DISTRIBUTED DATA MINING AND KNOWLEDGE MANAGEMENT WITH NETWORKS OF SENSOR ARRAYS

Maurice Dixon

Computing, Communications Technology and Mathematics,
London Metropolitan University, 31 Jewry Street, London, EC3N 2EY, UK

M.Dixon@Londonmet.ac.uk

Simon C. Lambert and Julian R. Gallop

CCLRC Rutherford Appleton Laboratory,
Chilton, Didcot, Oxon, OX11 0QX, UK

S.C.Lambert@rl.ac.uk

J.R.Gallop@rl.ac.uk

Abstract Environmental pollution control relies heavily on human expert judgment supported by historical data and scientific models. Telemonitoring by networks of arrays of heterogeneous sensors provides the opportunity for data mining models to be constructed from the historical data to supplement human expertise. This paper reports some progress made in the TELEMAC project by data mining. TELEMAC is concerned with enhancing the efficacy of anaerobic digestion in potentially unstable digesters. In the laboratory using full instrumentation it is possible to derive a good description of the digester state. With data mining it is possible to identify some constraints on sensor choice. This paper examines this data mining work from the perspective of a three layer Grid architecture to see what implications and requirements arise that could benefit the exercise of expert judgment. After placing the specific TELEMAC situation in a generic Grids context, we present a classified approach to attributes for metadata and indicate some examples of model resource discovery.

Keywords : Anaerobic digestion, Data mining, Grids, Telemonitoring and control, Wastewater treatment.

1 Introduction

Networks of sensor arrays, measuring properties of multiple instances of some physical process, raise some important issues in the context of Grids. An example is provided by the TELEMAC project [1], a European Union funded project on anaerobic wastewater treatment, in which individual treatment plants are equipped with a variety of sensors. The aim of TELEMAC is to improve the monitoring and control of digesters from a central telemonitoring and control centre, TCC [2]. The control of these plants could benefit from data mining and the leveraging of knowledge through the TCC.

Although the TELEMAC project was not conceived as a Grids project, nonetheless there is clear potential for applying Grid technologies. The focus is on three levels of grids: knowledge, information, and data rather than computation. Issues that arise include:

1. data heterogeneity;
2. the data mining methods themselves;
3. time-based issues, such as the updating of data mining models;
4. the role of human expertise.

In TELEMAC the user interacts with a heterogeneous environment of data stores and data collection sensors. Grid technology could provide a standard framework for the interoperation of the distributed sites. Jeffery emphasised metadata, agents, and brokers as key architectural components of Grids [3]. Paraphrased here are observations relevant to TELEMAC on:

"**Metadata:** Most examples of metadata in use today are neither structured formally nor specified formally so tend to be of limited use for automated inter-operations and consequently require human interpretation."

"Agents/Brokers: Agents use metadata to take action; the can provide a monitoring function. Brokers act as go-betweens for agents"

This paper is structured as follows. Section 2 considers the industrial context and associated biochemical processes. Data mining in TELEMAC is discussed in Section 3; the issues of sensor arrays, the role of sensor ranking, and diversity of sensors are addressed in a data mining context. Examples of data mining results are presented. In Section 4 we consider TELEMAC from the perspective of the three layer Knowledge, Information, and Computation/Data Grids. It is here we address the issue of leveraging knowledge and grid resources. The role of human expertise in providing knowledge management in the plant monitoring and control cycle is presented and this shows the way the three Grids interact in this type of environment. In Section 5 we identify some specific attributes that are useful in the metadata for our data mining models and resources. A short summary of our conclusions finishes the section. Some issues of heterogeneity are considered in the Appendix.

2 Industrial context

Anaerobic wastewater treatment is an important technology for the disposal of certain kinds of waste, in particular the by-products from alcohol production in wineries and distilleries [4]. It has great advantages such as efficiency, low production of sludge, and the possibility of energy recovery through cogeneration. However it is an unstable process which is difficult to monitor and control with the consequence that plant is operated at low efficiency. Expert knowledge is required for efficient operation of the plant but that expertise is unlikely to be locally available at small, possibly remotely located, individual plants. Therefore the role of the TCC is crucial here in supporting expert human knowledge by a range of analysis and prediction techniques.

