Predictive Modelling of Plankton Dynamics in Freshwater Lakes using Genetic Programming

P.A. Whigham
University of Otago, Department of Information Science, Dunedin, New Zealand
Friedrich Recknagel
University of Adelaide, Department of Soil and Water, Waite Campus,
Glen Osmond, South Australia 5064

Abstract Building predictive time series models for freshwater systems is important both for understanding the dynamics of these natural systems and in the development of decision support and management software. This work describes the application of a machine learning technique, namely genetic programming (GP), to the prediction of chlorophyll-a. The system endeavoured to evolve several mathematical time series equations, based on limnological and climate variables, which could predict the dynamics of chlorophyll-a on unseen data. The predictive accuracy of the genetic programming approach was compared with an artificial neural network and a deterministic algal growth model. The GP system evolved some solutions which were improvements over the neural network and showed that the transparent nature of the solutions may allow inferences about underlying processes to be made. This work demonstrates that non-linear processes in natural systems may be successfully modelled through the use of machine learning techniques. Further, it shows that genetic programming may be used as a tool for exploring the driving processes underlying freshwater system dynamics.

1. INTRODUCTION

This paper describes the application of a genetic programming (GP) system to predict the timing and magnitudes of algal blooms for Lake Kasumigaura, in the South-Eastern part of Japan. This data has previously been studied using an artificial neural network (Recknagel, 1997; Recknagel et al., 1998; Recknagel and Wilson, 1999) which demonstrated the potential for these tools to predict highly nonlinear phenomena such as blue-green algal blooms in freshwater lakes. The purpose of this paper is to compare the predictions for this system developed by genetic programming with the previous neural network approach, and to demonstrate that the GP system allows the underlying processes for this system to be studied.

1.1 Data Characteristics of Lake Kasumigaura

Lake Kasumigaura is situated in the South-Eastern part of Japan. It is a large, shallow water body where no thermal stratification occurs. Water temperatures vary widely, from 4°C in the winter to 30°C in summer. The lake has high external and internal nutrient loadings and therefore primary productivity is high. As algal succession changes species abundance year by year, it is very difficult to predict algal blooms or develop causal models of the lake algal behaviour. This has become an important issue due to the need for good predictions of the growth of harmful blue-green algae such as *Microcystis spp*, *Oscillatoria* and *Anabaena flos aquae*.

1.2 The Ecology of Freshwater Phytoplankton

Phytoplankton include representatives of several groups of algae and bacteria. They are usually distinguished by being freely floating and dependent on water movement for maintenance and transport (Reynolds, 1984). Many factors affect their population dynamics and they vary depending on the type of phytoplankton under consideration. However, all algae species rely on light as a basic input for photosynthesis and require nutrients such as nitrogen and phosphorus for growth and reproduction. Factors such as water temperature, turbidity, mixing, competition and grazing are also relevant to the population dynamics of algae. Therefore seasonal patterns are normally evident in the cycles of population density, however these signals are often dramatically shifted due to nutrient loadings, other species dominance and other less easily identified factors. Even though much work has been done on phytoplankton, there are still difficulties with developing reliable predictive models for algal growth. This is mainly due to the highly nonlinear behaviour of the population as a whole, which depends on both variable climatic conditions as well as the relationship between components of aquatic food chains.

2. GENETIC PROGRAMMING

The field of Genetic Programming (Koza, 1992) developed from the evolutionary population-based search methods used with Genetic Algorithms (GA) (Holland, 1992). GP extended the fixed-length approach of GA's to allow basic computer programs to be evolved in the form of functional LISP expressions. This extended the GA concepts by allowing the size and shape of the evolved solutions to change, thereby offering the possibility for the system to discover a program which generalised a set of training examples. GP has been applied successfully to many problems (Koza, 1990; Roston and Sturges, 1995; Gruau, 1996; McKay, 1997) and has been previously shown to be useful in developing time series expressions (Whigham, 1999). The GP system used with this study was designed specifically for time series analysis by allowing the user to select sub-portions of the training data during the evolution of the system and to incrementally adjust the parameters that control the search algorithm. The system used a steady-state population which incrementally added new population members while probabilistically removing the weakest.

