

09 PH
W 74 62



Short Term Forecasting of Algal Blooms in Drinking Water Reservoirs using Artificial Neural Networks

A thesis submitted for the award of Doctor of Philosophy

Hugh Edward Campbell Wilson
Discipline of Environmental Biology
School of Earth and Environmental Sciences
The University of Adelaide

April 2004

Abstract

Artificial neural networks (ANNs), trained to make short term forecasts of algal blooms in lakes and rivers, are potentially useful decision making tools for the operational management of eutrophication. This thesis addresses the question of whether a standardised, generic ANN model representation can be developed to achieve this goal. It is argued that four requirements need to be addressed; i) compatibility of models with existing water quality monitoring regimes, ii) stability and repeatability of training outcomes, iii) realistic and meaningful estimates of model performance and iv) explanation of predictions.

ANN model inputs were represented as *summary statistics of sliding time windows*. This approach was shown to increase the compatibility of typical time-series ANN model structures with datasets compromised by missing values and uneven sampling intervals. To improve stability, models were represented as an ensemble of ANNs trained on bootstrap samples of data (ie *bagging* (Breiman, 1994)). It was shown that the average prediction of the bagging ensemble was relatively unaffected by variance of the individual member models. Validation set representation was maximised by use of leave-*k*-out methods. Comparative error measures were devised to illustrate model performance characteristics relative to “naive” controls. A *sensitivity analysis through time* approach was utilised to explain the relative importance of input variables and to account complex interactions between variables.

Training data was available from six sites including Lake Biwa (Japan), Burrinjuck Dam (NSW, Australia), Darling River (NSW, Australia), Lake Kasumigaura (Japan), Myponga Reservoir (SA, Australia) and Lake Soyang (South Korea). These datasets were found to differ significantly from each other in terms of environmental characteristics and data availability. Models were developed to make one and two week forecasts. Predicted variables included chlorophyll *a* concentration and cell counts of the three most abundant algal species for each dataset. Experimental results showed that site/output specific input layers lead to better performance than site/output generic models. Furthermore, it is evident that ANNs capable of non-linear processing generalise better over local (short term) time scales, whereas perceptron models constrained to linear decision boundaries perform better over global (long term) scales.

Contents

1	Introduction	1
2	ANN Model Development	7
2.1	Introduction	7
2.2	Knowledge Representation and Inference by ANNs	7
2.2.1	ANN Structure and Information Processing	7
2.2.2	Supervised Learning by ANNs	12
2.2.2.1	Historical Context	12
2.2.2.2	Backpropagation	12
2.2.2.3	Alternatives to Backpropagation	14
2.2.3	Unsupervised Learning	15
2.3	An ANN Model Development Process-Model	15
2.3.1	Step 1 – Model Design	16
2.3.2	Step 2 – Model Approximation (Training)	19
2.3.2.1	Numerical Conditioning	19
2.3.2.2	Incremental vs Batch-Mode Training	20
2.3.2.3	Training Meta-Parameters: Learning Rate and Momentum	21
2.3.2.4	Local Minima	22
2.3.3	Step 3 – Generalisation	23
2.3.4	Step 4 – Model Validation	26
2.3.5	Step 5 – Knowledge Discovery	27
2.4	ANN Models of Eutrophication Variables	30

2.4.1	Introduction	30
2.4.2	Model Design	30
2.4.2.1	Inputs Describing Nutrient Availability and Chemical Properties	31
2.4.2.2	Inputs Describing Physical Conditions	33
2.4.2.3	Inputs Describing Biological Factors	34
2.4.2.4	Modelling time-series Interactions	34
2.4.3	Model Inference	35
2.4.3.1	Approximation	37
2.4.3.2	Generalisation	37
2.4.4	Validation	39
2.4.5	Knowledge Discovery	40
2.4.6	Discussion and Conclusions	41
2.4.6.1	Choice of Input Variables	41
2.4.6.2	Modelling Time Series	42
2.4.6.3	Approximation and Generalisation	44
2.4.6.4	Validation	45
2.4.6.5	Knowledge Discovery	46
2.5	Proposals for ANN Model Representation	47
2.5.1	An “Input Window” Model Representation	47
2.5.2	Improving Generalisation Qualities by Bagging	48
2.5.3	Model Validation by Rotation Performance Estimators	49
2.5.4	Sensitivity Analysis Through Time	51
2.5.5	“LakeNet” – a Platform for ANN Model Implementation	52
2.6	Conclusion	54
3	Study Sites and Data	57
3.1	Introduction	57
3.2	Study Sites	59
3.2.1	Lake Biwa	59
3.2.2	Burrinjuck Reservoir	62

3.2.3	Darling River	65
3.2.4	Lake Kasumigaura	68
3.2.5	Myponga Reservoir	72
3.2.6	Lake Soyang	75
3.3	A Comparison of Trophic State	79
3.3.1	Discussion and Conclusions	85
3.4	Model Design	87
3.4.1	A Generic ANN Model Design	87
3.4.2	Case Specific ANN Model Design	89
3.5	Conclusion	93
4	Model Complexity and Bagging	95
4.1	Introduction	95
4.2	Methods	96
4.2.1	Model Inputs and Outputs	96
4.2.2	Model Inference	98
4.2.2.1	Training Algorithms	98
4.2.2.2	Numerical Conditioning	99
4.2.2.3	Hidden Layer Configuration	99
4.2.3	Model Validation	100
4.2.4	Computational Platform	101
4.2.5	Experimental Treatments	101
4.2.6	Summary	102
4.3	Results and Discussion	102
4.3.1	Effect of Model Complexity	102
4.3.1.1	Model Error Rates	102
4.3.1.2	The 0 Hidden Unit Models	105
4.3.1.3	Model Variance	105
4.3.1.4	The Overfitting Index	109
4.3.1.5	Reservations	109