The anaerobic digester plants operate on a range of engineering principles such as upflow sludge blankets, lagoons, upflow fixed-beds and continuous stirred tanks, CSTRs. Within TELEMAC there is a preponderance of CSTRs at the industrial level with typical volumes of 500->5000m^3. The chemical oxygen demand, COD, of the wastewater is one measure of the outflow quality; organic loading rates within the digesters vary between 2kg and 20kg COD m^{-3}d^{-1}. Measurement of COD is generally not available on-line [2].

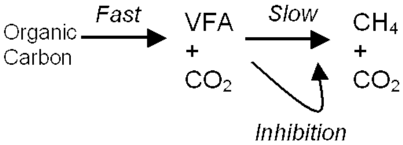

Figure 1. The biological process for anaerobic waste water treatment

The biological process has two main steps; these are shown in Figure 1. In the first step a set of acidogenic bacteria generate volatile fatty acids and carbon dioxide. This conversion proceeds at a fast rate. Volatile fatty acids themselves are acetates and acetic acid or similar. The second step is a slow conversion of the volatile fatty acids to methane and more carbon dioxide by methanogenic bacteria. The problem is that a build up in the concentration of volatile fatty acids inhibits the methanogenic bacteria. This can lead to suppression of the second stage and ultimately to irreversible destabilisation of the digester; then it could take a period of several weeks or even several months to recover. A converse problem occurs if the digester is hydraulically over-loaded and the biomass is washed out.

3 Data mining in TELEMAC

3.1 Introduction

The biological and chemical processes involved in anaerobic digestion are complex but there is good qualitative understanding of the main features. Although analytical models have been developed [5], there is still much scope for data mining of sensor data to complement them. Data mining helps to answer both static and dynamic questions, such as which sensors form the minimum set required for accurate estimation of key variables like concentration of volatile fatty acids, or what is the likely future value of such a variable given the current state of the digester plant [6].

3.2 Sensor ranking and diversity

A wide range of sensors are commercially available for use with anaerobic digesters. These are summarized below.

Sensor Types:

Classical plant instruments such as gas and liquid flow meters, pressure and temperature gauges.
Titrimeter: to measure acid and base concentrations (up to 4 variables)
Infra-Red spectrometer (up to 5 variables)
TOCmeter to measure total organic content
Thermal conductivity sensor for CO_2

Sensor Mode:

Online sensors return a value at measurement time
Offline chemical analysis returns a measurement significantly later and may be different in value than from an online measurement

Sensor Problems:

Sensor reliability – failure due lack of precision, saturation, lag in recovery of measuring capacity, foaming in digester
Sensor accuracy – calibration; contamination; standard setting

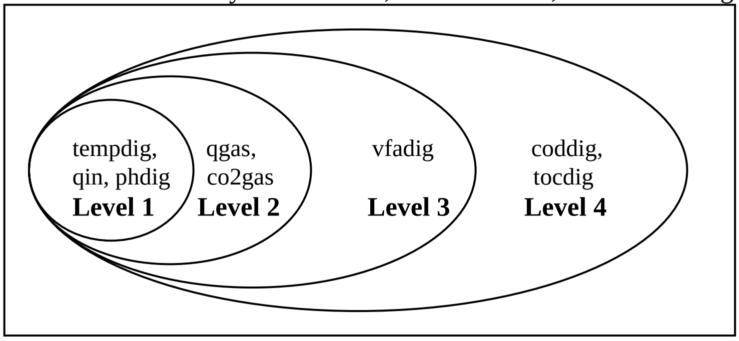

Figure 2. Venn diagram showing sensor ranking. The sensors are defined with their Level. Suffix *dig* indicates measurement of the digester content.

In Figure 2 *tempdig* is the temperature, *qin* is the influent liquid flow rate, *phdig* is the pH, *qgas* is the biogas flow rate, *co2gas* is the percentage of carbon dioxide in the biogas, *vfadig* is the concentration of volatile fatty acids, *tocdig* is the concentration of total organic content, and *coddig* is the concentration of chemical oxygen demand. With a full set of sensors it is possible to get a fairly complete chemical description of the current digester state. The figure shows expert judgement of the ranking of sensors by expected availability/reliability, with the simplest and most robust in the inner ring. These four levels of sensor are relevant when dealing with operational industrial systems which would lack such full instrumentation.

