Input Variable Selection and Calibration Data Selection for Storm Water Quality Regression Models

Siao Sun and Jean-Luc Bertrand-Krajewski
Outline

• Problem statement
• Case and data
• Methods
• Results and discussion
• Conclusions
Problem statement

• Storm water quality models are a useful tool in storm water management.
• Interests grow in analyzing existing data to develop models.
• Regression for storm water quality modeling is a common method.
Problem statement

Context: Regression model for modeling storm water quality.

With numerous measured events, questions are:

1) Inputs selection
2) Calibration data selection
Case and data
Case and data

- **Saône River**
- **Rhône River**

- **Ecully catchment**
  - Residential
  - Combined sewer
  - 245 ha
  - 60 active ha
  - 42% impervious
Case and data

- 239 storm events between 2004-2008
- Event TSS load (kg) as storm water quality index
- Regression model

\[ y = \sum_{i=1}^{N} b_i x_i + b_0 + e \]

- Output TSS
- Inputs

56 potential explanatory variables
Shall we use all 56 variables as model inputs?

- Calibration (simulation ability) and verification (prediction ability) were performed.
- When splitting data, uncertainty due to calibration data selection exists.
- Cross validation was used.
Results - Input selection

7 variables are selected
Results - Input selection

7 variables are selected

- Antecedent dry period from the last rainfall event exceeding 5 mm
- Antecedent dry period from the last rainfall event exceeding 30 mm
- Maximum rain intensity in 10 minutes during 12 hours before the event
- Maximum rain intensity in 10 minutes during 72 hours before the event
- Maximum rain intensity in 30 minutes during 4 hours before the event
- Maximum flow rate
- Total flow volume
Methodology - Calibration data selection

• Random selection

- The number of calibration events: 8-239

- For each number, calibration events are randomly selected for many times to study the uncertainty due to calibration data selection

- A calibrated model is evaluated by calibration data sets, verification data sets and all data sets
Results - Calibration data selection

- Random selection

(a) Calibration

(b) Verification

(c) All data

Number of Events used for calibration
Results - Calibration data selection

- Random selection

\[\text{Number of Events used for calibration} \]

\[\text{NS values} \]

(a) Calibration

(b) Verification

(c) All data

Siao Sun, Jean-Luc Bertrand-Krajewski
INSA Lyon, LGCIE, France
Methodology - Calibration data selection

• Select representative data for calibration using cluster method

  ➢ Divide all events into $n$ clusters if $n$ events is wanted for calibration

  ❖ A cluster contains data sets of similarity according to standardized Euclidean distance between data sets

  ➢ One data set is selected from a cluster to represent the cluster
Results - Calibration data selection

- Cluster selection

(a) Calibration

(b) Verification

(c) All data

Siao Sun, Jean-Luc Bertrand-Krajewski
INSA Lyon, LGCIE, France
Results - Calibration data selection

- Cluster selection

(a) Calibration

(b) Verification

(c) All data
Results - Calibration data selection

- Cluster selection

(a) Calibration

- Verification

(c) All data

Siao Sun, Jean-Luc Bertrand-Krajewski
INSA Lyon, LGCTE, France

9th International Conference on Urban Drainage Modelling Belgrade 2012
Results - Calibration data selection

• Cluster selection

(a) Calibration

(b) Verification

(c) All data

Siao Sun, Jean-Luc Bertrand-Krajewski
INSA Lyon, LGCIE, France

9th International Conference on Urban Drainage Modelling Belgrade 2012
Conclusions

1. Overfitting occurs when too many inputs are considered in a model

2. Data used for calibration can affect model behaviors

3. A cluster method can effectively aid choosing representative calibration data
Thank you for your attention!

Input Variable Selection and Calibration Data Selection for Storm Water Quality Regression Models

Siao Sun and Jean-Luc Bertrand-Krajewski

Siao Sun, Jean-Luc Bertrand-Krajewski
INSA Lyon, LGCIE, France