4.3.2	Model Aggregation	112
4.3.2.1	Effect of Bagging on Model Performance	112
4.3.2.2	Effect of Bootstrapping on Model Performance	115
4.3.2.3	Reservations	117
4.3.3	Effect of the Training Algorithm and Model Complexity	117
4.3.4	Validation Method	119
4.4	Conclusion	124
4.4.1	Model Approximation	124
4.4.2	Model Generalisation	125
4.4.3	Model Aggregation	125
4.4.4	Model Validation	126
5	The Generic ANN Model	127
5.1	Introduction	127
5.2	Methods	128
5.2.1	ANN Models	128
5.2.2	Model Inference	128
5.2.3	Model Validation	129
5.2.3.1	Continuous Error Measures	129
5.2.3.2	Classification Error Measures	130
5.2.4	Computational Platform	131
5.3	Validation Set Performance	131
5.3.1	Model Performance Evaluation	131
5.3.1.1	Lake Biwa	134
5.3.1.2	Burrinjuck Dam	135
5.3.1.3	Darling River	136
5.3.1.4	Lake Kasumigaura	137
5.3.1.5	Myponga Reservoir	139
5.3.1.6	Lake Soyang	140
5.3.1.7	Summary of Model Performance	140

5.3.2	Effect of Forecast Interval	142
5.3.3	Comparison with ANN Models from the Literature	144
5.4	Sensitivity Analyses	146
5.5	Discussion and Conclusions	149
5.5.1	Performance of the Generic ANN Model	149
5.5.2	Error Measures	151
5.5.3	Sensitivity Analysis	151
6	Identification of Lake Specific ANN Models	153
6.1	Introduction	153
6.2	Methods	155
6.2.1	Data Strip-Mining	155
6.2.1.1	The Initial Model	156
6.2.1.2	Feature Set Reduction	156
6.2.2	ANN Model Identification by Modified Forward Selection	157
6.2.3	Model Inference, Validation and Computation	159
6.2.4	Experimental Treatments	160
6.3	Experimental Results – Data Strip-Mining	160
6.3.1	Model Error Rates	160
6.3.1.1	Lake Biwa	162
6.3.1.2	Burrinjuck Dam	162
6.3.1.3	Darling River	162
6.3.1.4	Lake Kasumigaura	163
6.3.1.5	Myponga Reservoir	163
6.3.1.6	Lake Soyang	164
6.3.1.7	Effect of Input Layer	164
6.3.2	Model Structure	167
6.4	Forward Selection	168
6.4.1	Model Error Rates	168
6.4.1.1	Lake Biwa	168

6.4.1.2	Burrinjuck Dam	168
6.4.1.3	Darling River	169
6.4.1.4	Lake Kasumigaura	169
6.4.1.5	Myponga reservoir	170
6.4.1.6	Lake Soyang	170
6.4.1.7	Effect of Input Layer	170
6.5	Comparing Performance of Model Selection Approaches	174
6.6	Validation Set Performance of the Specific Model	174
6.6.1	Model Performance Evaluation	177
6.6.1.1	Lake Biwa	177
6.6.1.2	Burrinjuck Dam	179
6.6.1.3	Darling River	180
6.6.1.4	Lake Kasumigaura	181
6.6.1.5	Myponga Reservoir	181
6.6.1.6	Lake Soyang	182
6.6.1.7	Summary of Model Performance	182
6.6.2	Interaction of the effects of Input Layer, Hidden Layer and Validation Method on Model Performance	183
6.7	Discussion	187
6.7.1	Data Strip Mining	187
6.7.2	Forward Selection	189
6.7.3	Insights Regarding Time Series ANN Modelling	189
6.7.3.1	Modelling Time Series Interactions	189
6.7.3.2	The “Curse of Dimensionality”	190
6.7.3.3	Effect of Validation Method	191
6.7.3.4	The Effect of Hidden Layer Configuration	192
6.7.3.5	The Effect of Data Availability	194
6.7.4	Reservations about Models and Methods	194
6.7.4.1	Data Strip Mining	194
6.7.4.2	Forward Selection	195

6.7.4.3	Alternative Model Selection Methods	195
6.7.4.4	Consideration of Spatial Information	195
6.8	Conclusions and Recommendations	196
7	Conclusion	199
7.1	Summary of Findings and Recommendations	201
7.2	The Future	204
A	Tactical Responses to Algal Blooms	207
B	Box and Whisker Plots	209
C	Effect of Model Aggregation on RMSE	211
D	Effect of Model Complexity	219
E	Effect of Training Algorithm on RMSE	227
F	Generic Model Predictions	235
G	Classification Stats – 14 day forecasts	243
H	Classification Stats – 7 day forecasts	249
I	Generic Model Sensitivity	253
J	Starting Models for Strip Mining	259
K	Strip Mining – Error Rate Comparison	265
L	Forward Selection	271
M	Specific Model Predictions	277