2.1 GP Parameters

GP Starting Parameter	Value
Population Size	500
Initial Maximum Depth of Program	5
Crossover Rate	90%
Mutation Rate	5%
Available Functions	+,-,*,/ inv(x), ln(x), x^y, $\log_x y$, sinh(x), cosh(x), $\Re$

Table 1. The Initial GP Parameter Specification.

The GP system was initially defined by the set of parameters shown in Table 1. The values of Table 1 are used for each of the experiments described in this paper. The mathematical functions available as part of the evolved expression were the standard arithmetic operators, plus the inverse function, natural logarithm, power function, logarithm to an arbitrary base and the hyperbolic functions. The random real-number function $\Re$ was used to add random numbers to the equations. These numbers could then be adjusted during the evolution by a hill climbing algorithm. The problem specific arguments used in the experiments are described in Section 4. The Initial Maximum Depth parameter specifies the maximum tree depth that is allowed when randomly constructing the initial programs.

2.2 Crossover

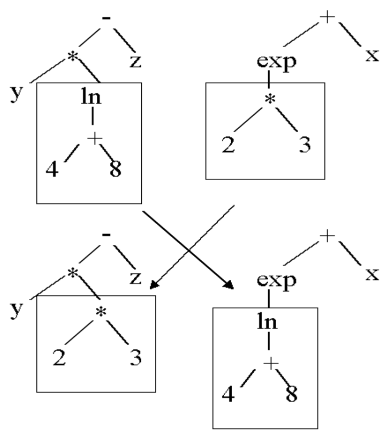

Figure 1. Crossover swaps subtrees between two programs.

Crossover with GP is performed by randomly swapping components from two programs, as shown in Figure 1. Here, the programs exp(2*3)+x and y*ln(4 + 8)-z are crossed to give the new programs y*(2*3)-z and exp(ln(4+8))+x. Crossover is designed to allow useful components of a fit partial solution to propagate throughout the population.

2.3 Mutation

Mutation with GP is performed by randomly deleting a subtree within a selected program, and generating a new, random, subtree based on the set of functions and arguments that have been defined for the problem. The new subtree is limited by the current maximum tree depth.

2.4 Hill Climbing Mutation for $\Re$

Random real numbers, represented as the variable $\Re$, are used as constants to allow the evolving mathematical expresions to adjust their scale and magnitude. These numbers are generated at random at the commencement of the evolution and are not tuned in any way with the final solution. To allow a fine tuning of an evolved expression, a hill climbing mutation for the random numbers contained in an expression is used. This operation can be applied to the current fittest solution during the evolution for a solution at any time by user control. It is typically used to tune the constant values when the evolution is complete.

3 PLANKTON DYNAMICS USING ARTIFICIAL NEURAL NETWORKS

The ANN setup and procedures have been previously described (Recknagel et al. 1998). A feed-forward architecture with back propagation for training was used for the plankton dynamics. The hyperbolic function was chosen as the transfer function to calculate the activation levels. The number of hidden layers, nodes and neurons as well as the learning rates and momenta were used as control parameters to find optimum training results.

4 TRAINING AND TEST DATA SETUP

Measured Factor	Av ± Std. Dev.	Units
Ortho phosphate	14.14 ± 25.71	mg/l
Nitrate	520.56 ± 503.4	μg/l
Secchi Depth	85.43 ± 44.57	cm
Dissolved Oxygen	11.2 ± 2.14	mg/l
pH	8.74 ± 0.59	-
Solar Radiation	1281 ± 671	MJ/m^2
Water Temperature	16.36 ± 7.79	°C
Chlorophyll-a	74.43 ± 42.51	μg/l

Table 2. Factors measured with the daily time series data.