3.3 Data mining techniques used

Classification and **Sequencing** A key aspect of data mining is the classification of digester states using cluster analysis. Analysis of the clustering results suggested that the cluster membership is stable as the number of clusters varies. A subset of variables in each cluster had a narrow spread in that cluster.[7] It has been possible to characterize state sequences and transition frequencies.

Regression models have been used for several purposes:

1. Models for predicting data values for missing/faulting sensors were constructed with associated predictions of confidence intervals. This has allowed both current prediction and short term forecasting of the concentration of dissolved and suspended organics during sensor failure.
2. Highly accurate short term forecasting is feasible using multivariate autoregression; with reliable sensors this could be used for plant control.
3. Predictions from auto-regression are of little use over extended time on occasions of sensor(s) faulting, a frequent occurrence, because the models depend on known target values at previous times. Non linear multivariate regression performs satisfactorily for current and imminent states.

The models need to be evaluated against an independent test set of data to ensure that the model training does not result in over-fitting to errors in the training data. Statistical tests for quality of fit need applying. such residuals, mean squared and mean errors, squared Pearson correlation function (R^2), and paired sample t-tests for means.

A range of models need to be deployed. Linear models can provide good starting pointers. In some circumstances they can be sufficient in themselves eg in the most extensively instrumented digesters. In other cases artificial neural net models provide a markedly superior model judged by out-of-sample test set estimates. Unit root tests aid a decision on whether to model in differences or levels.

3.4 Examples of work done

Data mining has shown that

- it is feasible to determine the ranking of sensors; for example in order to estimate a Level4 variable (Levels as in Figure 2) it is considerably better to have at least one Level3 sensor dataset (*coddig* requires either Level3 *vfadig* or Level4 *tocdig*)
- features between variables can mean that a second sensor adds little to the improvement of a model. E.g. strong colinearity means that if *tocdig* data is available then *vfadig* adds little additional modelling power

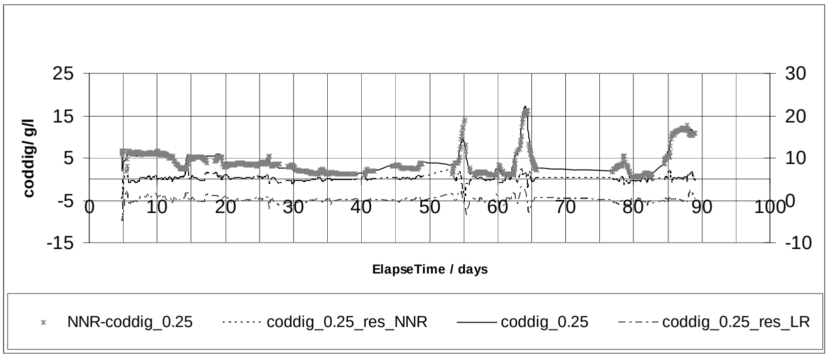

Figure 3a. Forward prediction of 0.25days. Prefix or suffix of NNR or LR indicates neural net or linear regression respectively.

Figure 3a shows a forward prediction of 0.25d for an INRA validated dataset using the sensor variables from Level1, Level2, and Level3 inputs to predict a Level4 variable, the concentration of *coddig* in g/litre. It compares the independent test set experimental data with the prediction of a neural net model and shows residuals on the left hand scale. The model had 8 logistic functions in two hidden layers (as 5+3) with *tempdig* eliminated; R^2=0.945, t-pair=1.3, mean residual error = 0.031, predicted mean square error=0.332. The residuals for a corresponding linear regression are shown on the right hand side scale. R^2=0.930(in sample R^2=0.928), t-pair=0.21, mean residual error = 0.014. Figure 3b compares the independent test set experimental data with the prediction of a neural net model. 97% of the actual experimental data points fall within the 95% prediction confidence band.

