River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): a case study at Klang River, Malaysia

ABSTRACT

A collective set of data over five years (2003 to 2007) in Klang River, Selangor were studied in attempt to assess and determine the contributions of sources affecting the water quality. A precise technique of multiple linear regressions (MLR) were prepare as an advance tool for surface water modeling and forecasting. Likewise, principle component analysis (PCA) was used to simplify and understand the complex relationship among water quality parameters. Nine principle components were found responsible for the data structure provisionally named as soil erosion, anthropogenic input, surface runoff, fecal waste, detergent, urban domestic waste, industrial effluent, fertilizer waste and residential waste explains 72% of the total variance for all the data sets. Meanwhile, urban domestic pollution accounted as the highest pollution contributor to the Klang River. Thus, the advancement of receptor model was applied in order to identify the major sources of pollutant at Klang River. Result showed that the use of PCA as inputs improved the MLR model prediction by reducing their complexity and eliminating data collinearity where $R2$ value in this study is 0.75 and the model indicates that 75% variability of WQI explained by the five independent variables used in the model. This assessment presents the importance and advantages poses by multivariate statistical analysis of large and complex databases in order to get improved information about the water quality and then helps to reduce the sampling time and cost for reagent used prior to analyses.

Keyword: Water quality; Multiple linear regression; Principle component analysis