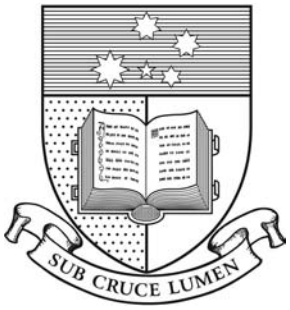

Developing Artificial Neural Networks for Water Quality Modelling and Prediction

Robert James May

BEng (Chem) Hons, BSc

Thesis submitted to The University of Adelaide
School of Civil, Environmental & Mining
Engineering in fulfilment of the requirements
for the degree of Doctor of Philosophy

Abstract

Modelling water quality within complex, man-made and natural environmental systems can represent a challenge to practitioners. Many conventional modelling tools are not capable of representing the complexities of physical and chemical processes often observed in these systems. Consequently, there has been a great deal of interest in the application of computational intelligence techniques, such as artificial neural networks (ANNs). However, "black-box" approaches, such as ANN modelling, are often criticised due to a perceived lack of transparency in the model development methodology. This research has therefore focussed on improving the tools and techniques that are used in the development of ANN models for water quality prediction and forecasting.

The body of research presented in this thesis is described by several peer reviewed articles. These articles describe the theoretical basis and practical context for the ANN model development techniques that have been proposed and applied as a part of this research. Specifically, the ANN development framework has been further enhanced by this research through the development of novel approaches to perform two key tasks: input variable selection (IVS) and data splitting.

The IVS problem is to select variables as ANN inputs from a number of potential candidates, so as to minimise the number of inputs, but maximise the predictive performance of the model. A forward-selection approach for IVS has been examined that is based on partial mutual information (PMI), which can identify an optimal set of variables to use as inputs to ANN models, given a set of candidate variables. Of particular concern is that the use of MI in place of the more traditionally used correlation, provides a more appropriate basis for the selection of inputs based on non-linear relevance. Moreover, the accuracy of MI estimates for a given sample size is difficult to determine. Quantifying the accuracy of MI estimates is necessary to determine critical values of MI, since this forms the basis for of the termination criterion that stops the forward selection process.

Novel termination criteria were developed that alternatively determine the optimum number of candidate input variables. In comparison to the existing ap-

proach, which is based on a computationally expensive, yet potentially inaccurate bootstrap approach, the alternative criteria were found to both reduce the computational requirements and increase selection accuracy of the PMI-based IVS approach, resulting in a much improved algorithm.

Data splitting is an essential part of ANN model development, as the available modelling data must be partitioned into subsets for training, testing and validation. Depending on the data splitting method employed, the data split can have a significant effect on model performance, or reduce confidence in performance assessment. A popular method based on clustering of the self-organizing map (SOM) was examined. The approach was found to be sensitive to SOM size and the manner in which samples are drawn from within the SOM units. However, despite an optimal number of partitions, the SOM can generate partitions that are non-uniformly distributed, and which differ in size and shape. Although conventional rules to increase the sampling rate within larger clusters can reduce variance, the remaining variance can still be significant.

A hybrid algorithm called SOMPLEX was developed, which combines clustering on the SOM, and the DUPLEX algorithm used to perform intra-cluster sampling. DUPLEX is a fully deterministic algorithm that generates a representative sample, regardless of the size or distribution of data within a SOM cluster. For several example applications to predicting water quality, SOMPLEX was found to generate representative data for training, testing and validation, with no variation. The hybrid SOMPLEX approach combines the strengths of the two individual data splitting algorithms, in that the clustering on the SOM reduces the operational complexity, and the DUPLEX sampling improves on random sampling of SOM units to reduce sample variability and increase the representativeness of datasets generated.

In terms of the overall ANN development framework, the outcomes of this research have been an increased understanding of how to best implement ANN techniques, and an appreciation for their place within the context of a water quality modelling toolkit, which comprises both conventional and non-conventional modelling approaches. It was also observed that although the ANN modelling paradigm is quite powerful, it is not without limitations. Many of the limitations and problems encountered with ANN model development are more indicative of the application, rather than the modelling approach itself.

Statement of Originality

I, *Robert James May*, hereby declare that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: ..... Date: .....

Acknowledgement

Above all, I wish to express my profound gratitude towards my supervisors, Professor Graeme Dandy and Professor Holger Maier, for their invaluable guidance, support and encouragement throughout the course of this research.

The author is grateful for the financial support provided by the CRC for Water Quality and Treatment, the technical assistance from the research team for Project 2.5.0.1, and the project leadership of Mr Michael Holmes and Dr Christopher Chow. I thank United Water for their support of this research, in particular Dr John Nixon for his contribution as my industry supervisor, and Dr Stephanie Rinck-Pfeiffer, for her patience while I completed this thesis.

Thanks to the all the staff within the School of Civil, Environmental & Mining Engineering, in particular Dr Stephen Carr for his assistance with software development. The author also wishes to acknowledge that a significant part of this research was aided by the facilities provided by the South Australian Partnership for Advanced Computing.

Many thanks to all of my fellow post-graduate students within the School, for the Friday afternoon drinks and carpet bowls—I blame you all for making the experience so enjoyable that I lost all motivation to finish. Thanks especially to Ms. Gayani Fernando, for sharing a mutual love of neural networks; Dr Matthew Gibbs and Mr Darren Broad, for the more than occasional Tuesday drinks; and I am truly indebted to Mr Michael Leonard for a year of Saturdays spent writing up, and many other enjoyable discussions over the years.

