The use of Data Mining for the Monitoring and Control of Anaerobic Waste Water Treatment Plants

Julian Gallop [1], Maurice Dixon [2], Jerome Healy [2], Simon Lambert [1], Laurent Lardon [3], Jean-Phillipe Steyer [3]

[1]Business and Information Technology Department,

CCLRC Rutherford Appleton Laboratory,

Chilton, Didcot, Oxon. OX11 0QX, UK

J.R.Gallop@rl.ac.uk

phone +44 1235 445569

fax +44 1235 445831

[2]London Metropolitan University, UK

[3]Laboratoire de Biotechnologie de l'Environnement - INRA, France

Key words: waste water treatment, data mining, linear regression, non-linear regression, neural nets, clustering, TELEMAC, sensors, decision support

Abstract. This paper describes the role of data mining in the EU-funded TELEMAC project. TELEMAC provides SMEs with tele-monitoring and communications support for the deployment of anaerobic digesters for the treatment of waste water from the alcoholic beverages industry. Anaerobic digesters are very efficient but can become unstable so require expert knowledge for their handling. TELEMAC aims to enable the expert to work remotely and expertise to be shared. The purpose of data mining in this context is to provide support for the fault detection and isolation system and guidance for the experts. In this paper we discuss the role of data mining and indicate how it will be deployed within TELEMAC. We report work on identifying digester process states, sensor depletion

or malfunction, and prediction intervals and prediction risk associated with neural net models. We report on using clustering and visualization to identify process states. Good estimates of chemical oxygen demand can be made using neural nets; the prediction risk and local error bars are seen to be qualitatively satisfactory.

1. Introduction

Anaerobic digesters have well-understood advantages in terms of efficiency in waste water treatment; they offer a high yield way of processing organic pollutants. They are rapid, can degrade concentrated and difficult substrates, offering the possibility of energy co-generation while producing very few sludges. They are biological systems which can become unstable, leading to a breakdown of the process resulting in a shutdown and a lengthy restart process. Avoidance of the instability needs expert supervision and leads to plants being run at much lower efficiency than could be achieved. (Bernard, 2004) provides a useful discussion and bibliography.

TELEMAC is an EU funded project[1] with 15 partners in 7 countries. The project involves treatment of waste water from the alcoholic beverages industry. The production plants are small to medium size units and are widely distributed. One aim is environmental impact reduction. Another aim is to provide a biogas quality suitable for cogeneration. Its objective is to provide a set of tools to assist a remote expert centre in managing multiple anaerobic waste water treatment plants (WWTPs) through the internet; the centre is known as a Telemonitoring and Advanced Telecontrol Centre, TCC. A discussion of the problems, issues, and plans was presented to the ECAI workshop on Binding Environmental Sciences and Artificial Intelligence. (Lardon et al, 2002).

Data mining is seen as a complementary technology within the telemonitoring and control centre. Its purpose is to analyse the expanding history of data from sensors and its correlation

with normal or abnormal conditions of the plant. The expected result is diagnostic help and a better understanding of the underlying mechanisms. Data mining supplements the simulation models from biomathematical equations. (Bernard et al 2001). There is a need for the TCC to re-evaluate the rules, classifications and predictions in the light of the experts' perception of digester behaviour as reflected by the accumulating data. A longer term aim is the generalization of expertise across plants.

This paper discusses the contribution that data mining can make to assisting in the deployment of anaerobic digesters. We have investigated data from several contributory anaerobic digesters within TELEMAC. The data from INRA-LBE pilot scale digester is best suited for the present investigation because of the quality, quantity, and expert understanding of the anaerobic digester and of the data. The data has been filtered by experts. We also had available simulation data from one INRA-LBE experiment. We present some findings from the deployment of data mining with data from the INRA-LBE digester.

Specifically we found:

1. Initial analysis of clustering results suggest that the cluster membership is stable as the user-directed number of clusters varies, which reduces the chances of arbitrary cluster determination. A further method of post-cluster analysis shows that certain sensor variables (including pH and volatile fatty acids in the digester) possess a dominatingly small variance in more than 50% of the clusters.

2. Linear regression gave a fair model for the prediction of the chemical oxygen demand in the digester but tests of the functional form showed that it was mis-specified. We present predictions of the chemical oxygen demand with local errors bars from non linear regression with neural nets There was some evidence that the prediction bounds were under-estimated by about 10% of the predicted value for the best neural net models.

[1] IST-2000-28156

2. Role of data mining in TELEMAC

2.1 TELEMAC as an application of Data Mining

The TELEMAC project includes several partners who are responsible for WWTPs being investigated in the project. The plants are diverse in their chemical principle of operation. Other differences result from the purpose of the WWTP which can be industrial, pilot or experimental. The active volume varies between 2 litres (experimental) and 5 million litres (industrial). Furthermore, the treatment plants operated by small to medium enterprises (SMEs), have few sensors since some are expensive, which contrasts with the experimental situation at INRA which has a full range. Some chemical measurements are taken manually at the industrial sites, resulting in sparser datasets. It is therefore of great interest to know which sensors may be most effective when balanced by the need for economy.

The Fault Detection and Isolation method, FDI, is crucial to the stable running of the digesters. The FDI runs at sites responsible for individual digesters but the calibration of process states is the responsibility of the TCC. Data mining is aimed at assisting this FDI system. Currently experts are banding manually sensor readings then classifying digester Process States; the procedure can be improved if data mining techniques continuously check first the validity of the banding and then check the validity of the expert rules. One objective of data mining is to seek a better classification of the digester states. Visualisation techniques comprise part of this approach. There is particular interest in using data mining models to estimate Process States and for situations where a sensor fails or is suspected of misreading, or absent because of cost.

