate on a selection of the most sensitive indicator species. A total of 10 taxa were selected: Gammaridae, *Pisidium, Sialis*, *Erpobdella, Helobdella, Lymnaea,* Asellidae, *Tubificidae*, Chironomidae group *thummi-plumosus* and *non thummi-plumosus.* No Trichoptera taxa were modelled because of their very low occurrence (maximal 8 sampling sites).

### 2.2 Genetic Algorithms

Genetic algorithms are general purpose search algorithms inspired by Charles Darwin's principle of the 'survival of the fittest' to solve complex optimisation problems (Holland, 1975; Goldberg, 1989). A population of competing solutions evolve over time to converge to an optimal solution. Although it is not guaranteed to find the optimum, the use of a population helps to avoid local maxima. A solution is represented by a chromosome, consisting of several genes. A genetic algorithm starts off with an initial population of randomly generated chromosomes. During successive iterations, called generations, the initial chromosomes advance towards stronger chromosomes by reproduction among members of the previous generation. New generations are created by three genetic operators: selection, crossover and mutation. Selection of the best chromosomes makes sure that only the best chromosomes can crossover or mutate by rating the individual chromosomes by their adaptation or associated fitness.

There are numerous variations of genetic algorithms. The one presented here is the simple one outlined by Goldberg (1989). This is a powerful algorithm despite its simplicity. It was applied to find an optimal set of input variables for the prediction of the presence or absence of benthic macroinvertebrate taxa in unnavigable watercourses in Flanders. The chromosomes consisted of 17 genes, each representing an input variable, with a binary encoding. This meant that a particular variable was either selected (represented by '1') or not (represented by '0'). Each chromosome of a particular generation is allocated a piece of a roulette wheel, according to their fitness. By spinning the roulette wheel, a chromosome is selected for reproduction. In this manner, chromosomes with high fitness have a higher chance of being selected for the next generation. Crossover is performed by splitting two genes at a randomly chosen position and reconnecting the gene sequences. In this study, crossover is set at a probability of 60 percent. Mutation changes the value of a gene of an individual with a given probability to introduce a degree of random noise into the procedure. This helps to avoid local optima. In this experiment, mutation occurs with a probability of 3 percent. The initial population consisted of 20 chromosomes that were evolved through minimal 30 generations. As a criterion for accepting a solution, it was assumed that a solution was already present in the population during the last 5 generations. This procedure was repeated for 10 different initial populations by applying different 'seeds' to find the optimal input variable subsets for the macroinvertebrate taxa. The fitness of a subset of input variables was assessed by means of an ANN.

### 2.3 Artificial Neural Networks

In the experiments with genetic algorithms, a multilayered feed-forward artificial neural network (ANN) was used for evaluation. The processing elements of the models, called neurons, are arranged in a three-layer network. The first layer, called the input layer, connects with the input variables. There is one neuron for each of the input variables. As a result, the first layer consisted of maximally 17 neurons. The last layer, called the output layer, connects with the output variables. There are two neurons in the output layer, which account for the two categories into which individuals have to be classified, namely present and absent. The layers between the input layer and the output layer are the hidden layers. In our case, an ANN with one hidden layer of 10 neurons was used. Neurons are connected with the neurons in the adjacent layers. Through these connections, signals are sent from the input layer to the output layer through the hidden layer. The intensity of the transmitted signal is determined by the weight of the connections. In the training phase of the model build-up, the connection weights are adjusted to optimise the number of correctly classified instances (CCI). This CCI score, expressed as percentages of individuals correctly classified over the total number of examined individuals, was used to quantify the capability of the models to produce the right answer through the learning procedure. The back-propagation algorithm (Rumelhart et al., 1986) was applied to train the ANN with a learning rate of 0,2. After training is stopped, the performance of the network has to be tested. In our study, model training and validation was based on 10-fold cross-validation.

The outline of our model set-up is visualized in Figure 2. A genetic algorithm searches for an optimal variable subset. These subsets are evaluated by ANNs, based on CCI. Eventually, the CCI of the final subset is calculated. In this study, the predictive power of the selected subset will be compared with the predictive power of the initial variables to evaluate the results of the variable selection stage.

