Automatic data quality assessment: Practical application using in situ measurement stations for river water quality monitoring

World Water Congress & Exhibiton

Busan, Korea

16-20 SEP 2012

Janelcy Alferes, Pascal Poirier and Peter Vanrolleghem

Canada Research Chair in Water Quality Modeling

UNIVERSITÉ LAVAL

modelEAU

Problem definition

Effective management of water bodies

↓

Reliable water quality information

↓

Integrated evaluation

↓

Monitoring and assessment strategy

↓

Modelling, decision making, control

Problem definition

- In situ **monitoring stations**
  - Information-rich data sets ✓
  - Pollution dynamics ✓
  - Reduce costs ✓
  - Huge/complex data sets X
  - Errors and uncertainties X
  - Reliability of sensors insufficient X

➡ **Data evaluation/validation is crucial**

In situ monitoring stations

- Urban river (Canada)
- Water quality variables:
  - pH, µS/cm, O_2, NH_4, NO_3, TSS...
- Sample time: 5-60 sec
- Practical issues:
  - Maintenance, fouling, clogging...

⬇

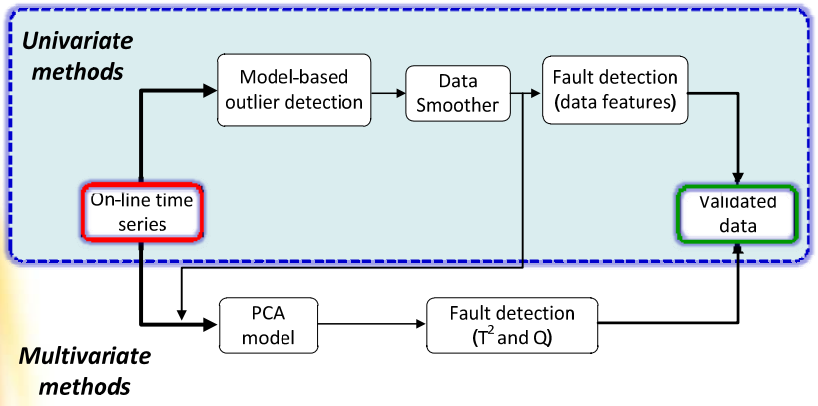
Representative data??

Data quality assessment tools

- Current practice --» manual procedures
- Automatic data quality evaluation:
  - Corrupted, doubtful, unreliable
  - Outliers
  - Noise
  - Sensor faults

Using time series information!

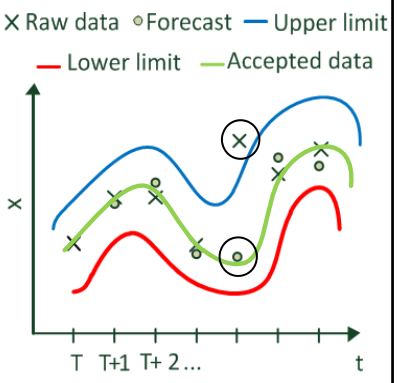
Data quality assessment tools



Data quality assessment tools

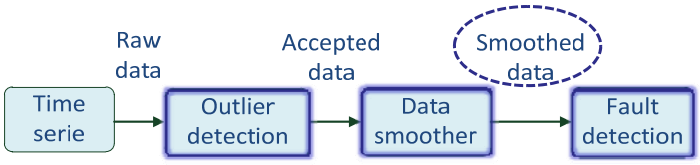
- Univariate methods
  - Outlier detection
  - Autoregressive models
  - At T forecasting (T+1):
    - variable $\hat{x}$
    - std of error $\hat{\sigma}_e$
  - Prediction interval:

$$x_{\lim} = \hat{x} \pm K \cdot \hat{\sigma}_e$$



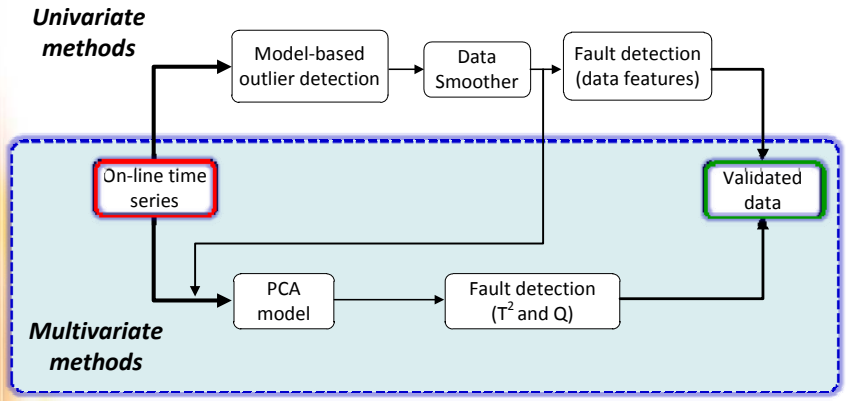
Data quality assessment tools

- Univariate methods



  - Fault detection
    - % replaced data --» data goodness
    - Slope --» rate of change
    - Residuals correlation --» good fit of model
    - Residual std (RSD) --» variance

Data quality assessment tools

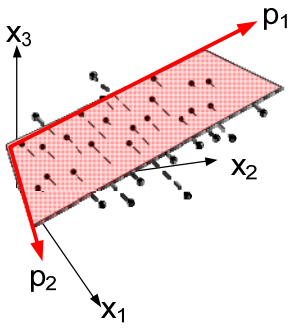


Data quality assessment tools

- Multivariate methods
  - Reduce dimension of X identifying "key" variables.
  - Ex. Let be 3 variables (x_1,x_2,x_3)
  - New variables (p_1,p_2) as linear combinations:

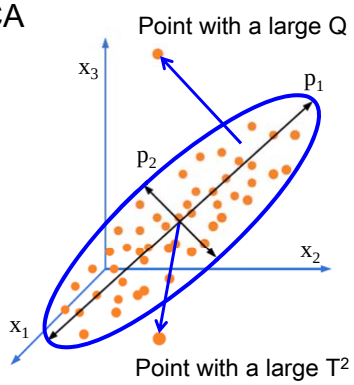
$$p_1 = c_{11}x_1 + c_{12}x_2 + c_{13}x_3$$
$$p_2 = c_{21}x_1 + c_{22}x_2 + c_{23}x_3$$

  - Axes of a new coordinate system
  - Directions of max. variability



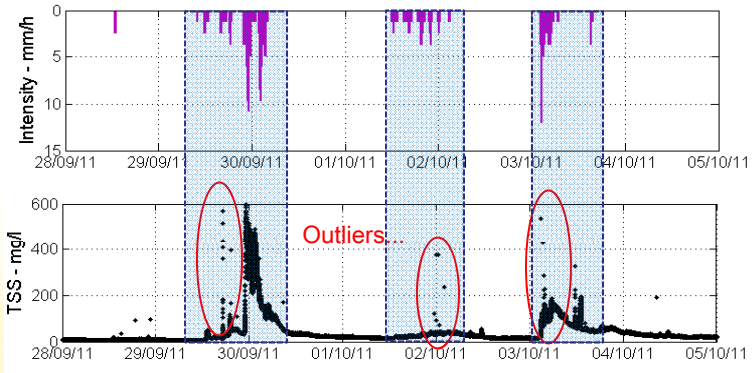
Data quality assessment tools

- Multivariate methods
  - Fault detection within the PCA space
    - **T^2**: normalized sum of scores: variations within the model
    - **Q**: sum of squared residuals: goodness of fit of samples to the model
    - Detection limits are defined.



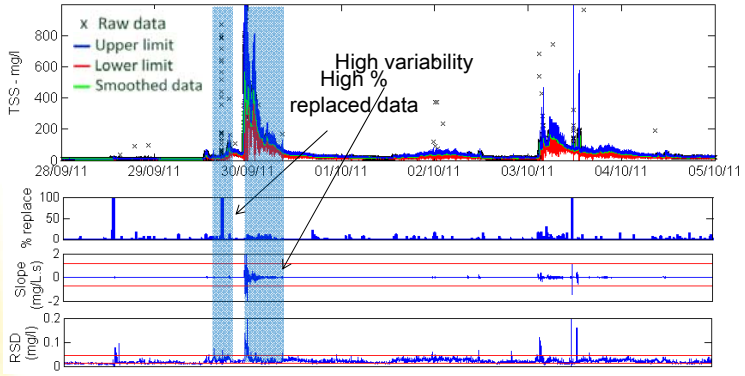
Results

- Univariate methods (TSS)