2.2 Acquisition and availability of data

The data set used in this work is drawn from experiments conducted in the $1m^3$ pilot up-flow anaerobic fixed bed digester at INRA-LBE. (Steyer et al, 2002a). We have selected 90 days

of data measurements sampled at approximately 0.5 hour intervals for the work reported here; giving 4100 time stamped records for the study. This is an extensively instrumented digester; in this study we consider a subset of the sensors. The data from the sensors were filtered to ameliorate common sensor problems such as bias, drift and sticking and incomplete records were omitted from this set. There are missing data in the elapsed time ranges, {(0.6, 0.46), (15.6,18.4), (28,29), (67.7,76.8), (83.6,84.4)} days.

Sensor diagnosis is operated in two steps: single-variable analysis for the most obvious failures, multi-variable analysis for the others. A discussion of the approach has been given elsewhere (Lardon et al, 2004).

Table 1: Descriptive Statistics for Filtered Sensor Data

Sensor	Venn Level	Mean	Min	Max	Standard Deviation
tempdig / ^{o}C	1	35.2	27.3	37.5	1.07
qin / (1.0E-3 m^3/hour)	1	24.6	0	54	13.1
pHdig	1	7.2	5.25	8.76	0.4
co2Gas / %age	2	33.3	3	55	6.6
qgas / (1.0E-3m^3/hour)	2	203.2	0	481.	96.8
vfadig / (g/1.0E-03 m^3)	3	1.809	0	8.37	1.576
tocdig / (g/1.0E-03 m^3)	4	1.160	0	4.53	0.858
coddig / (g/1.0E-03 m^3)	4	3.9	0	17.2	3.05

The table (Table 1) shows basic statistics for the subset of 8 sensors and also shows the Venn diagram level, which is explained below. It is already known that tocdig bears a strong relationship to coddig, so for some of our work it is omitted.

The temperature of the digester, tempdig, the inflow rate, qin, and the pH of the digester, pHdig are derived from classical online instrumentation. The same is true of the biogas flow rate, qgas, and the %age composition of carbon dioxide in the gas, co2gas. The concentrations in the digester of volatile fatty acids, vfadig, total organic carbon, tocdig, and chemical oxygen demand, coddig require advanced and modified sensors (Steyer et al, 2002b).

A corresponding set of time derivatives was supplied with the data. The average hydraulic retention time was about 40 hours.

There are other extensive datasets originating from the same digester which can be used for evaluation. A more difficult task is the generalisation of results to other plants, which will become plausible when sufficient data becomes available from other types of plant.

2.3 Sensor ranking

Figure 1 shows the relative ranking of sensors as variables used here. The sensors have been ranked by digester experts into four levels shown in the Venn diagram based upon availability, expense, and reliability.

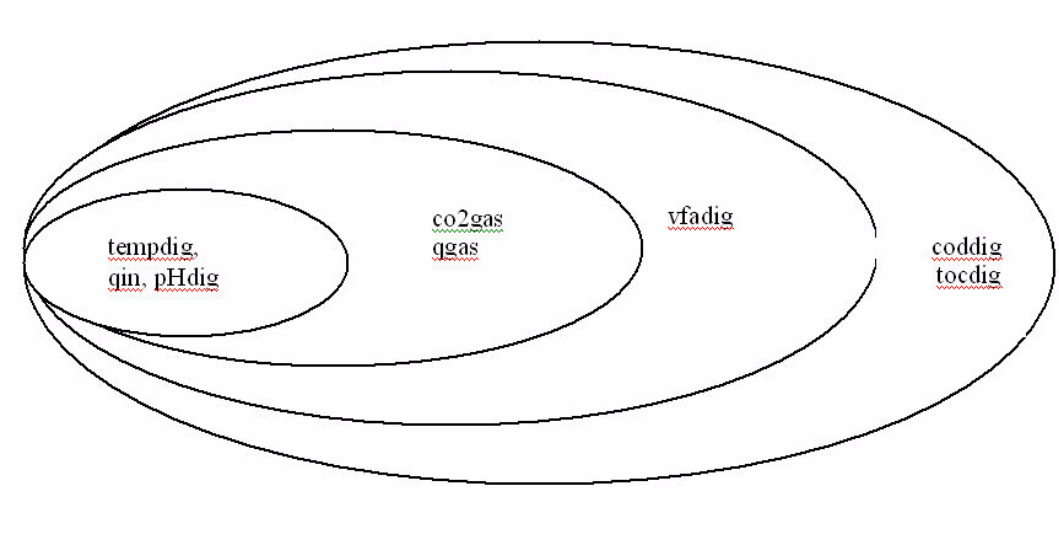

Figure 1: Venn diagram showing Sensor Ranking

2.4 Pairwise characteristics

Characteristics of the data are shown in Figure 2. This is a multiway scatterplot of 8 variables, using the XMDV multidimensional visualization tool (Ward, 1994); elapsed time and the subset of 7 sensors without tocdig are shown. The chart shows an individual scatterplot of each pair of variables. Each individual scatterplot is reciprocated by a reversed plot on the opposite side of the major diagonal, which also shows each variable plotted against itself. Thus the leftmost column shows plots of all the sensors against elapsed time and the plot of any sensor against another is independent of time. In this picture, some points are highlighted in black using XMDV's brushing tool and reveal projections of all data records for which codDig is above a certain threshold.