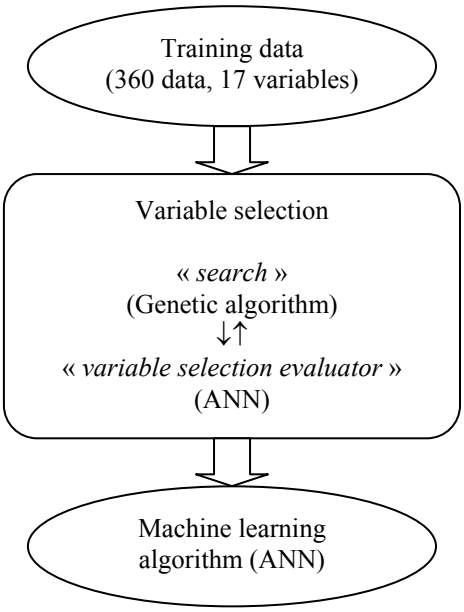


**Figure 2.** Modelling set-up of the developed input variable selection scheme.

## 3. RESULTS

The first step was to find the optimal variable subset to build ANNs by means of a genetic algorithm (see Figure 2). The second step consisted of comparing the results of the ANNs with and without a preceding variable selection procedure based on a genetic algorithm. The genetic algorithms were run with the full dataset.

Figure 3 shows the population maximum, minimum and average merit for seed 4 of *Pisidium* over 40 generations. The merit of an individual of a generation is calculated as the ratio of the number of incorrectly classified instances to the total number of instances. It is obtained by 5-fold cross-validation on the data. The average merit of a population started off at 0.25 and was quickly reduced to around 0.13 at generation 15. The lowest merit was 0.086, reached at generation 23. The maximum merit reveals the poorest solution in a generation. No clear trend could be detected in the maximum merit. This means that the probabilities for cross-over and mutation were set high enough to ensure that the probability of getting trapped in a local maximum was avoided.

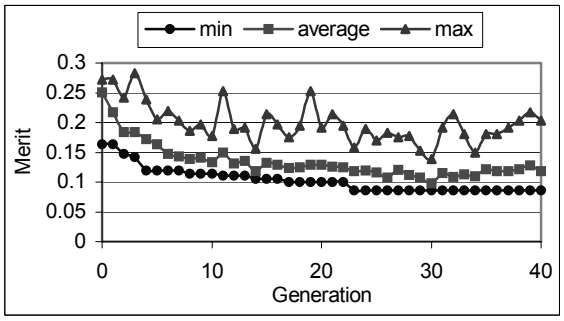


**Figure 3.** Evolution of the minimal, maximal and average merit of a variable subset for every generation for *Pisidium*.

Ten different seeds (or initial populations) were fed to the genetic algorithm to obtain a representative input variable subset. This was done because certain seeds resulted in a slight variation of few selected variables. This could be avoided by increasing the number of generations and individuals considerably or applying different seeds and average the results, which was opted for. The number of variables that were selected for *Pisidium* by introducing 10 different seeds to the genetic algorithm are represented in Figure 4.

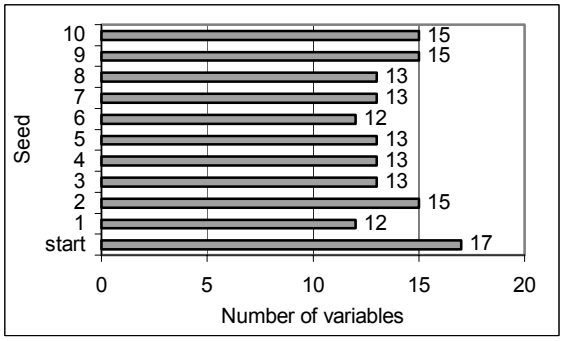


**Figure 4.** Number of selected variables by a genetic algorithm in 10 different seeds for *Pisidium*.

In a second step, the selected variables of the ten different seeds were put together to find out how to select the optimal variable subset. In this phase, the predictive power of different input sets was evaluated by an ANN with the same properties as the ANN in the genetic algorithm. In this evaluation phase, a 10-fold cross-validation was used. Five different cases were compared. The initial starting set (17 variables) was compared with the variables that were selected in respectively 50% or more, 80% or more, 90% or more and 100% of the 10 seeds. The result for *Pisidium* are shown in Figure 5 in which no 90% cases were found. Considering all 10 selected taxa, the highest CCI was found when an ANN was build with the variables which were selected in all seeds to which were added those that were just not selected (the 90% cases or the 80% cases when no 90% cases were found). This may be explained by the possibility that the evolution of a particular seed reaches a good solution but not the optimal one. Only Gammaridae had a higher CCI with variables that were selected in all seeds.

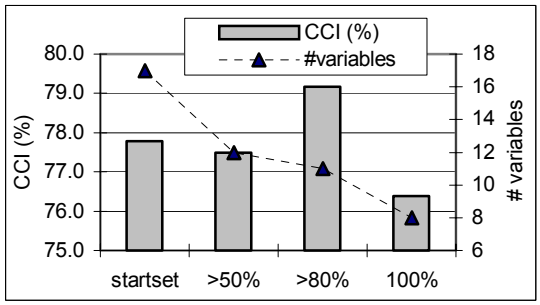


**Figure 5.** Comparison of predictive performance of variable subsets and number of selected variables for *Pisidium* by using different variable selection criteria.