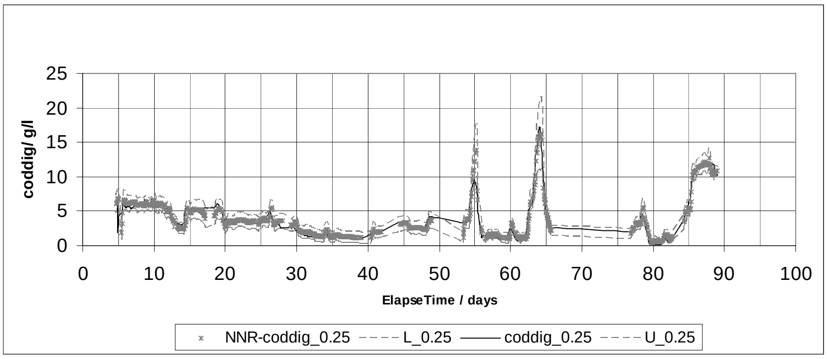

Figure 3b. Lower(L_0.25) and Upper(U_0.25) 95% prediction confidence bands [6] from forward prediction of 0.25days for a validated dataset for the concentration of *coddig* in g/litre.

4 The Grids context

4.1 Telemonitoring and control:the TELEMAC concept

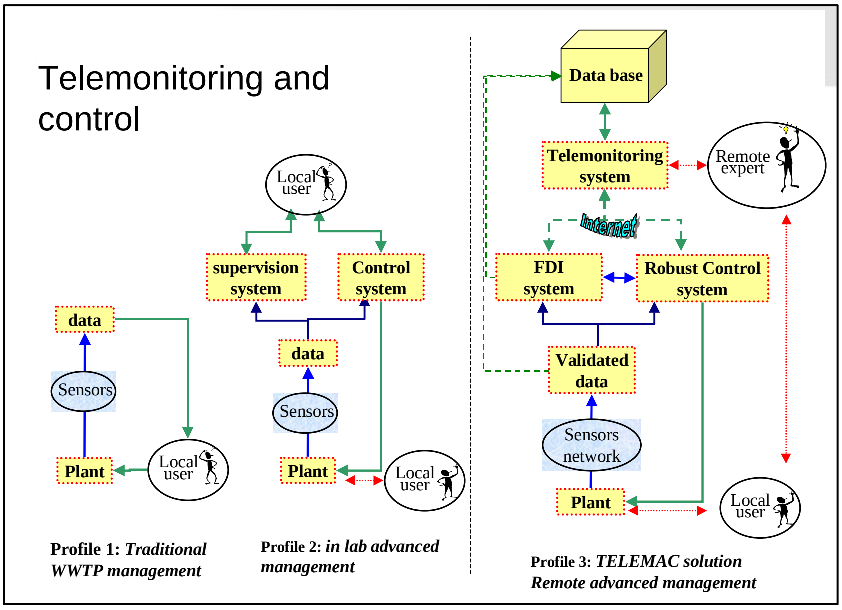

Figure 4. The evolution of monitoring and control of wastewater treatment plants.

Figure 4 shows how the TELEMAC project represents an important advance in remote monitoring and control of wastewater treatment plants. Profile 1 shows the traditional practice on an isolated plant. Profile 2 shows how TELEMAC laboratory prototypes evolved. Profile 3 shows the full TELEMAC solution to sharing expertise while maintaining local control. The Database, Telemonitoring system and Expert are based at the Telemonitoring and Control Centre (TCC) and are remote from the local user. Other components in Profile 3 are local to individual plants. The icons for the local user show the transition from puzzled in Profile 1 to enlightened in Profile 3.

Profile 3 introduces the monitoring and control of multiple plants from a single remote centre, the TCC. This is a step towards a full Grids-based system, though there is as yet no concept of identifying and combining resources according to specific needs: the system components and linkages are predefined and inflexible.

It is possible to abstract the essential components of the above model so as to prepare for a Grids-based solution.

Local User: Needs to be able to operate the plant in normal mode and receive warning of possible excursions. They will seek and receive advice from a Remote Expert.

Remote Expert: Needs to monitor each individual plant, compare with reference models, issue advice and alerts to local operators/users. The remote experts service a TCC.

Local Plant: Different plants have different arrays of sensors, have different volumes and operating principles. Each plant has its own data validation and consistency check for fault detection and isolation. Individual variables and combinations of variables are validated. In laboratory prototypes, multiple sensor consistency for the same variable can be used for calibration. Outputs of the FDI are used to provide robust control guidance. Each plant is serviced by a TCC; of course a single TCC may service multiple plants.

Telemonitoring and Control Centre: Receives and stores validated data from local plants. It provides advice from monitored data in response to enquiries. It pools models to generalise expertise. It revises models as new situations are recognised. The TCC is responsible for holding the models and data for its plants. There are mathematical analytical and simulation models as well as data mining models.