3. Modelling Approach to Regression

The first method to be described is Regression. For this and for the other methods described in this paper, the SPSS Clementine commercial, general purpose data mining package is used. For exporting data mining models to other platforms, either C code or the XML-based Predictive Model Markup Language (PMML, 2003) can be used.

The original dataset was split randomly into three approximately equal sized sets labelled TRAIN1, TRAIN2, TESTSET. Model training was done on TRAIN1 (see below for use of TRAIN2) and testing was on the TESTSET. Time was not used as a modelling variable in this work but it was retained for record labelling and visualisation.

Exploration of the data was carried out using linear regression to provide an indicative ranking of the relative importance of each sensor's data in the set. The ranking was done systematically by forward stepwise regression using F^* = (mean squared residual/mean squared error). The E-views statistical package was used for some of the diagnostics for heteroskedasticity, suitability for a linear functional form, normality of residuals, and serial correlation.

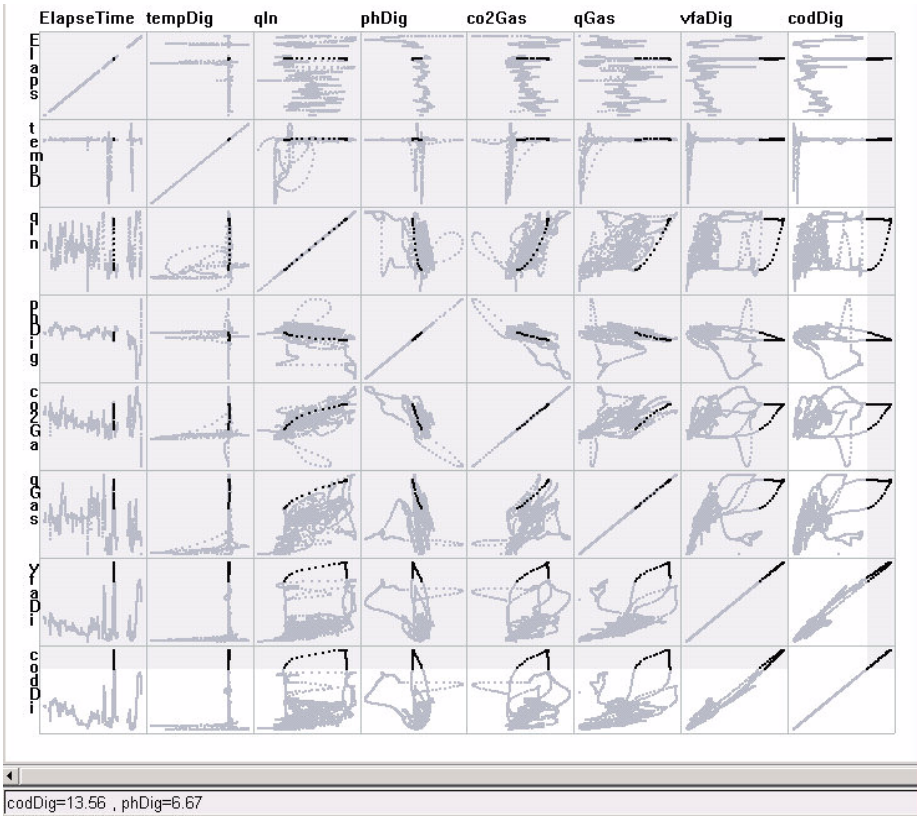

Figure 2: Multi-variate scatterplot of 7 sensors using the XMDV visualization tool

Initial neural net models were constructed using 1 hidden layer of 5 logistic functions; monitoring showed that a fixed time exit on the best model achieved gave a satisfactory

convergence in the target variable. In the modelling, TRAIN1 was randomly split into a training set and a validation set with 50% of records in each; this is called Phase1 training. The Phase1 neural net model was then used to predict a set of squared residuals for each record in dataset TRAIN2. In Phase2 another neural net is fitted to the input data using a least squares cost function for a joint target of the target variable and the squared residual. Again 1 hidden layer is used but with 7 logistic functions in it. This approach enables us to make a joint estimate of a variable and its prediction intervals for unknown target values given unseen data. The theory for this is given in (Healy et al, 2003a) and a demonstration of its use was given in (Healy et al, 2003b); it is appropriate where the residuals are heteroskedastic. The method is robust, does not require bootstrapping or calculation of the Hessian matrix. Nor does it require the assumption of a Gaussian noise model.

$$P \approx d^*(\mathrm{x}_\mathrm{i}) \pm t_{(1-\alpha/2),(n-k-1)} \sqrt{\frac{n\sigma^{*2}(\mathrm{x}_\mathrm{i})}{(n-k-1)}}$$

t is the standard Student-t function, n is the number of data points in the training set, k is the number of degrees of freedom(variable parameters for the neural net), x_i is the input record vector (omitting timestamp values). $d^*(x_i)$ is the estimated value and $\sigma^{*2}(x_i)$ is the corresponding estimated point variance.