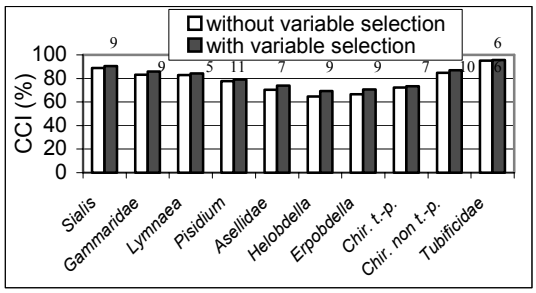


**Figure 6.** Predictive performance and number of selected input variables for 10 benthic macroinvertebrate taxa.

Comparing the CCIs before and after selection shows that the CCI increased after variable selection from 0.6% for *Tubificidae* to 4.5% for *Helobdella*. The number of variables could also be considerably reduced from 17 to 11 for *Pisidium* and the Chironomidae group *non thummi-plumosus* and to 5 for *Lymnaea* (Figure 6).

The taxa are ranked from a low abundance to a high abundance: *Sialis* was found in only 9% of the sampling sites, while *Tubificidae* were sampled in 95% of the sites. Based on this ranking, a clear relation between the CCI and the ratio present/absent is revealed. It can thus be concluded that the CCIs can be slightly increased by including a variable selection stage, but that the overall performance is highly dependent on the present/absent ratio.

## 4. DISCUSSION

Being well acquainted with the available data is very important. Data pre-processing can have a significant effect on the model performance. Looking at the data revealed that the sampling took place throughout the year, but water temperature for example is highly dependent on the season. Maier and Dandy (1996) investigated the effect of input data with and without seasonal variation on the performance of ANN models. Their results indicated that, with the aid of their hidden layer nodes, ANNs have the ability to cope with irregular seasonal variation in the data. To assist the ANN model to deal with seasonal variation, it was opted to add a variable day, ranging from 1 (January 1) to 365 (December 31).
Next to data-propressing, also the selection of model inputs is very important. Until now however, little attention has been paid to this task in relation to model performance. ANNs are clearly a data-driven approach. The structure of the model does not need to be determined first before the unknown parameters can be estimated. A data-driven approach determines itself which model inputs are critical. (Maier & Dandy, 2000).
Presenting a large number of variables to ANN models clearly increases the network size. This leads to a decreasing processing speed (high computational cost) and more data are required to estimate the connection weights efficiently. The number of nodes is fixed by the number of input variables, whereas the number of nodes in the output layer equals the number of model outputs. In our models presented here, there were two outputs values, that is present and absent. The number of input nodes was determined by the results of the genetic algorithm, while a fixed number of hidden layer nodes was applied (10). Adding a variable resulted therefore in 10 extra connection weights to be calculated. To ensure a good generalisation ability of the ANN models, a number of empirical relationships between the number of training samples and the number of connection weights have been suggested in literature. Some of these are based on the rule of thumb that the number of weights should not exceed the number of training samples. Others are based on the rule that the ratio between the number of training examples to the number of connection weights should be 2 to 1 (Maier & Dandy, 2001). Nevertheless a relatively high ratio was encountered when all variables were used. This ratio could for example be increased from 1.9 to respectively 2.8 and 5 for *Pisidium* and *Lymnaea*. An optimal network geometry can therefore be described as the smallest network that adequately captures the relationships in the training data.
The results showed a higher model performance for all tested taxa when the modelling was preceded by a variable selection stage. This is at first surprising because a loss in information in a data-driven approach should result in a loss in performance. The variables that were not selected can be seen as irrelevant for a particular taxon (Witten & Frank, 2001). The higher performance can be explained by the connection weights that also use the irrelevant information. The useless information is indeed also sent through the nodes and can as such slightly alter the connection weights. Removing variables makes ANN more transparent then when a variable selection is performed. Sensitivity analysis on the resulting model would give further information on the importance of the different variables.
The number of input variables was reduced by 6 to 12, depending on the taxon considered. Instead of genetic algorithms, other techniques could be used for variable selection in modelling, such as correspondence analysis or principal component analysis (Roadknight *et al.*, 1997). These statistical methods however have some basic limitations, e.g. a normal distribution and linearity of the data are a priori requirements. Linear relationship between the variables are indeed rare in ecology. These techniques are therefore considered ineffective in detecting the importance of variable combinations that are insignificant on their own due to non-linearity of ecological data.
Also a stepwise method based on progressively removing variables with the lowest impact until the model performance decreases significantly could be introduced. This approach was for example applied by Walley & Fontama (1998). This method also does not take interactions between variables into account, because the impact of every variable is calculated seperately. For example, the variable ‘day’ was implemented into our models to account for the seasonal variation of certain variables among which the river water temperature. When removing the variable ‘day’, the impact of the variable ‘river water temperature’ can grow significantly because it brings information about seasonal variations or it can drop because the patterns in the data cannot be recognized anymore. This will not be represented

