



Input variable selection and calibration data selection for storm water quality regression models

Siao Sun¹, Jean-Luc Bertrand-Krajewski²

¹ University of Lyon, INSA Lyon, LGCIE, France, siao.sun@insa-lyon.fr

² University of Lyon, INSA Lyon, LGCIE, France, jean-luc.bertrand-krajewski@insa-lyon.fr

ABSTRACT

Storm water quality models are a useful tool in storm water management. Interests have grown in analyzing existing data for developing models for urban storm water quality evaluations. It is important to select appropriate model inputs when many candidate explanatory variables are available. Model calibration and verification are essential steps in any storm water quality modelling. This study investigates input variable selection and calibration data selection in storm water quality regression models. The two selection problems are mutually interacted. A procedure is developed in order to fulfil the two selection tasks in order. The procedure firstly selects model input variables using a cross validation method. An appropriate number of variables are identified as model inputs to ensure that a model is neither overfitted nor underfitted. Based on the model input selection results, calibration data selection is studied. Uncertainty of model performances due to calibration data selection is investigated with a random selection method. An approach using cluster method is developed in order to enhance model calibration practice based on the principle of selecting representative data for calibration. The comparison between results from the cluster selection method and random selection shows that the former can significantly improve performances of calibrated models. It is found that the information content in calibration data is important in addition to the size of calibration data.

KEYWORDS

Calibration, cluster method, input selection, overfitting, regression, storm water quality modelling