3.1 Modelling with linear regression

In tables 2, 3 and 4 are presented a series of linear regression models trained on data set TRAIN1. The aim was to estimate the sensor ranking levels (as shown in Table 1) of sensors required to predict coddig. Using R^2 mean squared errors, Akaike's information criterion (AIC) and residual plots, we conclude that the linear models Level4+3+2+1, Level3+2+1 give fair estimates of coddig. On the basis of these it is reasonable to infer that the prediction of coddig requires either a Level4, tocdig, or a Level3, vfadig, sensor. Level1 and Level2

sensor data are not themselves able to predict coddig within the linear approach. Elapse time is not included as an explicit modelling variable and tempdig was eliminated as being indistinguishable from 0. The Ramsey RESET test indicates that the linear model has a functional form which is mis-specified for the prediction of coddig; the Jarque-Bera statistic indicates the residuals are not normal; the White test indicates heteroskedasticity; the Breusch-Godfery test indicates there is serial correlation of the residuals. Pearson correlations were used to identify high multicolinearity between tocdig, coddig, and vfadig. For this dataset, the correlation between tocdig, and coddig was 0.97, between vfadig and coddig was 0.95, and between tocdig and vfadig was 0.988.

Table 2: coddig prediction by forward stepping linear regression from Venn Level4+3+2+1

Addition order->	**tocdig**	**Qin**	**pHdig**	**co2gas**	**qgas**	**vfadig**		
R^2	0.940	0.948	0.952	0.961	0.977	0.979		
-AIC	522.1	667.8	750.5	946.7	1508.	1626.		
Addition order->	**tocdig**	**Qin**	**pHdig**	**co2gas**	**qgas**	**vfadig**	**dco2gas**	**dqgas**
R^2	0.940	0.948	0.952	0.961	0.977	0.979	0.980	0.980
-AIC	522.1	667.8	750.5	946.7	1508.	1626.	1632.	1635.

For Level4+3+2+1 the order of selection of sensor data for addition remained the same when the set of derivatives were included in the data set. vfadig is highly correlated with tocdig so is included late in the selection. The adjusted R^2 values are effectively identical to R^2. The ratio of Akaike's to Schwarz-Bayes Information Criterion remained close to 1.025 . Including the time derivatives in the inputs did not significantly modify these qualitative expectations.

Table 3: coddig prediction by forward stepping linear regression from Venn Level3+2+1

Addition order->	**vfadig**	**pHdig**	**Qgas**	**co2gas**	**qin**				
R^2	0.912	0.933	0.944	0.972	0.972				
-AIC	113.3	399.8	581.5	1293.	1302.				
Addition order->	**vfadig**	**pHdig**	**Qgas**	**co2gas**	**dvfa dig**	**tempdi g**	**dco2gas**	**dQgas**	**dpH dig**
R^2	0.912	0.933	0.944	0.972	0.972	0.972	0.972	0.973	0.973
-AIC	113.3	399.8	581.5	1293.	1320.	1325.	1329.	1335.	1337.

In Figure 3 we compare the experimentally determined chemical oxygen demand, coddig, given in TESTSET with the predictions of the models. Missing time ranges are linearly joined in the graph. Level4+3+2+1 refers to predictions from the model in Table 2 without the derivatives. Resid_Level4+3+2+1 is the corresponding predicted residual. Also in Figure 3 is plotted the residual from the linear regression model in Table 3, Resid_Level3+2+1 which omits the time derivative terms.

Table 4: coddig prediction by forward stepping linear regression from Venn Level2+1

Addition order	**co2gas**	**pHdig**	**Qgas**			
R^2	0.257	0.271	0.288			
-AIC	-2115.	-2097.	-2074.			
Addition order	**co2gas**	**dqgas**	**Qin**	**pHdig**	**dpH dig**	**qgas**
R^2	0.257	0.282	0.297	0.308	0.314	0.317
-AIC	-2115.	-2082.	-2061.	-2048.	-2040.	-2037.

3.2 Non Linear Regression

A set of neural net models have been trained using the data set TRAIN1 as used above; however all the input variables from Table 1 at each Venn level were used. The target was a least squares cost function for coddig. The hidden layer nodes were logistic functions. Initial models were produced using the software's heuristic algorithm for number of nodes and training exit. For the results reported here, a fixed number, five, logistic nodes were used in one hidden layer for Phase1. TRAIN1 was randomly split into a training and a validation set containing an equal number of records; pruning of input variables and the number of nodes was suppressed. Training was terminated on time with the best model chosen on the criterion of best fit to the validation set. A set of squared residuals for dataset TEST2 were generated using the model from Phase1 with the residual being the difference between the input and the predicted value of coddig. This set of residuals constituted the local error bar model which we aimed at constructing. For Phase2 the target was a least squares cost function for coddig and

its squared residual. Seven hidden layer nodes were used. Pruning was suppressed, the best network was selected on the basis of the performance on the validation set constructed from TEST2 with time used to determine the exit.

