UNIVERSITY of OULU
OULUN YLIOPISTO

DATA QUALITY AND RELIABILITY

Antti Koistinen
antti.koistinen@oulu.fi

Control Engineering Research Group
Faculty of Technology, University of Oulu

OUTLINE

- Data sources
- Data for decision making
- Recursive outlier detection method
- Summary

UNIVERSITY of OULU
OULUN YLIOPISTO

MEASUREMENTS

- Measurement uncertainty is a vital part of measurement data
  - Uncertainty model
  - Reproducibility, Representativity, sampling errors, variations…
  - Sampling can be the largest uncertainty component in biological and chemical measurements
- Sensor placement must be carefully assessed
- Sensor type needs to be suitable for the task
  - Accuracy, fouling, calibration needs
- Reliable measurement data can be connected to modelling and open data for creating new information
- Continuous and reliable monitoring requires maintenance

SIMULATION AND MODELS

- Measurement data can be used in combination with simulations and models to form predictions
  - User must be aware of the uncertainties considering the model
- Simulation can aid in estimating of water stream composition further away from the process
  - Use of simulation data in place of missing measurements
  - Indication for more accurate inspection
- Simulated projections can be used for preventive actions (early warning)

CROWD SOURCING

- Data acquired through Crowd sourcing must contain enough measurement points to be reliable
- Measurement system must be such that different users can easily take comparable measurements
- Single measurement points are not reliable, but several that are properly sampled from describing locations form a solid monitoring method
- System can provide a good and fast indicator of changes in vast areas

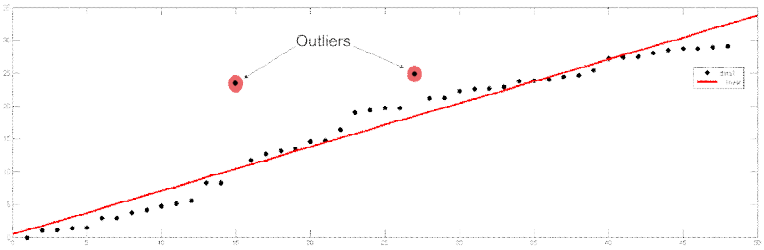
DATA FOR DECISION MAKING

- Decision making needs valid describing information that enables the choosing of valid actions
  - The user must confirm that the data is suitable for its intended use
  - Data should be inspected for outliers
  - Measurement uncertainty information may need to be connected to values when used in decision making
- Data processing method can have a key impact on the decision forming
- Data should be presented in an understandable way
  - User must have knowledge about the meaning of numerical values
  - Graphic indicators or words such as “high” or “low”

7

OUTLIERS

- Outlier detection is done in order to find data points generated by a different mechanism
  - Statistical, Depth-based, deviation-based, distance-based, density-based, high-dimensional…



8

OUTLIER DETECTION

- Recursive statistical method for open data outlier detection
  - Six-sigma-rule used for the control limits
  - Control limits are created each time step for each process variable
  - Correlation between signals are taken into account with recursive alarm formula
  - Coefficients can be chosen according the data

9

OUTLIER DETECTION

- Moving average

$$\mu_t = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}(x_t - x_k)}$$

x_t=signal value
x_k=k[th] data point of the sliding window
N=frame size
μ_t=moving average

10

OUTLIER DETECTION

- Moving standard deviation

$$\sigma_t = \sqrt{\frac{1}{N-1}\sum_{k=0}^{N-1}(x_t - x_k - \mu_t)^2}$$

x_t=signal value
N=frame size
μ_t=moving average

OUTLIER DETECTION

- Control limits

$$CL_t = c_t[d \cdot \sigma_t + d_{DB} \cdot \tilde{x}_{DB} + d_{MM} \cdot (max_{DB} - min_{DB})]$$

$$UCL_t = \mu_t + CL_t$$
$$LCL_t = \mu_t - CL_t$$

$\tilde{x}_{DB}$ = median of moving standard deviations
d, d_{DB}, d_{MM} = user defined coefficients
max_{DB}, min_{DB} = maximum and minimum values of signal