Finally, I acknowledge that credit for any of my past, current and future achievements must be shared with my parents, brother, sister, and my partner, Shannon; for I could accomplish very little without their love and support.

Robert May
Adelaide, 14 October 2009

Contents

Publications

Book chapters

- May, R. J., H. R. Maier, and G. C. Dandy, Development of artificial neural networks for water quality modelling and analysis, in *Modelling of Pollutants in Complex Environmental Systems*, edited by G. Hanrahan, vol. 1, pp. 27–62, ILM Publications, London, UK, 2009.

Journal articles

- May, R. J., G. C. Dandy, H. R. Maier, and T. M. K. G. Fernando, Nonlinear variable selection for artificial neural networks using partial information, *Environmental Modelling and Software*, *23*, 1312–1326, 2008.
- May, R. J., H. R. Maier, G. C. Dandy, and J. B. Nixon, Application of partial mutual information-based variable selection to ANN forecasting of water quality within water distribution systems, *Environmental Modelling and Software*, *23*, 1289–1299, 2008.
- May, R. J., H. R. Maier, and G. C. Dandy, Data Splitting for Artificial Neural Networks Using SOM-based Stratified Sampling, *Neural Networks*, *20*, 283–294, 2010.

Journal articles under review

- May, R. J., H. R. Maier, and G. C. Dandy, SOMPLEX: a hybrid SOM-DUPLEX data splitting algorithm for ANN development, *Submitted to Water Resources Research*

Peer-reviewed conference articles

- May, R. J., G. C. Dandy, H. R. Maier, and T. M. K. G. Fernando, Critical values of a kernel-density based mutual information estimator, in *IEEE International Joint Conference on Neural Networks*, pp. 9997–10,002, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006.

List of Figures

List of Tables

List of Abbreviations

ACF	Auto-correlation function
ACO	Ant colony optimisation
AIC	Akaike Information Criterion
ANN	Artificial neural network
ARMA	Auto-regressive moving-average
ARX	Auto-regressive with exogenous inputs
BIC	Bayesian Information Criterion
BPA	Back-propagation algorithm
CV	Cross-validation
CVI	Cluster validity index
DBS	Density biased sampling
DSS	Data subset selection
EA	Evolutionary algorithm
EANN	Evolutionary neural network
GA	Genetic algorithm
GRIDA	GRNN input determination algorithm
GRNN	Generalised regression neural network
ICA	Independent component analysis
IMC	Inverse model control
IVS	Input variable selection
JMI	Joint mutual information
KDE	Kernel density estimation
MAD	Median absolute deviation from the median
MAE	Mean absolute error
MPC	Model predictive control
MCS	Monte Carlo simulation
MI	Mutual information
MIFS	Mutual information feature selection
MLP	Multi-layer perceptron
MR	Maximum relevance

continued on next page

mR	Minimum redundancy
mRMR	Minimum redundancy–maximum relevance
MRE	Mean relative error
MSE	Mean squared error
PACF	Partial auto-correlation function
PC	Principal component
PCA	Principal component analysis
pdf	Probability density function
PMI	Partial mutual information
PMIS	Partial mutual information-based selection
QE	Quantisation error
RI	Relative importance
RMSE	Root mean squared error
SBSS	SOM-based stratified sampling
SCE	Shuffled complex evolution
SOM	Self-organizing map
SRS	Simple random sampling
SVR	Single variable regression
UVA	Ultra-violet absorbence
VQ	Vector quantisation
WDS	Water distribution system

List of Symbols

General

$a(z_j)$	Activation function
d	Number of variable dimensions
E	Expectation
$f(z_j)$	Hidden node transfer function
$\hat{f}(x), \hat{f}(x, y)$	Density function estimate
F(x)	Process/model transfer function
p	Number of model parameters
r	Coefficient of determination
R	Linear correlation
R'	Partial correlation
V	Variance
W	ANN weight matrix
w_i	ANN weight
X	Random independent/ANN input variable
x	Realisation/observation of X
$\hat{x}$	Model estimate of x
Y	Dependent/ANN output variable
y	Realisation/observation of output variable Y
$\hat{y}$	Model estimate of y
z_j	training input vector (GRNN) / hidden node input (MLP)

Input Variable Selection

B	Bootstrap size
C	Candidate input variable
G	Gaussian kernel function
h	Kernel bandwidth

continued on next page

h_G	Scott reference bandwidth
I	Mutual information
I'	Partial mutual information
k	Number of selected input variables
K_h	Kernel function
$p(x), p(x, y)$	Probability density function
S	Hat matrix defined by $\hat{y} = S(y)$

Data Splitting

$C^{(i)}$	i^{th} cluster
$\mathcal{D}$	Modelling data set
H	Number of strata
k	Number of SOM units $= m \times n$
m	Number of SOM rows
N	Length of dataset D
n	Number of sampled data / number of SOM columns
$n_{\mathcal{S}}$	Number of training data
$n_{\mathcal{T}}$	Number of test data
$n_{\mathcal{V}}$	Number of validation data
S	Silhouette coefficient
$\mathcal{S}$	Test data set
$\mathcal{T}$	Training data set
$\mathcal{V}$	Validating data set

Greek Symbols

ε	Error term
ϕ	Partial auto-correlation
ψ	Golden Ratio ($\sim$1.6)
σ	Standard deviation / GRNN smoothing parameter
μ	Mean