UNIVERSITY of OULU
OULUN YLIOPISTO

MODELLING THE EFFLUENT QUALITY UTILIZING OPTICAL MONITORING

10TH FINNISH ENVIRONMENTAL MONITORING DAYS 17.9.2015
Viikinmäki wastewater treatment plant

INTRODUCTION

- The optical monitoring variables together with process measurements were utilised to predict the suspended solids in biologically treated wastewater.
  - Variables were selected from as an early stage of the process as possible.

  -> Proactive information on the quality of biologically treated wastewater.
- The optimal subset of variables for model development was searched using five variable selection methods.
  - Correlation-based selection, Forward selection, Stepwise regression (Matlab), Successive projections algorithm (SPA), Genetic algorithm (GA).
  - Selections were performed based on mathematical ground only, any deterministic models or chemical or biological knowledge were not taken account.
  - Manual selection was based on the visual inspection of data, expert knowledge and trial and error.
- The data was collected from a period of over one year.
  - Seasonal changes (temperature, heavy rains, melting snow, variations in the quality and quantity of wastewater).
  - Optical monitoring was carried out daily, laboratory measurements only a few times a week. The missing laboratory data was not interpolated.

  -> The total amount of data was 97 points.

NONLINEAR SCALING

- The dataset was scaled between **[**−2, +2] using the nonlinear scaling method based on generalized moments, norms and skewness.
- The feasible range
  - Support area: the min and max of the values of the variable *x*.
  - Core area [*cl, ch*]
  - Central tendency value *c* divides the value range.
    - The point where the skewness changes from positive to negative.
    - The estimates of the corner points are the points where the direction of the skewness changes for the lower and upper data set.

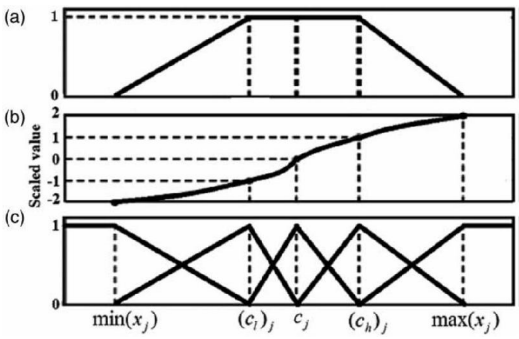


Figure. (A) The feasible range, (B) scaled value and (C) membership functions.

[Juuso E. Integration of intelligent systems in development of smart adaptive systems: linguistic equation approach [dissertation].Oulu: Acta Universitatis Ouluensis. Series C, Technica 476. Dissertation. 258. 2013. http://urn.fi/urn:isbn:9789526202891]

VARIABLE SELECTION

- One of the most important steps in the data analysis and model development.
- Only significant variables must be selected.
  - Over-fitting, increased computational complexity in training and bad prediction results.

- Wrapper methods
  - A subset of variables is assessed according to their usefulness to a given predictor.
  - Wrapper methods wrap around an appropriate learning machine which is employed as the evaluation criterion, such as prediction or classification error.
  - Wrappers often give better results but are slower than filters.
  - For example forward selection and genetic algorithms.
- Filter methods
  - Variables are selected or deleted according to the formed ranking which is based on the correlation coefficients.
  - Very efficient but the model is seldom optimal.
  - For example correlation-based selection and successive projections algorithm (SPA).

CROSS VALIDATION

- The quality of a developed model depends highly on the quality and length of the dataset.
- Typical resampling method cross-validation is one way to predict the fit of a model for a validation set when dataset is small and an explicit validation set is not available.
  - Leave one out (LOO)
  - Leave multiple out (LMO)
  - *k*-fold
    - The whole data set is used for training and validating the model.
    - The original dataset is randomly partitioned into *k* subsets of equal size.
    - One subset is used as a validation data for testing the model and the remaining *k*–1 subsamples are used as training data. The process is repeated *k* times and each of the subsets is used only once as the validation data.
    - A single estimation is produced by combining (averaging) the *k* results.
    - Optimally *k* is between 5 and 10.

RESULTS

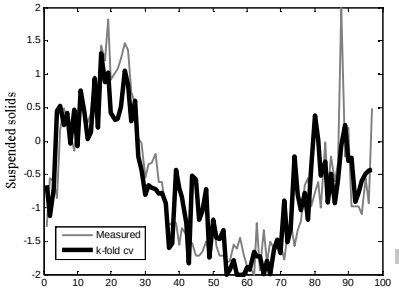
- Altogether, 14 variables were selected in different subsets.
- Three of five variable selection methods gave similar subsets:
  - Fractal dimension, influent total nitrogen and sulphate, mechanically treated wastewater iron and nitrate nitrogen, and anoxic proportion.

  -> The optimal model with this dataset(?)

Correlation analysis	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Stepwise selection	Manual selection	Variables	
3	3	3	1	3	3	1	Total floc area
13	13	8	13	13	8	2	Amount of filaments
6	12	12	9	12	12	3	Fractal dimension
2	8	9	12	8	9	4	Aspect ratio
4	9	11	11	9	11	5	Median area of objec
12	11	13	4	11	13	6	Number of small obj
14	7		2			7	(I) Suspended solids
5			10			8	(I) Total nitrogen
						9	(I) Sulphate
						10	(M) total nitrogen
						11	(M) Iron
						12	(M) nitrate-nitrogen
						13	Anoxic proportion
						14	Temperature

(I) influent, (M) mechanically treated wastewater

Variable selection method	R^2	RMSE
Correlation analysis	0.71	0.55
Forward selection	0.77	0.49
Genetic algorithms	0.76	0.49
Successive projections algorithm +GA	0.71	0.55
Stepwise selection	0.77	0.49
Manual selection	0.77	0.48



RESULTS

- The model can be used to predict the level of the suspended solids and to show the quality of the biologically treated wastewater hours in advance in comparison to laboratory analysis.
- The result was different than in the earlier study.
  - R^2 0.80 and RMSE 0.53; the amount of filaments, mean and median area of objects and incoming wastewater iron. [Tomperi J., Koivuranta E., Kuokkanen A., Juuso E., Leiviskä K. (2015) Real-time optical monitoring of the wastewater treatment process, Environmental Technology, DOI:10.1080/09593330.2015.1069898]
  - The seasonal changes add variation and noise to data but longer dataset gives a more reliably analysis of the process operation because many factors affecting the quality of sludge and purification process are dependent on the temperature.
    - Certain variables are always in the important role, but models are not generalizable and the models should be actively updated.
- The result can be considered satisfactory because the optical monitoring was done only from one of nine heterogeneous process lines whereas the suspended solid samples included wastewaters from all lines.
- The objective, continuous and fast on-line optical monitoring method is a valuable tool for monitoring the wastewater treatment process, receiving new information and combined to predictive modelling it has potential to be used in the process control, keeping it in stable operating conditions and avoiding environmental risks.

THANK YOU!

Jani Tomperi
jani.tomperi@oulu.fi

Control Engineering
P.O Box 4300
90014 University of Oulu