Table 2 shows the measured variables used for developing the models. For all experiments, 8 years of daily data (’84, ’85, ’87, ’88, ’89, ’90, ’91, ’92) were used for training and 2 years of daily data (’86 and ’93) for testing the GP system and the neural network. The root mean square error (RMSE) was used as the fitness function for the training data and as a measure of accuracy for the test data. A lower RMSE was taken to indicate a better prediction of the test data. When comparing 2 different learning techniques a lower RMSE for the unseen (test) data implied that the learning system had better generalised the patterns found in the training data.

4.1 Data Preparation

The NN approach normalised each input and output variable to the range {0…1}, where 0 → minimum value and 1 → maximum value. Since this normalisation incorporates the additional information of the range for each variable this encoding would be expected to allow a better model to be developed. In the case of the GP system, the exponential, power and hyperbolic functions will overflow with typical values for many of the variables if normalisation does not occur. Issues of scaling are considered in Section 5.

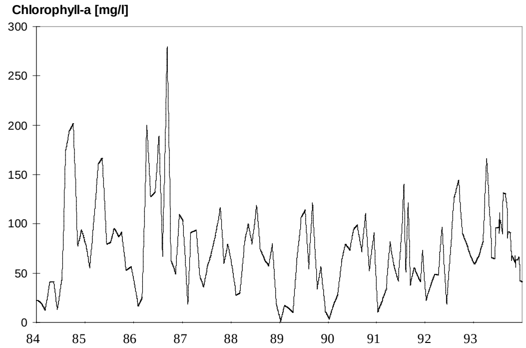

Figure 2. Daily Time Series data for Chlorophyll-a for all years.

5. PREDICTING CHLOROPHYLL-A

Chlorophyll-a is used as a sampling technique for estimating the total biomass of the phytoplankton community in a waterbody. Hence the driving factors for chlorophyll-a tend to represent the overall behaviour of the plankton community. The daily time series data for chloroplyll-a is shown in Figure 2. Note that the validation (test) year 1986 has a far larger concentration measure than any of the training years. The other validation year (1993) is more typical of the training years.

5.1 Results using Non-Normalised Data

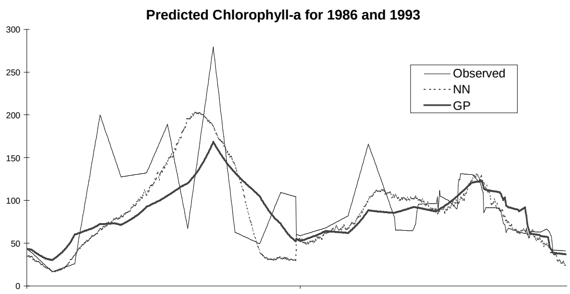

Figure 3. Prediction of Chlorophyll-a Part I (RMSE = 41.83)

Equation (1) shows the evolved expression for chl-a prediction, and Figure 3 gives the response for the two test years.

$$(4600.42/S) + T \qquad (1)$$

where

S = secci-depth and T = water temperature.

The RMSE for (1) based on the 2 years of test data was 41.83 versus the NN error of 41.78. For our purposes these results are comparable.

A different run using GP produced Equation (2), which had a lower RMSE (40.67) for the test data. However the equation was more complex.

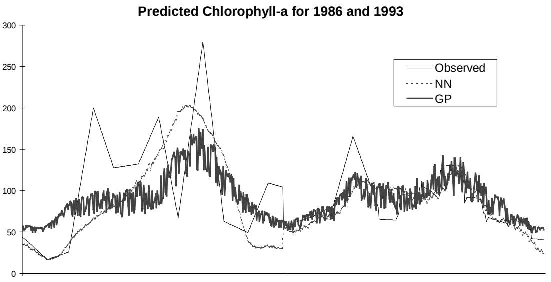

Figure 4. Prediction of Chlorophyll-a Part II (RMSE = 40.67)

$$E_1 + \frac{25.786 + P + E_2}{(0.098N / T + 0.107T)} \qquad (2)$$

where

$$E_1 = 33.174 + \frac{L + 893.75 + 9.35T}{S} + \frac{69.51}{S + 42.28}$$

$$E_2 = \frac{151.94N / T + 95.63T + 893.75N + 574.68 + 893.75P}{(9.356N + 95.527T)}$$

and L = solar radiation, N=nitrate, T = water-temperature and P = phosphorus.

