City University of New York (CUNY)

CUNY Academic Works

International Conference on Hydroinformatics

2014

Accurate Prediction Of Ecological Quality Ratio With Product Unit Neural Networks

Arie de Niet

Erwin Meijers

Sebastiaan Schep

[How does access to this work benefit you? Let us know!](#)

ACCURATE PREDICTION OF ECOLOGICAL QUALITY RATIO WITH PRODUCT UNIT NEURAL NETWORKS

A.C. DE NIET (1), E.M.MEIJERS (2), S.SCHEP (1)
(1): Witteveen+Bos, Deventer, The Netherlands
(2): Deltares, Utrecht, The Netherlands

This paper shows how to use product unit neural networks (punn) to derive a data-driven model for the prediction of ecological quality in surface water. In a comparison with other approaches the punns provide by far the best prediction. Moreover they reveal the underlying relations between characteristics and ecological quality.

INTRODUCTION

Encouraged by the European Water Framework Directive [1] many measures are taken by Dutch government to improve water quality en ecological quality of surface water. Although it is well known which type of measures are most effective in what cases, it is uncertain what the total effect of a measure is. One way to find out this is to make use of collected data. Over the past ten years data has been collected of important characteristics and the corresponding ecological quality of Dutch water bodies. This dataset is used for the development of a data-driven model with so called product unit neural networks.

In this paper we first describe the dataset in more detail, introduce product unit neural networks and show the good results in a comparison with alternative methods.

DATASET OF ECOLOGICAL QUALITY

For Dutch surface water ecological quality is summarized in 4 ecological quality ratios (EQR's): phytoplankton, macrofauna, aquatic flora, fish. Each of the EQR's is a number between 0 and 1. The EQR's are part of the Dutch implementation of The Water Framework Directive (see [2] for more details). For natural water types the EQR can be divided in the classes bad (0-0.2), insufficient (0.2-0.4), fair (0.4-0.6), good (0.6-0.8) and very good (0.8-1.0). The dataset used for this research consists of measurements of ecological quality (EQR scores) in Dutch surface water and related characteristics like water quality, hydromorphology and some other aspects. Table 1 gives an overview of the characteristics that are measured as well as some explanation about unit, scale and properties. Some characteristics are semi-discrete, some are continuous.

Table 1. List of Characteristics

Characteristic	Short	Unit	Min	Max	Values
Banks	Ba	-	1	3	1 = green, 2 = steep, 3 = flat/muddy
Level Control	L	-	1	3	1 = unnatural, 2 = stable, 3 = natural
Maintenance	Ma	-	1	2	1 = intensive, 2 = extensive
Connectivity	Co	-	1	3	1 = isolated, …, 3 = open connection
Meandering	Me	-	1	5	1 = straight, … , 5 = free meandering
Weirs	W	-	1	3	1 = weirs, 2 = fish pass, 3 = no weirs
Shading	Sha	-	1	3	1 = no shadow, 3 = lots of shadow
Shipping	Shi	-	1	2	1 = intensive, 2 = hardly any
BOD5	Bo	mg O_2/l	0.5	20.0	Summer average (April - September)
Chloride	Cl	mg Cl/l	100	15000	Summer average (April - September)
Total Phosphorus	P	mg P/l	0.01	10.0	Summer average (April - September)
Total Nitrogen	N	mg N/l	0.67	100.0	Summer average (April - September)

It depends on the water type what characteristics are most important. For example the ecological quality of fast flow streams is determined by other characteristics than the ecological quality of deep lakes. For simplification all Dutch water bodies are merged into 8 water type clusters, each with its own characteristics. In table 2 an overview of water type clusters is presented as well as the characteristics that have a significant influence on the EQR.

Table 2. Overview of water type clusters and the characteristics that influence the EQR

Water type cluster	Ba	L	Ma	Co	Me	W	Sha	Shi	Bo	Cl	P	N
Slow flowing brook					x	x	x		x		x	x
Fast flowing brooks					x	x	x		x		x	x
Ditches	x	x	x								x	x
Canals	x	x	x					x			x	x
Shallow lakes	x	x									x	x
Deep lakes	x	x									x	x
Brackish waters	x	x	x	x						x	x	x
Brackish to saline waters	x	x	x	x						x	x	x

The dataset as used in this paper was provided by STOWA, the knowledge center of the Dutch Water Boards. For each water type cluster the dataset has 100-200 measurements.

The dataset conceals information on the effect of measures for improvement of the ecological quality. This information can be unrevealed by a data-driven model. There is a wide range of options to develop such a model with the data. In the next section we propose an alternative approach via product unit neural networks.

Since watertype cluster is a categorical variable (with 8 categories) it is not straightforward to incorporate it as a variable in any model. Usually the model quality suffers from attempts to do so. Since the aim is to create a good prediction model in all cases, each watertype cluster requires a separate set of EQR models. Since the first three watertype clusters in Table 2 do not have an EQR for fytoplankton the total number of models that needs to be derived from the data is 29.