Coefficient c_t is updated automatically depending of alarm frequency in signal i

LCL_t, UCL_t = lower and upper control limits

OUTLIER DETECTION

- Control limit

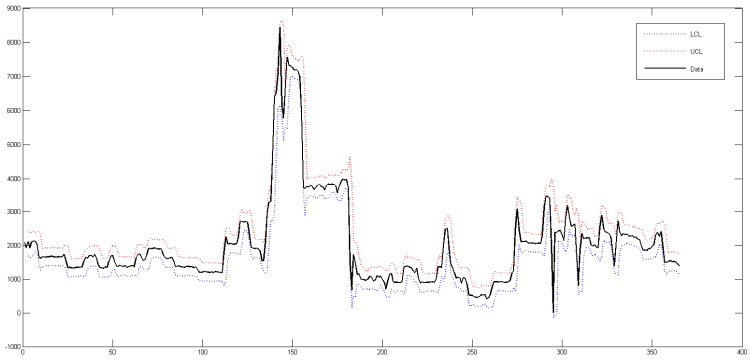
$$c_t = \frac{1 + A_{all}(t) + A_i(t)}{1 + A_{all}(t)}$$

A_{all} = all alarms in certain process area
A_i = alarms of the variable i in a certain time period

13

OUTLIER DETECTION

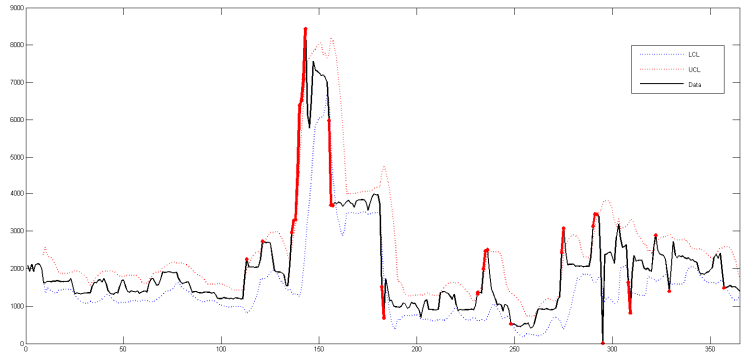
- Control limits formed with 3 previous data points



14

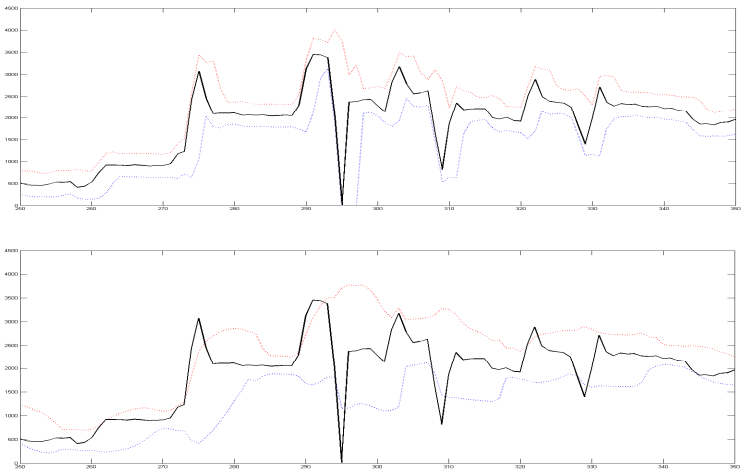
OUTLIER DETECTION

- Control limits formed with 9 previous data points



15

OUTLIER DETECTION



16

SUMMARY

- Statistical methods can be used in evaluating of data quality
  - Selecting of method and parameters play important role
- Decision making should use all the relevant data available in order to get better understanding of related events
- Less accurate measurements and simulated projections can indicate changes in trends and need for further inspection