The results against the test data are shown in Figure 4. Note that Equations (1) or (2) do not use exponential or power functions and rely on constants to scale the data.

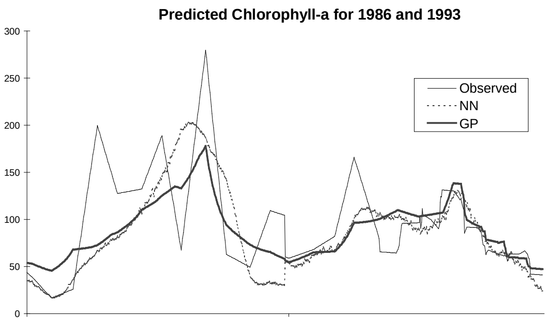

Figure 5. Prediction of Chlorophyll-a Using Normalised Data (RMSE = 37.08).

5.2 Results using Normalised Data

To demonstrate that different solution forms were created based on the scaling of the data, a series of runs were performed using normalised data. When attempting to produced a small, generalised model, equation (3) was discovered, which had a low RMSE (37.08). Note however it did not predict the peak measure of Chl-a. The resulting prediction is shown in Figure 5.

$$\frac{\cosh T}{(6S + 2N + 2.250836)} \qquad (3)$$

where

S = secci-depth, N = nitrate, T = water temperature.

The simplicity of this equation resulted from forcing the GP system to only allow equations to be produced to a depth of 6. Additionally, at certain times the maximum allowed depth was decreased to break apart the basic partial solutions and then the system was allowed to reconstruct them. This forced building blocks to be gradually created resulting in (3). Note that the lower RMSE for (3) shows that this strategy has helped to produced a more general, yet more accurate, solution.

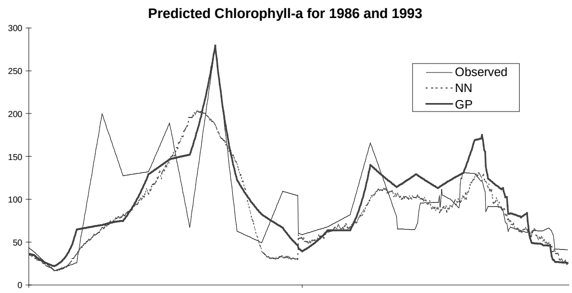

Figure 6. Prediction of Peak Chlorophyll-a using Normalised Data (RMSE = 37.08)

5.4 Modelling the Peak Chlorophyll Response

The GP system allowed the user to alter the range of training data during the evolution. By selecting a subset of the training data from 1984 to build the initial equation elements the final equation (based on all of the training data) modelled the peak chlorophyll response, as shown in Figure 6. The RMSE was 37.08 which compared favourably with the other evolved solutions. The peak prediction is modelled by Equation (4).

$$e^{-\left[2.795 * S + N + S / e^{N}\right]} \qquad (4)$$

6. PROCESS-BASED MODELLING

A difference equation model for algal growth (Equation (5)) was used to compare the previous data-driven techniques that have been described. This equation was developed based on current process understanding (Recknagel and Benndorf, 1982). The constants in (5) were set based on values that had been discovered from laboratory tests and fieldwork. Using this original equation the prediction for the test period is shown in Figure 7. The RMSE was 91.46 which is significantly worse than either GP or NN.