PRODUCT UNIT NEURAL NETWORK

For the development of the data-driven model a product unit neural network (punn) was used. A punn is a special type of an artificial neural network that uses products instead of weighted sums. Usually a neural network consists of connected nodes of the form

$$\sum_{u=1}^{n_x} w_u x_u \quad (1)$$

In a standard feed forward neural network the output of a series of nodes in one layer is transferred by a transfer function and used as input for the next layer. Due to the complexity of even a simple network the model is a black box. An artificial neural network can provide a good model, but if so, it gives no insight in why it is a good model.

A product unit neural network (see [3]) uses

$$\prod_{u=1}^{n_x} x_u^{p_u} \quad (2)$$

instead of the sum in equation (1). Usually a punn is a summation of a number of these products. Punns have a number of advantages that we will shortly describe.

Great predictive power
Since products describe a much bigger space than sums, the predictive power of a punn is much bigger than of a standard neural network. Even a small network can provide a good prediction.

A product network allows for simplification
A trained punn gives insight in the importance of variables. A power close to zero gives a term in the product that is almost constant, hence one might consider to remove the particular variable from the product. Doing so in subsequent steps, meanwhile retraining the network, one can simplify the network. Simplification of a network is called pruning.

Insight in relation between inputs and outputs
A pruned product unit neural network gives insight in the relation between inputs and outputs as concealed in the dataset. Just the essential terms remain. If there is no relation the terms will drop out during the pruning process. On the other hand the terms that remain give a clue on the strength of the dependencies through the weights (powers) in the products. This makes a punn a powerful tool to uncover hidden relations in a dataset.

Transportable
Because of the predictive power and the possibility of simplification, a well trained and pruned product unit neural network is nothing but a simple formula of only a few terms that can easily be written down on a piece of paper. Or (more important) can be easily implemented in some code for use in simulations.

For training and pruning of product unit neural networks the authors at Witteveen+Bos developed an add-on for the neural network toolbox in Matlab. This toolbox was used in the experimental setup as described in the next section.

EXPERIMENTAL SETUP

The experimental setup is part of a kind of model contest for EQR. In order to find out what type of model would give the best results for EQR, three different companies and research institutes were asked to develop separate models for each water type cluster and each EQR. Afterwards the results of all models are compared. The approach with product unit neural networks is to two common approach related to the Water Framework Directive: regression trees (see e.g. [4]) and regular artificial neural networks (see e.g. [5]). In this section we describe the setup of the comparison of the models and more specific how the punns were applied to the dataset. For more details on the setup see [6].

For the development of the model the input is the measured value of the characteristics as listed in table 1. The output of the model is the corresponding value of the EQR for either phytoplankton, macrofauna, aquatic flora, or fish.

Model comparison
For a fair and clear comparison of the performance of the three different types of model the participants agreed on two rules a priori.

The first rule is a splitting of the dataset of each EQR in a part for model development and a part for model evaluation. The part for evaluation must not be involved anyhow in the development of the model.

The second rule is the way of evaluation of a model. For evaluation of the performance the following indicators were used:
- percentage of error less than 0.10;
- root mean squared error (RMSE);
- coefficient of determination (CoD).

The indicators are computed for the dataset for training and the dataset for evaluation separately.

Development of punns for EQR
Product unit neural networks require strictly positive inputs and give better results if they are of the same order of magnitude. As one can see in table 1 all the inputs are strictly positive, however they differ in scale. Hence all inputs were scaled a priori. The scaling can easily be reversed after training and pruning of a punn.

For each water type cluster and for each EQR the corresponding dataset (training part only) was used for the development of a punn via the following procedure
1. the initial punn consist of 4 products plus a constant;
2. 2000 punns are trained after random initialization;
3. the 30 best trained punns are pruned;
4. the single best pruned punn is delivered as EQR-model.

In order to prevent overtraining the training set has been further split in a part that is really used in training and a part that is used for validation, i.e. monitoring the progress of the training. If after a few consecutive training steps no progress is made on the validation set, the training is stops.

RESULTS

We will first focus on the outcome of the application of product unit neural networks. There-after we will show the performance of the punns in comparison to the other two model types.

Results of the punns

Due to limited length of this paper, we cannot give the results of all 29 models. However we will use the EQR's of Deep lakes as an example. In table 3 the performance-indicators for the training set and the evaluation set are given for the derived models for EQR for Deep lakes. It appears to be hard to build a good model for the EQR phytoplankton, but the models for aquatic flora and macrofauna are quite good.

