Desalination and Water Treatment
www.deswater.com
doi: 10.1080/19443994.2015.1112981

(2015) 1–13

Taylor & Francis
Taylor & Francis Group

Developing alternative regression models for describing water quality using a self-organizing map

Seo Jin Ki[a], Seung Won Lee[b], Joon Ha Kim[a,c,*]

[a]*School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea, Tel. +82 62 7153277; Fax: +82 62 7152434; email: joonkim@gist.ac.kr (J.H. Kim)*
[b]*Environmental and Plant Engineering Research Institute, Korea Institute of Civil Engineering and Building Technology, Gyeonggi-do 10223, Korea*
[c]*Sustainable Water Resource Technology Center, GIST, Gwangju 500-712, Korea*

Received 15 September 2015; Accepted 8 October 2015

ABSTRACT

Statistical models play an important role in elucidating the dynamic behaviors of surface water quality, given limited data on a large scale. In this study, we examine alternative approaches to develop regression models that predict fecal coliform (FC) concentrations in a river using different methods for selecting important variables provided by a self-organizing map (SOM). The raw data used as input to the SOM included 11 water quality, 6 meteorological, and 7 land use parameters that were monitored along the Yeongsan River in Korea on various time scales (from daily to half a decade) during 1996–2008. In both test and validation data sets, (multiple) regressions using backward elimination were compared against regression models via forced entry, which included a set of ranked variables simultaneously based on four indices in the SOM (i.e. structuring index, relative importance, cluster description, and Spearman's rank correlation). Results showed that the SOM effectively illustrated the complex relationship between FC and the remaining variables in the entire data set. This relationship was seen more clearly in homogeneous clusters, indicating that the regression models became more robust in each subdivided group. While the original backward elimination model (R^2 = 0.66) had much better performance than the models with four indices (R^2 = 0.40–0.45) in the test data set, its performance (R^2 = 0.42) was quite comparable to the relative importance model (R^2 = 0.38) in the validation data set. Based on this preliminary study, we recommend further investigation of these indices for a reliable regression analysis, as the t values currently used for the variable selection in regressions provide only a locally optimal solution for the final model. The proposed methodology, if verified successfully, would be useful in developing early warning models that control mortality or disease rates of fishes in high-density aquafarms via water quality.

Keywords: Regression models; Self-organizing map; Variable selection; Relative importance; Fecal coliform; Water quality data sets

*Corresponding author.