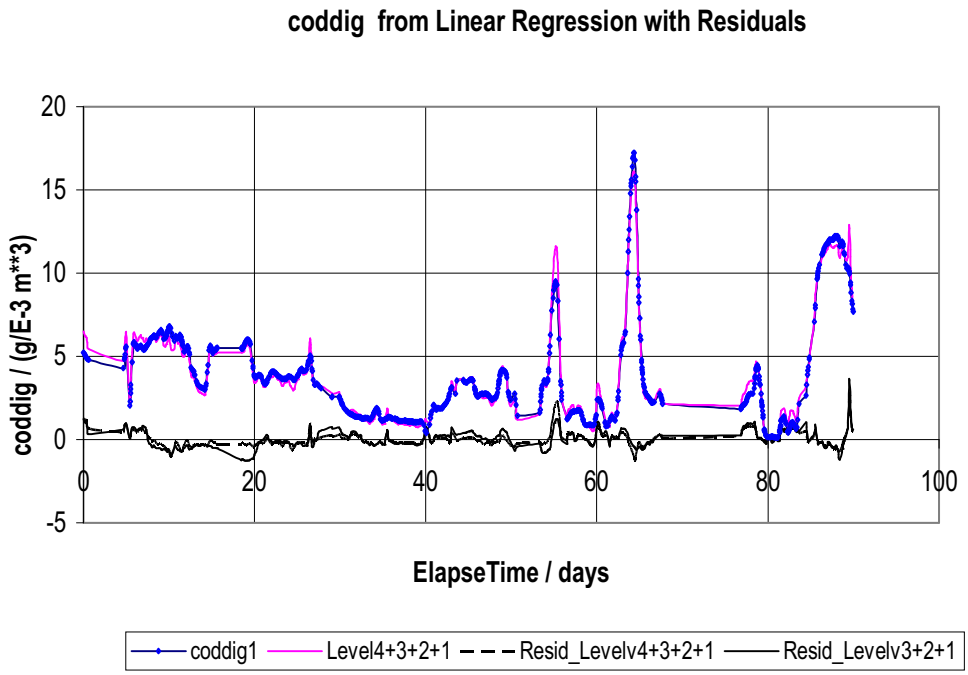

Figure 3: Prediction of coddig by linear regression

Figure 4 below is a plot of the predictions by non-linear regression from the neural net regression model of coddig from Phase2 as VLevel4+3+2+1 with the corresponding filtered experimental points as coddig. (In the figure, there is a suffix 1 which labels the sensor used) The corresponding residual is plotted as Resid_VLevel4+3+2+1; this is the actual difference between coddig and VLevel4+3+2+1. The upper and lower prediction interval bounds are plotted as Upper and Lower; these are based upon the neural nets prediction of the squared residual. The prediction interval has been estimated using a 95% confidence level for the t-value. In practice we found the bands needed to be (5%, 10%) of coddig wider in order to encompass (88%, 94%) of the filtered experimental points. There are linear joins of missing data points.

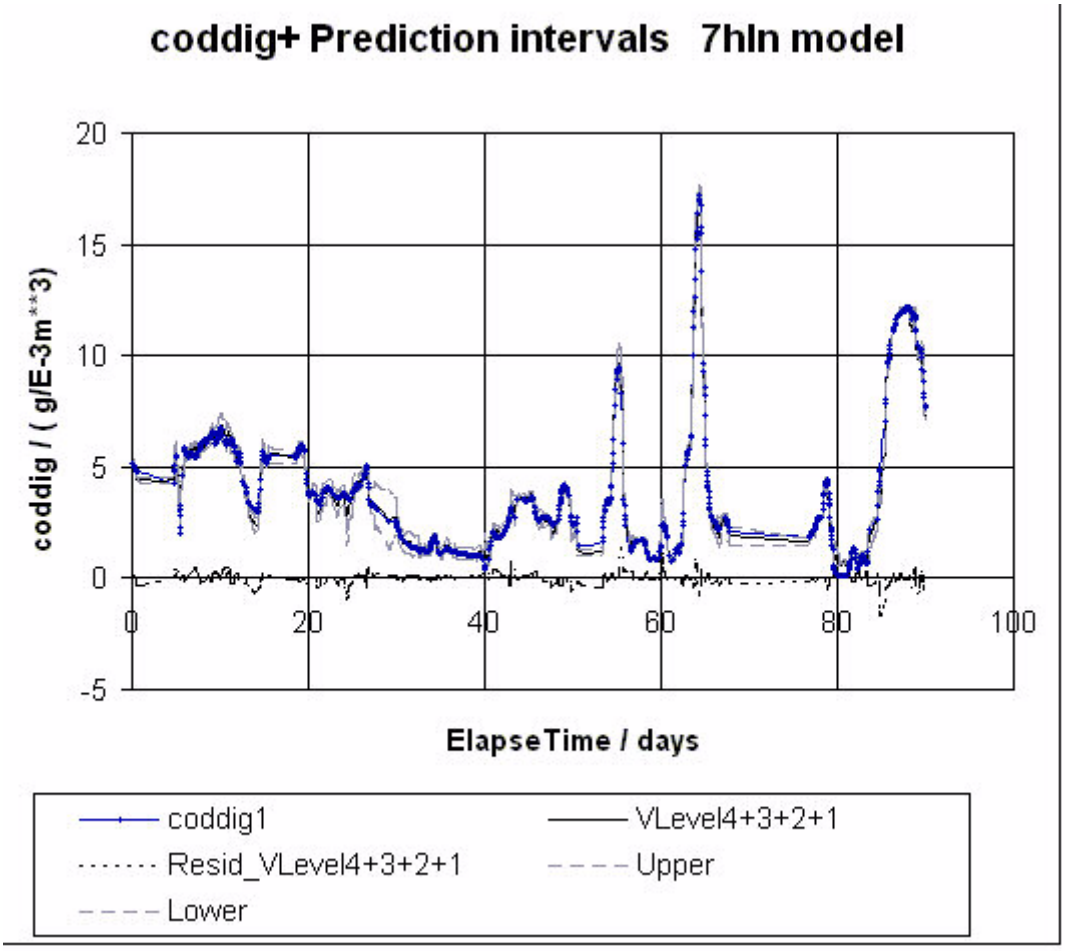

Figure 4: Predictions of coddig by non-linear regression with neural nets

Table 5: Summary of Phase2 Neural Net Regression

	VLevel4-3-2-1	VLevel3-2-1	VLevel2-1
R^2	0.993	0.981	0.921
E(MSE)	0.067	0.188	0.791
PRerror	-0.04	-0.12	+0.10
t-paired	0.246	1.352	2.911
t-dif-variance	0.0144	0.132	0.599