Table 3. Performance of EQR-punns for Deep lakes

Deep Lakes	Training (incl. validation)			Evaluation		
EQR	pct <0.1	RMSE	CoD	pct <0.1	RMSE	CoD
phytoplankton	50%	0.15	0.69	46%	0.17	0.63
aquatic flora	76%	0.086	0.85	73%	0.091	0.81
macrofauna	87%	0.066	0.86	85%	0.080	0.60
fish	90%	0.069	0.86	69%	0.12	0.69

To illustrate what the punns look like we give the formula's of the first three EQR's. The EQR of phytoplankton in Deep lakes is given by

$$EQR_{phyt} = -0.7347 + 0.9968 \frac{1}{N^{0.288}} + 0.002878 N^{1,535} + 0.09921 \frac{N^{0.3107}}{P^{0.4086}} \quad (3)$$

All variables in the EQR-formulas can be found in the second column of table 1. For phytoplankton Banks and Level Control in Deep lakes appear to be of no influence. The characteristics disappear in the pruning process. Only Total Nitrogen and Phosphorus remain in formula (3). For an expert this is not really surprising, however it is nice that the punn is able to recover this from the data itself.

Instead, for aquatic flora Total Nitrogen vanishes in the pruning process. The quotient between Level Control and Banks appears to be an important term. Furthermore the EQR appears to be strongly dependent on the inverse of Level Control (L). The formula for the EQR of aquatic flora in Deep lakes:

$$EQR_{aqfl} = 7.054 - 0.3623 \frac{L^{3.13}}{Ba^{2.456}} + 0.2406 \frac{L^{3.554}}{Ba^{3.148}} + \frac{0.5741}{L^{7.167}} - 7.474 \frac{P^{0.01563}}{L^{0.1776}} \quad (4)$$

Finally we give the EQR formula for macrofauna:

$$EQR_{macf} = 0.7516 - 0.1273 \frac{L^{0.6037} P^{0.3094} N^{0.8157}}{Ba^{0.1388}} + 0.08084 \frac{Ba^{1.107} L^{1.124}}{P^{0.03239} N^{0.2039}}$$

$$-0.3805 \frac{Ba^{0.3697} L^{0.6279} P^{0.05867}}{N^{0.1495}} + 0.03044\, Ba^{0.3636} L^{1.545} P^{0.4326} N^{1.041} \quad (5)$$

This one is much more complex and less easy to understand. It appears that pruning is not possible in this case. However it still can be written down in a single formula and incorporated in this paper. The EQR of fish in Deep lakes is given by a long and complex formula as well. Fortunately for other watertype clusters in many cases the model can be simplified by pruning.

Results for model comparison

For each model-type the average performance on the evaluation set of all 29 models is given in Table 4. From the table it is clear that the overall performance of punns is considerably better than the performance of the alternatives.

Table 4. Performance on evaluation set averaged over all 29 models (from [6])

model type	pct <0.1	RMSE	CoD
regression tree	63%	0.121	0.49
standard neural network	64%	0.129	0.41
product unit neural network	68%	0.106	0.60

A more detailed analysis reveals that the performance differs per water type cluster. There are several cases where the regression tree model or the standard neural network model performs better than the punn. However in 21 out of 29 models the punn has simply the best performance. Moreover, whereas the alternative approaches sometimes give a model with a coefficient of determination less than zero, i.e. the population mean is a better prediction than the model itself, the punns in all cases give a model of at least reasonable quality.

CONCLUSION

In this paper we showed that a product unit neural network is a powerful tool to develop a data-driven model for accurate prediction of the ecological quality ratio. In comparison with other methods the punns give a better performance. Moreover they have additional advantages as interpretability and transportability. With product unit neural networks the underlying relation between characteristics and ecological quality is unrevealed. The fact that the punn results in a simple formula allow the models to be more easily applied in a new context.

REFERENCES

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

[2] See http://krw.stowa.nl/Projecten/KRW_maatlatten.aspx?pId=49 (available in Dutch only)

[3] Durbin R., Rumelhart D.E., "Product units: A computationally powerful and biologically plausible extension to backpropagation networks", Neural Computation, Vol.1, No. 1 (1989), pp 133-142.

[4] Everaert G., Pauwels I., Goethals P. "Development of data-driven models for the assessment of macroinvertebrates in rivers in Flanders", Modelling for environment's sake : proceedings of the fifth biennial conference of the International Environmental Modelling and Software Society. 3.(2010), p.1948-1956.

[5] Gebler D.,Kayzer D., Szoszkiewicz K., "Artificial neural network modelling of macrophyte indices based on physico-chemical characteristics of water", Hydrobiologia (2013), doi 10.1007/s10750-013-1585-7, pp 1-10.

[6] Visser H., "De ecologische kwaliteit van Nederlands oppervlaktewater: een vergelijking van drie modelleringstechnieken" (available in Dutch only), PBL (2013), pp 1-21. publicwiki.deltares.nl/download/attachments/32079883/PBL+vergelijking+drie+modellen+2013.pdf