THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

Forecasting Water Resources Variables

Using

Artificial Neural Networks

By

Gavin James Bowden
B.E. Civil/Env (Hons)

Thesis submitted for the degree of
Doctor of Philosophy

The University of Adelaide
School of Civil and Environmental Engineering
Australia

February 2003

Contents

Abstract

Artificial neural networks (ANNs) have shown themselves to be a viable alternative to many traditional water resources models, particularly in the field of forecasting hydrologic variables. However, it has been acknowledged by a number of researchers that there is currently no systematic methodology available for the development of ANN models. In this research, a methodology is formulated for the successful design and implementation of ANN models for water resources applications. Attention is paid to each of the steps that should be followed in order to develop an optimal ANN model. These steps include: determining when ANNs should be used in preference to more conventional statistical models; dividing the available data into subsets for modelling purposes; deciding on a suitable data transformation; determination of significant model inputs; choice of network type and architecture; selection of an appropriate performance measure; training (optimisation) of the network's weights; and deployment of the optimised ANN model in an operational environment.

The developed methodology is successfully applied to two water resources case studies, namely, the forecasting of salinity in the River Murray at Murray Bridge, South Australia and the forecasting of cyanobacteria (*Anabaena* spp.) in the River Murray at Morgan, South Australia. These case studies are used to test and illustrate the methods proposed for each step in the methodology. Guidelines for the use of these methods are presented.

Two state-of-the-art statistical tests for nonlinearity are developed as tools for determining when it is appropriate to proceed with an ANN model, in preference to a conventional linear time series model. Using these tests, the salinity and cyanobacteria time series data are both found to exhibit significant nonlinear serial dependence, thereby justifying the use of ANNs for these case studies.

The way in which the available data are divided into training, testing and validation sets has a significant influence on the performance of an ANN model. Therefore, two methods for dividing data into representative subsets are developed, namely, a genetic algorithm (GA) and a self-organizing map (SOM). Both data division methods are able to lead to the development of ANN models that are capable of producing consistent forecasting results.

A number of data transformations are investigated to determine their effect on ANN performance. Three new transformations are proposed, including, a transformation for removing seasonality from data, a transformation for producing normally distributed

data and a kernel transformation for producing uniformly distributed data. The results from this research reveal that the most suitable transformation is case study dependent.

Three analytical techniques are developed to determine significant model inputs for ANNs. The first technique utilises a partial measure of the mutual information criterion to characterise the dependence between a potential model input and the output variable. The remaining techniques utilise unsupervised approaches (i.e. principal component analysis (PCA) and a SOM) to reduce the input dimensionality, and a hybrid GA and general regression neural network (GRNN) to select inputs that have a significant influence on the model's forecast.

Two types of feedforward networks are considered in this research, namely, the multilayer perceptron (MLP) and the GRNN. An evolutionary method for selecting an appropriate architecture for a MLP is investigated and these models are compared with GRNN models for the salinity and cyanobacteria case studies. The GRNN is found to give superior performance when tested on data within the training domain, whereas the MLPs are more robust when uncharacteristic data are encountered. Two architectural modifications are investigated for the GRNN model and a stochastic optimisation technique inspired by the behaviour of ant colonies is developed as an approach for training GRNNs with multiple sigma weights.

A framework is developed for deploying operational ANN models in a real-world environment. Various retraining strategies are investigated and two new methods for updating ANN models in real-time are developed. These methods are able to provide significant improvement in model performance when compared to the scenario of not retraining the model.

The developed methodology results in ANN models that are successful in providing salinity forecasts, 14 days in advance. ANN models are also successfully developed for predicting the timing and relative magnitude of significant growth events of *Anabaena* spp., 4 weeks in advance, provided that representative data are available in the calibration set. For the cyanobacteria case study, the causal variables are found to interact with concentrations of *Anabaena* in a very complex manner and the most important input variables are temperature, flow, pH and previous values of *Anabaena*.