Table 5 shows that there is a substantial increase in the mean squared error E(MSE) for known target values when tocdig and vfadig are omitted from the modeling variables. The model VLevel2-1 fails to satisfy the paired t-test. Although the R^2 for VLevel2-1 is high there is evidence from the residuals that the model is not satisfactory. Firstly it is substantially more noisy than the other neural net models over the whole range of coddig so the prediction intervals are much wider. Also it fails to deal with the higher coddig values, viz those >13g/1.0E-03m^3.

Sensitivity analysis gives the relative importance of input variables for predicting the joint target based upon the partial gradient of the target function with respect to the input variable. In procedures where pruning is permitted then the sensitivity analysis is used to determine input variable deletion. The figures in Table 6 (below) suggest that in the absence of tocdig and vfadig then the model makes much more use of qgas and qin. The sensor tempdig is found to contribute to the model. The sensitivities from Phase1 which estimates just coddig show some significant differences.

3.3 Summary of Regression

The linear regression models form a useful starting point for the modelling of coddig for this dataset. Both approaches satisfy t-tests for the consistency of means with the filtered experimental test data. However the Ramsey RESET test indicates that the linear model has the wrong functional form. The neural net models for the same inputs have better R^2 values and smaller residuals as evidenced by E(MSE). However when both tocdig and vfadig are absent from the inputs then the neural net appears to be able to achieve a better prediction than the linear model, although it fails the paired t-test for estimating the mean and is poor at high coddig values. The inclusion of the gradients does not appear to produce significant improvements in the modelling.

We have found it would be necessary to make a small semi-empirical adjustment to the predicted error bars, Upper and Lower, in order to encompass 95% of the coddig data points within them for the better neural net models. This appears to be caused by the neural net underestimating its prediction of small squared residuals. Further work is required on this. Moody (1994) has defined Prediction Risk as the expected value of mean squared error for unseen targets. The output from Phase 2 allows us to estimate this but the same features of squared residual error estimation disturb the value; these errors are recorded as PRerror. Over many modelling runs we have found that our estimate of Prediction Risk shadows the actual testset E(MSE) and so is a useful guide to the Prediction Risk.

Table 6: Sensitivity Analysis of Phase 2 Neural Net Models

Sensor	Venn Level	VLevel4-3-2-1	VLevel3-2-1	VLevel2-1
tempdig / ^{o}C	1	0.23	0.20	0.26
qIn / (1.0E-3 m^3/hour)	1	0.08	0.06	0.38
pHdig	1	0.41	0.32	0.43
Co2Gas / %age	2	0.46	0.13	0.25
qgas / (1.0E-3m^3/hour)	2	0.26	0.21	0.87
vfadig(g/1.0E-03 m^3)	3	0.85	0.62	---------
tocdig / (g/1.0E-03 m^3)	4	0.58	---------	---------
coddig / (g/1.0E-03 m^3)	4	xxxxxx	xxxxxx	xxxxxx

4. Clustering

The aim of using a clustering algorithm is to divide the data into partitions (clusters), where the data within each partition possesses some coherence. In general, cluster centres are identified and a data record is assigned to a cluster possessing the nearest centre.

4.1 The general approach

In TELEMAC, the goal is to identify distinct states of a WWTP digester, in order to diagnose and identify faults using the FDI, which is a straightforward, modular scheme. As already discussed, a partitioning of process states already exists in the current version of the TELEMAC FDI and it is desirable to find more effective partitionings if they exist.

4.2 Procedure and results

There is no single clustering algorithm, which will produce the best partitioning for all circumstances. The approach here is to use one available on the market (Two Step algorithm from the Clementine data mining system (SPSS, 2003)) and apply postprocessing techniques to provide insight into the results, which may require interpretation by the process experts and further iterations.

The information about the cluster data mining model generated by the algorithm includes the centre of each cluster and is output as PMML (DMG, 2003). This is transformed by XSLT, using a XML stylesheet, to a form suitable for combining with the cluster membership tagged to each record of the original dataset.

This algorithm allows the user to specify the numbers of clusters to be found or to leave that unspecified. When left unspecified for this data, this algorithm produced 2 clusters. Since currently the FDI uses 5, it was judged that 2 clusters provides insufficient discrimination of

process states. Therefore the cluster algorithms were run again, specifying 5 clusters and 15 clusters, the former corresponding to the number currently used in the FDI.

Given a set of records, each containing a value for each of variables v_1, …. v_j, a clustering algorithm generates an identifier u for the cluster, which can be and is here a number from 1 up to and including the number of clusters being generated on this execution of the algorithm. Since the algorithm is being executed repeatedly with a different number of clusters produced on each execution, the result is a new dataset which, for each original datum, contains: v_1, …. v_j, u_1,,…u_k .

Figure 5 below shows a multivariate scatterplot, but this time with the cluster identifiers for a 5-cluster and a 15-cluster run included at the end of the other variables. The discrete property of the identifiers is apparent from the last 2 columns and bottom rows.

