DATA MINING TO SUPPORT ANAEROBIC WASTE WATER TREATMENT PLANT MONITORING AND CONTROL IN THE TELEMAC PROJECT

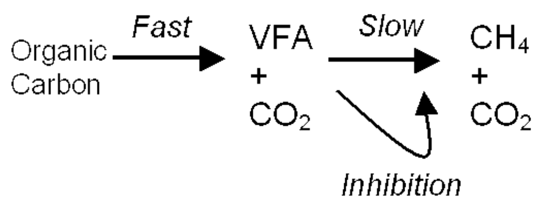
Maurice Dixon [**], Julian Gallop [*], Simon Lambert [*], Laurent Lardon [***], Jean-Phillipe Steyer [***], and Jerome V. Healy [**]

** Business and Information Technology Department, CCLRC Rutherford Appleton Laboratory, UK*
*** London Metropolitan University, UK*
**** Laboratoire de Biotechnologie de l'Environnement - INRA, France*

ANAEROBIC DIGESTION AND TELEMAC PROJECT

- TELEMAC aims to improve the process of treating waste products from alcohol production processes.
- Anaerobic digestion offers a rapid rate with high throughput.
- It degrades concentrated and difficult substrates.

The chemistry of anaerobic digestion



But:

- Risk of unstable states in digester.
- So typically operated at low efficiency to avoid problems.
- Expert knowledge required.

The promise of data mining:

- Characterisation of current and imminent digester states, especially consequences of organic overload/underload and hydraulic overload.
- Sensor ranking/modelling in cases of sensor omission or failure.
- Sharing and adaptation of rules/expertise between plants.

PREDICTION AND SENSOR VALUES

- Sensor availability is affected by expense and reliability.
- Is it possible to substitute for some sensor values by others?

Data mining methods:

- Data filtering in a plant specific fault detection and isolation system. Applied to single sensors and to consistency between multiple sensor readings.
- Data visualization to provide pairwise multivariable displays.
- Linear Regression using both forward and backward stepping regression to rank sensors in terms of incremental improvement of prediction.
- Non linear regression with neural nets using inputs ranked by expert judgment.

Key results:

- Very high linear correlation between COD, TOC and VFA in the digester, eg COD can be predicted from VFA alone with an R^2=0.91.
- Non linear regression is needed to enable prediction of VFA using data from more readily available sensors; R^2=0.95 (R^2=0.45 for linear regression). The corresponding figures for COD are R^2=0.92 and R^2=0.28.
- Prediction risk for COD increases substantially unless more specialised sensors are present.

COD – chemical oxygen demand;	qin – input flow rate
VFA – volatile fatty acid	pHdig – pH in digester
TOC – total organic carbon	qgas – output gas flow

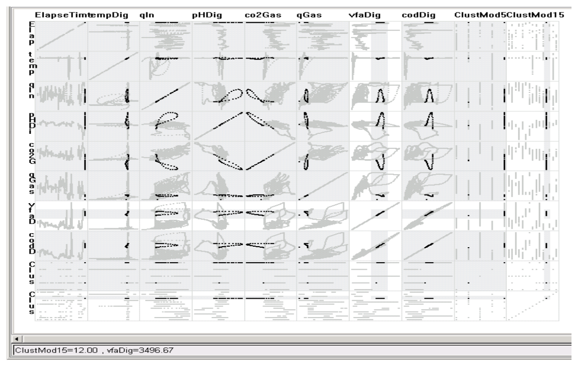
ACKNOWLEDGEMENTS

The TELEMAC project is funded under the European Commission's IST programme as project IST-2000-28156.

CLUSTER ANALYSIS

Cluster analysis identifies subsets showing strong self-similarity. We measure variable compactness, inclusion and precedence.

Clusters from two different runs



MODELLING ERROR BOUNDS

Prediction Intervals:

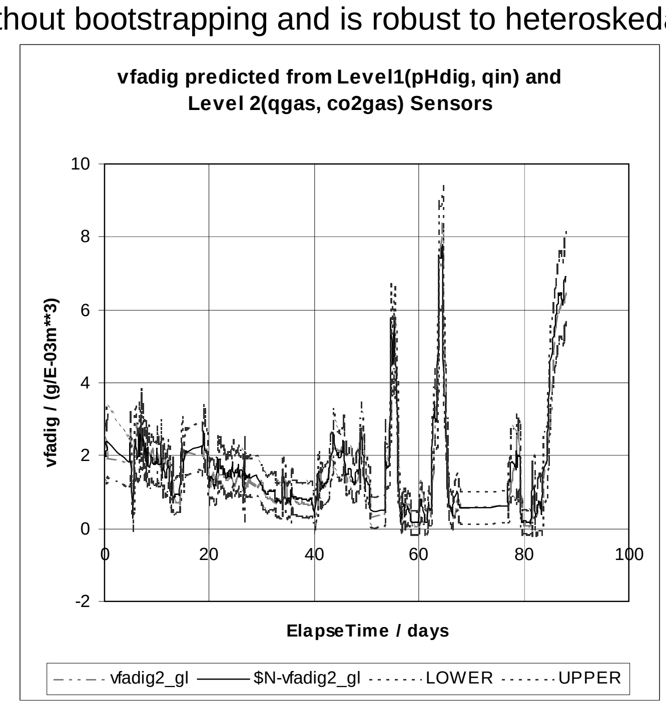
- A Neural Net with two output nodes can estimate the mean and variance of the conditional distribution of a target.
- One node is trained to fit a target value and the other is trained to fit squared residuals.
- This gives the prediction interval as:

$$PI(\mathrm{x_i}) \approx d^*(\mathrm{x_i}) \pm t_{(1-a/2),(n-k-1)} \sqrt{\frac{n\sigma^{*2}(\mathrm{x_i})}{(n-k-1)}}$$

where d^* is the estimated target for input row $\mathrm{x_i}$, t is the Student's t-distribution; n is the number of rows used in training, k is the number of applicable degrees of freedom, a is the significance level, $\sigma^{*2}(\mathrm{x_i})$ is the estimate of the variance of d for row $\mathrm{x_i}$.

Key findings:

- Good prediction of vfadig for independent test set from pHdig, qin, qgas, CO_2gas (% of CO_2 in gas). This contrasts with linear regression modelling.
- Has coefficient of determination,R^2 = 0.95 compared to linear regression with R^2 = 0.45.
- 96% of filtered test set experimental values lie in the 95% prediction band.
- Demonstrated a method which gives prediction intervals without bootstrapping and is robust to heteroskedasticity.



Neural Net estimates with prediction confidence intervals