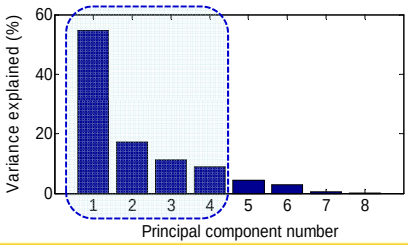
Results

- Univariate methods (TSS)



Results

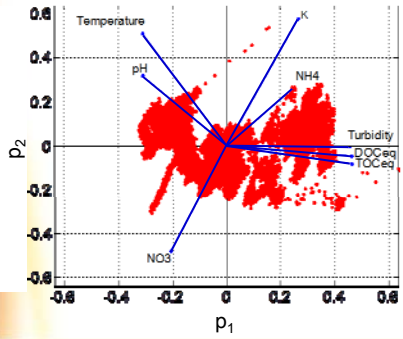
- Multivariate methods
  - Dataset with 8 variables (8 dimensional data space)
    - NTU, NO_3, TOC, DOC, pH, K^+, NH_4 and °C
  - Firsts 4 components (p_1...p_4) > 90% variability



Some preliminary results

- Multivariate methods

Representation of data in the new space

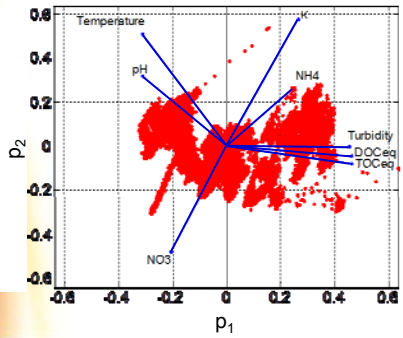


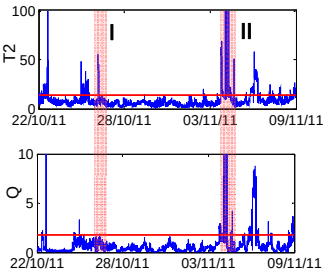
- Vectors represent the variables and their contribution to p_1 and p_2
- Each point corresponds to a sample in the new space (scores)

Some preliminary results

- Multivariate methods

Representation of data in the new space

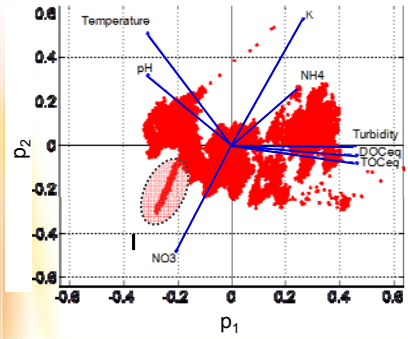


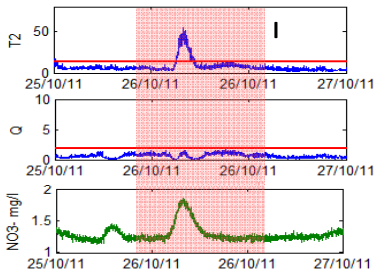


Some preliminary results

- Multivariate methods

Representation of data in the new space

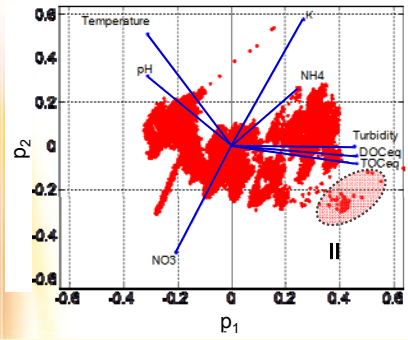


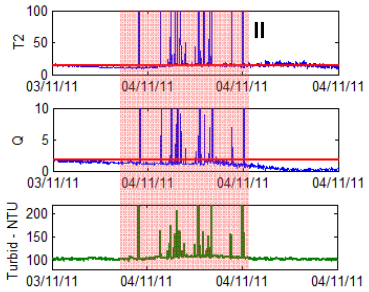


Some preliminary results

- Multivariate methods

Representation of data in the new space





Conclusions

- Data quality assessment tools satisfactorily validated
- Univariate methods allowed creating “good” time series and detecting individual faults
- Multivariate methods allowed dimension reduction and detection of multiple faults

Acknowledgement

Canada Research Chair in Water Quality Modeling

Primodal
Earth Essential Technologies

Fondation canadienne pour l'innovation
Canada Foundation for Innovation