Since the clustering algorithm can produce results whichever number of clusters is specified, it is desirable that there is some stability, without which there would be some suspicion of the results. A plausible measure of this is, for example, to determine the degree to which each of the 15 clusters is contained wholly, or at least largely, within one of the 5 clusters. More precisely given one of the 15 clusters, with which of the 5 clusters does it share the greatest overlap and what is that proportion expressed as a percentage, which we refer to as the inclusion of the 5-cluster identifier with the 15-cluster set. An inclusion measure of 100% would mean that the 15-cluster identifier is wholly contained within one of the 5-clusters Having defined this, its calculation involves a straightforward use of an SQL database.

For the 15 clusters, the minimum measure is 76% which could be regarded as borderline and requiring further investigation and, for all the other clusters, the measure is 88% or above which is satisfactory.

For greater confidence in the stability of the results, further cluster executions with different numbers of clusters need to be made.

A further investigation is to determine how compact the clusters are. If some variables have a narrow variance within a particular cluster, it is then possible not only to begin to determine the process states, but also to determine which variables and what ranges of those variables are responsible for a cluster algorithm identifying that cluster.

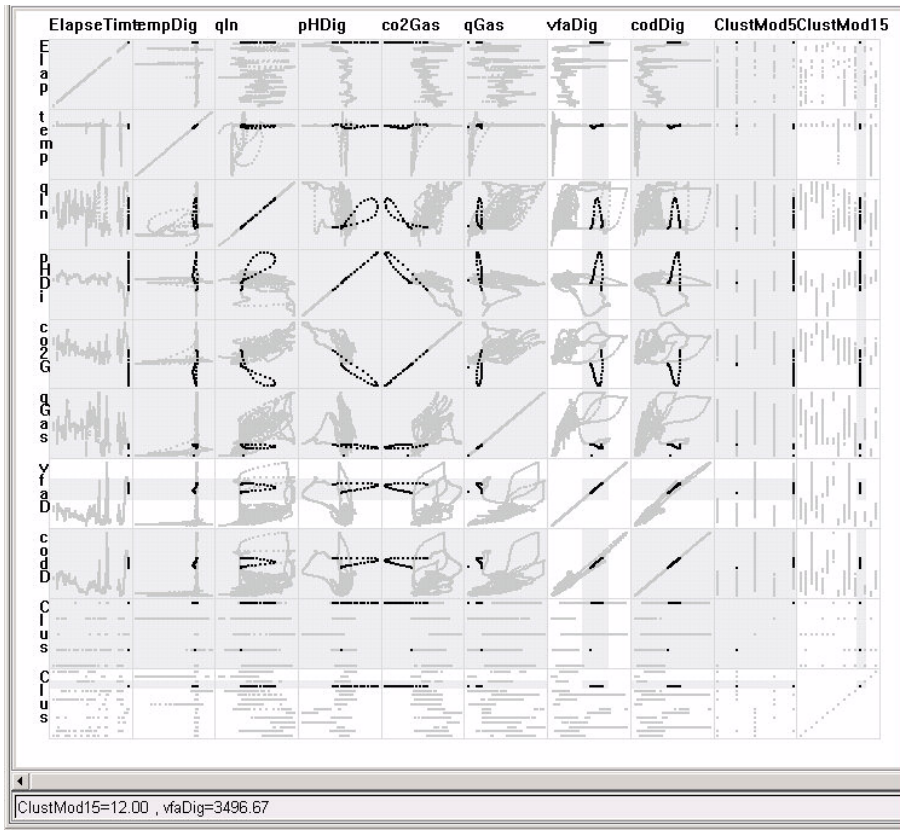

Figure 5: Multiway scatterplot including 2 sets of cluster identifiers

.

For a given variable within a given cluster, we use a per-cluster width measure and express that as a percentage. This is based on the standard deviation of that variable within that cluster, by comparison with standard deviation of that variable within the whole dataset.

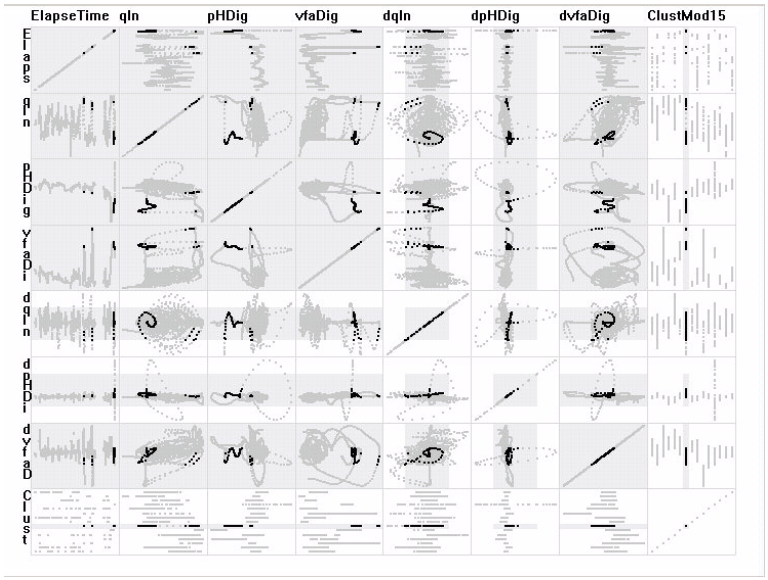

Figure 6: Multiway scatterplot of selected variables, their derivatives and cluster identifiers

In a 15-cluster run, we note that all but 3 clusters contain at least one variable or derivative whose per-cluster width measure within the cluster is less than 40%. 7 out of 15 have at least one variable or derivative whose per-cluster width measure is less than 25%. Potentially these clusters and variables within them are candidates offer potential for identifying criteria for process states.

