AOGS 2005

Data-driven and data-mining approaches for selecting significant input variables

NITIN MUTTIL[1] and AMITHIRIGALA W. JAYAWARDENA[2]

[1]*Research Associate, Department of Civil Engineering, University of Hong Kong, Hong Kong*
[2]*Senior Lecturer, Department of Civil Engineering, University of Hong Kong, Hong Kong*

Data is a valuable resource. Data with high quality, i.e., which is representative enough to adequately describe the physical phenomena, provides the key to the use of data-driven approaches. Thus, in any data-driven model development process, familiarity with the available data is of the utmost importance, a major step of which consists of selecting significant variables to be used as input. Lack of optimal choice of input variables for the optimization of appropriate model structure would lead to use of spurious inputs, which would contaminate the model with noise. Maier and Dandy [1] have reviewed 43 international journal papers, which used the popular data-driven approach, artificial neural networks for modelling and forecasting of water resources variables. They concluded that in many papers reviewed, the modelling process is carried out incorrectly and one of the main areas of concern included arbitrary selection of model inputs.

The choice of input variables is generally based on a priori knowledge of causal variables and physical insight into the problem being studied and if the relationship to be modeled is not well understood, then analytical techniques can be used. In this study, the application of various artificial neural network and genetic programming models are presented to select ecologically significant variables for the prediction of chlorophyll-a, a measure of algal biomass, in Tolo Harbor, Hong Kong, which is well known for frequent occurrences of algal blooms and red tides. The water quality data used is the comprehensive biweekly data collected as part of the routine water quality monitoring programme of the Hong Kong Environmental Protection Department (EPD). A detailed analysis of the different models, which include the study of the weights of a trained neural network, the interpretation of genetic programming equations and use of statistical techniques are employed. The study shows that the used models correctly identify the key input variables in accordance with ecological reasoning.

Keywords: Data-driven models; significant input variables; algal blooms

References

[1] H. R. Maier and G. C. Dandy, Neural networks for the predication and forecasting of water resources variables: a review of modelling issues and applications. Env. *Modelling & Software*, **15**, 101-124, (2000).