4.2 Knowledge, information, computation/data Grids

A general architecture has been proposed for structuring knowledge, information, and data/computation in a Grids context [8]. This architecture, shown in Figure 5, represents the conversion of data to knowledge and then using the knowledge to exercise control. Explicitly the control is over the data and its processing but ultimately it is concerned with changing the data in the real world. Homogeneous access to heterogeneous distributed data occurs in the information layer. As well as including data mining technology the knowledge layer encompasses human experts and decision makers. This model is therefore compatible with the approach taken in TELEMAC.

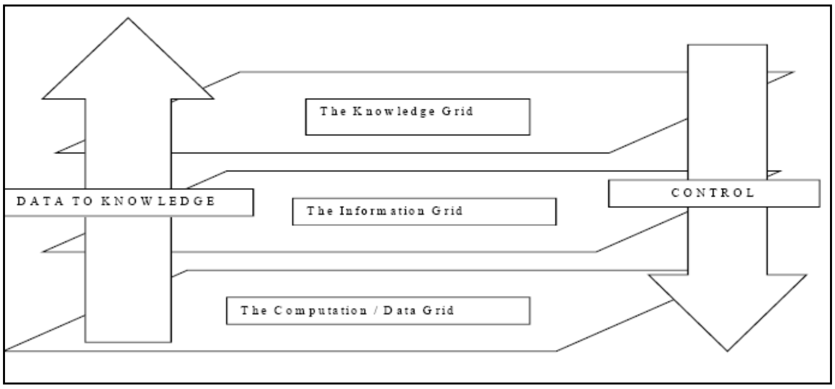

Figure 5. The Knowledge, Information, Computation/Data Grids (taken from [8])

4.3 The Grids perspective for leveraging knowledge

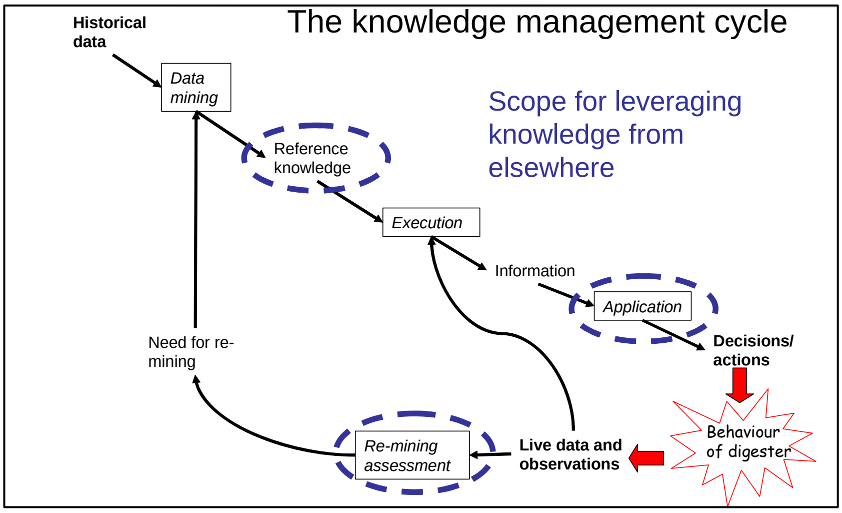

Figure 6. The knowledge management cycle.

Figure 6 shows the mining of historical data to produce reference knowledge and models that can be applied to current behaviour of the digester plant. The cycle is closed by the observation of the resulting behaviours leading to a need for re-mining, if there are deviations from what was expected. The ovals with broken lines indicate opportunities for leveraging knowledge obtained from elsewhere. For example, reference knowledge obtained about the behaviour of a digester in a state of hydraulic overload might be generalisable to other digesters of the same class, and usable in managing such states in future.

Large companies are likely to opt for an intra company TCC while the many small wineries might collaborate through geographically local TCCs. In either case there is scope for leveraging knowledge that has been derived about a particular situation by applying it in other circumstances, typically to a different plant. This leveraging should be done in a transparent way. It is therefore anticipated that a Grids infrastructure will provide the appropriate user transparency for this to proceed because it provides access to resources. Now it is not necessary for the Remote Experts to be located at every TCC.