by the calculated impact analysis before. The more variables that are introduced, the more complex the search area becomes because of local minima and maxima, due to the importance of certain variable combinations. The stepwise method does not have the possibility to find a better solution when getting trapped in a local maximum. This is when the advantage of using a genetic algorithm becomes clear, because variable combinations are compared and cross-over and mutation prevents a premature convergence to a local maximum. Furthermore, the technique remains labour intensive. A genetic algorithm on the other hand is more intensive computationally.

The selection of input variables is not only important for improving model performance, but also for policy and management objectives. Questions arise about the effect of measures that are or will be taken in the future. It is therefore essential that key variables can be found which determine the presence or absence of an indicator taxon. In this paper, it is demonstrated that the developed automated variable selection scheme can trace these key variables. The developed method could also make data collection more effective because some variables may be irrelevant in case few indicator species are to be considered. The problem is that when the river water quality improves, other variables that were ignored before could become essential in the future. For this reason, expert-knowledge may still be useful when it comes to the construction of sustainable and robust models.

## 4. CONCLUSIONS

The application of ANN models is an advantage if relations between environmental input variables are unknown, very complex or non-linear. In combination with a specific procedure for the selection of the most important impact variables by means of a genetic algorithm, the complexity of the models could be reduced. This causes an improvement of the generalisation of the induced models, which will finally result in a simplification and a better understanding of the underlying relationships in the data. By means of a genetic algorithm, in our case study, the number of input variables could be reduced from 17 to 5-11, depeding on the taxon. In addition, the prediction success increased with maximum 5 percent due to removal of irrelevant information.

## 5. ACKNOWLEDGEMENTS

The authors are indebted to the Science Foundation Flanders (FWO) for its financial support (project 3G01.02.97). The data used in our study was collected in the context of several projects related the establishment of the TRIAD methodology for the assessment of the river sediment quality in Flanders. The first author is a recipient of a grant of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

## 6. REFERENCES

De Cooman, W., M. Florus, M. Vangheluwe, C. Jansen, S. Heylen, N. De Pauw, E. Rillaerts, P. Meire, and R. Verheyen, Sediment characterisation of rivers in Flanders. The Triad approach. In: G. De Schutter (Ed.), Characterisation and treatment of sediments CATS 4, proceedings, PIH Antwerpen, Belgium, pp. 351-363, 1999.

De Pauw, N., B. Beyst, and S. Heylen, Development of a biological assessment method for river sediments in Flanders, Belgium, *Verh. Internat. Verein. Limnol.,* 27(5), 2703-2708, 2002.

Goldberg, D.E., Genetic algorithms in search, optimization and machine learning, Addison-Wesley Publishing Company, Reading, MA, 412p., 1989.

Holland, J.H., Adaptation in natural and artificial systems, University of Michigan Press, Ann Arbor, MI., 1975.

Maier, H.R., and G.C. Dandy, Neural network models for forecasting univariate time series, *Neural Network World*, 6(5), 747-771, 1996.

Maier, H.R. and G.C. Dandy, Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, *Environmental Modelling & Software*, 15, 101-124, 2000.

Norris, R.H., and M.C. Thoms, What is river health?, *Freshwater Biology*, 41, 197-209, 1999.

Recknagel, F., Applications of machine learning to ecological modeling, *Ecological Modelling*, 146, 303-310, 2001.

Roadknight, C.M., G.R. Balls, G.E. Mills, and D. Palmer-Brown, Modeling complex environmental data, *IEEE Transactions on Neural Networks*, 8(4), 852-862, 1997.

Rumelhart, D. E., and G. E. Hinton, Learning representations by back-propagation errors, *Nature*, 323, 533-536, 1986.

Walley, W.J., and V.N. Fontama, Neural network predictors of average score per taxon and number of families at unpolluted sites in Great Britain, *Water Research*, 32, 613-622, 1998.

Witten, I.H., and E. Frank, *Data mining: practical machine learning tooals and techniques with Java implementations,* Morgan Kaufmann Publishers, San Francisco, 369p., 2000.