$$Chla_{t+1} = Chla_t + Chla_t * (Phot - \mathrm{Re}\,sp) - Chla_t * (Cop + Clad) * 0.0001 \tag{5}$$

where

$$Phot = (0.068 * T) * (\frac{0.025 * L}{28 + 0.025 * L}) * (\frac{P}{Chla_t} / (\frac{1.7}{X} + \frac{P}{X} + \frac{1.7}{Chla_t} + \frac{P}{Chla_t}))$$

$$X = 5.76 * Chla_t \text{^} 0.41$$

and

$$\mathrm{Re}\,sp = (0.00228 * T) + 0.3 * Phot$$

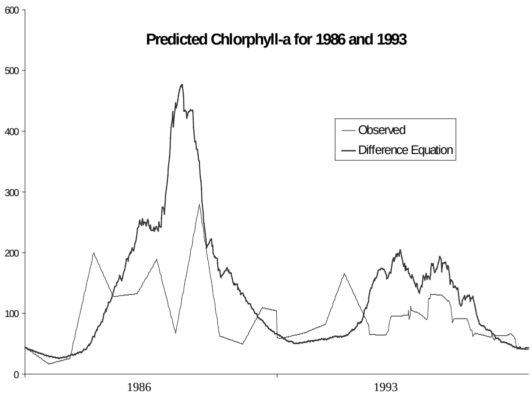

Figure 7. Prediction of Chlorophyll-a using a difference equation (RMSE = 91.46).

In an attempt to improve the performance of (5) the constants were tuned based on the training data. Each constant could vary within a range of ±20% based on the 8 years of training data. The constants were varied using a hillclimbing mutation. The RMSE was used as the fitness measure, however only values of chl-a that were above 75 mg/l contributed to the error measure in an attempt to force the equation to better model the peak events. The prediction for the test period is shown in Figure 8 and the modified equation is shown as Equation (6). This equation had a dramatically improved RMSE of 46.75 which approached the GP and NN model accuracies.

$$Chla_{t+1} = Chla_t + Chla_t * (Phot - \mathrm{Re}\,sp) - Chla_t * (Cop + Clad) * 0.00008 \tag{6}$$

where

$$Phot = (0.07634 * T) * (\frac{0.0295 * L}{22.4 + 0.0299 * L}) * (\frac{P}{Chla_t} / (\frac{1.36}{X} + \frac{P}{X} + \frac{1.36}{Chla_t} + \frac{P}{Chla_t}))$$

$$X = 4.608 * Chla_t \text{^} 0.328$$

and

$$\mathrm{Re}\,sp = (0.00273 * T) + 0.2696 * Phot$$

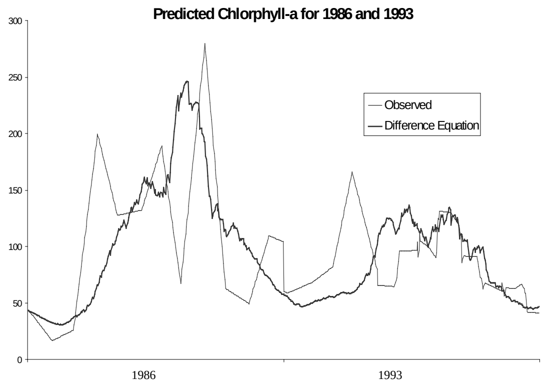

Figure 8. Prediction of Chlorophyll-a using a difference equation with constants calibrated using the training data (RMSE = 46.75).

The success of this equation demonstrates that the constants of process-based models may be successfully calibrated for new freshwater conditions using simple machine-learning techniques.

7. DISCUSSION

The previous studies have demonstrated that models can be developed for the non-linear dynamics of phytoplankton. However a number of basic issues have arisen due to this work.

This work has demonstrated that scaling data affects the form and accuracy of evolved solutions. Developing a general understanding of how scaling affects machine learning techniques seems a valid and relevant research topic that is currently not well understood.

The GP model used the same training and test data setup as the original NN study so that the results could be validly compared. However, there are other approaches to using data sets such as 10-way cross validation, which allow all of the data to be used for developing models and estimating error. Further work is required to determine whether these approaches would improve the overall performance of the learning system.