In Figure 6, one of these clusters is shown, but restricting the variables to those whose per-cluster width measure is low.

4.3 Remarks

Further investigation of these relationships would strengthen our knowledge of the underlying process. Further clustering algorithm execution needs to be run with other user-directed numbers of clusters.

It is also necessary to complement the clustering with other methods. Having determined sets of clusters, applying an associative rule induction algorithm would infer some criteria for membership of that cluster.

One source of complexity in this application is the presence of more variables than can be displayed at once, except by sophisticated 2D charts. Some work described above helps the number of relevant variables be reduced. Another way to reduce the dimensional problem is to use Principal Components Analysis. This is another technique that will be used to strengthen our knowledge of the process state.

5. Conclusions

This work has demonstrated that accurate predictions can be made of digester's chemical oxygen demand from sensor datasets which include either a Level4, tocdig, or a Level3, vfadig sensor variables. The local error bars corresponding to 95% prediction intervals for unseen data were estimated but comparison with the actual TESTSET target indicates that there can be an underestimate of the width of the prediction interval. The method used for obtaining prediction risk and prediction intervals was robust to heteroskedasticity and did not require bootstrapping.

Some cluster post-processing techniques have been used in combination, which increases their effectiveness. These include the use of appropriate measures to assess the stability and relevance of the clusters. The measures are straightforward to calculate and require the combination of information from a number of sources and so can be done with an SQL DBMS. These, with appropriate visualization, show that subject to further applications of the algorithm specifying larger numbers of clusters, there is stability in the results. They also show that certain sensor variables, particularly pH and VFA in the digester, possess a small variance in more than 50% of the clusters.

This initial analysis of the clustering results suggest that the cluster membership is stable as the user-directed number of clusters varies, which reduces the chances of arbitrary cluster determination. A further method of post-cluster analysis shows that certain sensor variables

(including pH and volatile fatty acids in the digester) possess a dominatingly small variance in more than 50% of the clusters.

Our understanding of data was greatly enhanced through the use of the multivariate visualisation tool XMDV.

Further work will be undertaken on sensor depletion and on forward time prediction.

6. Acknowledgements

We wish to acknowledge the support from the European Commission's IST programme under the TELEMAC project (IST-2000-28156).

J V Healy was supported by a teaching research bursary from the London MU CCTM.

7. References

Bernard, O., et al. TELEMAC: *An integrated system to remote monitor and control anaerobic wastewater treatment plants through the Internet [TELEMAC contribution #1]*, to be published in the conference proceedings of AD10, 2004

Bernard, O., Hadj-Sadok, Z., Dochain D., Genovesi A., Steyer J.P. Dynamical model development and parameter identification for anaerobic wastewater treatment process. *Biotechnology and Bioengineering,* 75(4), 424-439, 2001.

DMG (2003), web pages on PMML at http://www.dmg.org/v2-0/GeneralStructure.html

Healy, J. V., Dixon, M., Read, B. J., and Cai, F. F. Confidence in Data Mining Model Predictions from Option Prices. *Proceedings of the IEEE, IECON03*, 1926-1931, 2003a

Healy, J. V., Dixon, M., Read, B. J., and Cai, F. F. Confidence and prediction in generalised

non linear models: an application to option pricing, *International Capital Markets Discussion Paper* 03-6, 1-42, 2003b.

Lardon, L., Punal A., Steyer, J.P., Roca, E., Lema, J., Lambert S., Ratini, P., Frattesi, S., and Bernard, O. Specifications of Modular Internet-Based Remote Supervision Systems for Wastewater Treatment Plants *Proceedings of BESAI* 45-50, 2002.

Lardon, L., Punal, A. and Steyer,J.P.,On-line diagnosis and uncertainty management using evidence theory—experimental illustration to anaerobic digestion processes, *Journal of Process Control*, Volume 14, Issue 7, October 2004, Pages 747-763, 2004

Moody, J.E., Prediction Risk and Architecture Selection for Neural Networks : *From Statistics to Neural Networks:Theory and Pattern Recognition Applications* (Eds: Cherkassky, V., Friedman, J.H., and Wechsler, H.) NATO ASI Series F, Springer Verlag, 1994.

Steyer, J.P, Bouvier, J.C., Conte, T., Gras, P., and Sousbie, P. Evaluation of a four year experience with a fully instrumented anaerobic digestion process. *Water Science and Technology*, 45(4-5), 495-502, 2002a.

Steyer, J.P., Bouvier, J.C., Conte, T., Gras, P., Harmand, J. , Delgenes, J.P. On-line measurements of COD, TOC, VFA, total and partial alkalinity in anaerobic digestion processes using infra-red spectrometry. *Water Science and Technology*, 45(10), 133-138, 2002b.

Verstraete, W., Vandevivere, P. New and broader applications of anaerobic digester. *Critical Reviews in Environmental Science and Technology*, 29(2), 151.

Ward, M.O. XmdvTool: integrating multiple methods for visualizing multivariate data,

Proceedings IEEE Visualization 1994, 326-333, 1994