The fitness function used the standard RMSE. Other error measures for time series, such as the coefficient of efficiency and the mean absolute percentage error may improve the developed models by changing the shape of the fitness landscape. In particular, the large bloom events are being overfitted because of the way RMSE is calculated.

Phytoplankton dynamics are complex and often have different phases based on climate and other population histories. This study has not considered any mechanism for detecting these different phases. A more suitable approach may be to evolve a rule-based system as demonstrated by Bobbin and Recknagel (1999). It allows different phases of growth and decay to be recognised and for different equations to be used for each phase. These phase changes are shown by the inability for our models to predict well for both test years. It is evident that different processes are occuring for each of these years and the generalised model is only able to account for one type of behaviour.

The current study did not exploit the time series nature of the data by allowing past values as input to the evolved equations. Additionally, the growth and decay of phytoplankton are related to the current population of phytoplankton. Equations for population growth which include the current population as part of the next time step prediction are likely to produce more understandable models. Extensions of the GP approach to allow generalised forms of Equation (5) would also be of interest. This approach is likely to produce equations that have underlying process explanation and can be extended to other systems.

8. CONCLUSION

This study has demonstrated that both GP and NN are capable of producing predictive models for ecological time series data. The paper has highlighted issues with scaling data for machine learning and the difficulty involved with producing understandable models. The calibration of a determinstic process-based model using machine learning techniques has been demonstrated. A number of areas for future research have been highlighted.

9. REFERENCES

Bobbin, J. and F. Recknagel, 1999. Mining water quality time series for predictive rules for algal blooms by genetic algorithms. Proc. of the Int. Conference MODSIM 99 (in press).

Gruau, F. 1996. On using Syntactic Constraints with Genetic Programming. In:P. a. K. Angeline, Jr., K.E., (Editor) *Advances in Genetic Programming* 2. 402-417.

Holland, J. H. 1992. Adaptation in Natural and Artificial Systems. Cambridge, Mass.: MIT Press

Koza, J. R. 1990. Concept Formation and Decision Tree Induction Using the Genetic Programming Paradigm. In:H. P. a. M. Schwefel, R., (Editor) Parallel Problem Solving from Nature. 124-129.

Koza, J. R. 1992. Genetic Programming:on the programming of computers by means of natural selection. Cambridge, Mass.:MIT Press

McKay, R. I., Pearson, R.A. and Whigham, P.A. 1997. Learning Spatial Relationships: Some Approaches. In GeoComputation '97. R. T. Pascoe, (Editor), University of Otago, Dunedin, New Zealand. 69-79.

Recknagel, F. 1997. ANNA - Artificial Neural Network model for predicting species abundance and succession of blue-green algae. *Hydrobiologia.* 394:47-57.

Recknagel, F., and J. Benndorf. 1982. Validation of the ecological simulation model SALMO. *Int. Revue ges .Hydrobiol.* 67:113-125.

Recknagel, F., T. Fukushima, T. Hanazato, N. Takamura, and H. Wilson. 1998. Modelling and Prediction of Phyto- and Zooplankton Dynamics in Lake Kasumigaura by Artificial Neural Networks. *Lakes and Reservoirs: Research and Management.* 3:123-133.

Recknagel, F., and H. Wilson. 1999. Elucidation and prediction of aquatic ecosystems by artificial neural networks. *Ecological Modelling.* (in press).

Reynolds, C. S. 1984. The ecology of freshwater phytoplankton. Press Syndicate of the University of Cambridge, New York

Roston, G., and R. Sturges. 1995. A Genetic Design Methodology for Stucture Configuration. *ASME Advances in Design Automation.* DE 82:73-90.

Whigham, P. A., Crapper, P.F. 1999. Time series modelling using genetic programming: An application to rainfall-runoff models. In:L. Spector, Langdon, W.B., O'Reilly, U. and Angeline, P.J., (Editor) Advances in Genetic Programming 3. . MIT Press, Cambridge, MA, USA. 89-104.