4.4 Grid resources

With reference to Figure 6, it is possible to identify a number of classes of resource that can enable the leveraging of knowledge. These are:

Data mining tools. A selection of tools and methods such as those mentioned in section 3.3 may be available at the TCC. Not every TCC will have the same set, so there is potential for offering the tools themselves for use as a resource.

Datasets. Data sets from sensor data are steadily accumulated at the TCC and constitute the raw material for data mining that is a valuable resource in its own right. An ontology for resolving heterogeneities needs to be included.

Mined data. The results of the data mining, in the form of neural nets, rules, clustering parameters are obviously of potential value in dealing with situations on other plants. This is the classic example of transferring ‘knowledge’ from one plant to another.

Human expertise. It is important not to forget that the expertise of the remote experts is itself a kind of resource that can benefit the operation of multiple plants in a Grid.

5 Grids based approach to TELEMAC

5.1 Generalising the problem

From a Grids perspective we can consider each network to consist of a set of nodes (in TELEMAC each of these is the local computer associated with a digester) and a set of decision centre nodes (in TELEMAC a TCC). Figure 7a. shows a network of sensor arrays. The sensors are labelled for reporting variables A,B,C,D,E etc. A1, A2 are two different sensors reporting on variable A. The plant’s local computer acts as the node, validating the data for that plant, and passing it to the TCC. In some circumstances control action may be passed from the TCC to a node for action on the plant controls. In Figure 7b the Decision support centre node comprises the remote experts, the validated data and models, and the data mining and knowledge investigation, tools, (DMKI)

.

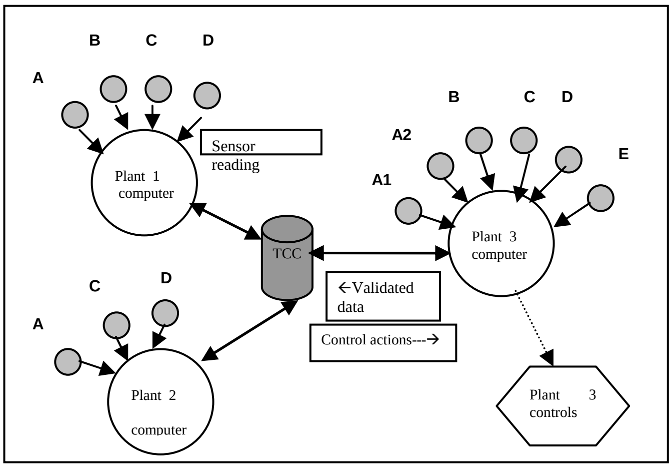

Figure 7a. A network of sensor arrays.

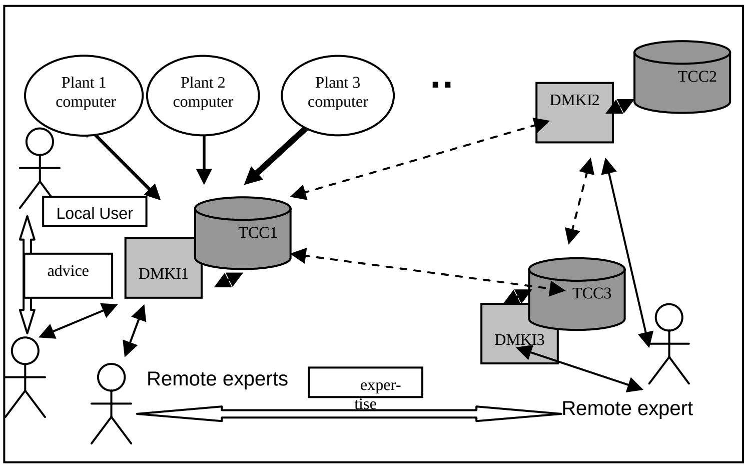

Figure 7b. A network of data, information and knowledge sharing.

.

5.2 Metadata

Metadata is required in a Grids system to represent properties of the Grid resources and allow reasoning over them to locate and deploy resources. The terms applicability, transformability, and reliability emerged as important metadata attributes for reworking TELEMAC in a Grids architecture. These terms are discussed in relation to the data mining models and resources.

Data Mining Models

Applicability: this class of metadata identifies circumstances under which the model can be deployed with confidence on the basis of the model generation eg the type of process, and the range of sensors available. It would normally be based on expert knowledge.

Transformability: this metadata identifies expert judgement about whether the models estimates can be used in (gu)estimating different regimes.

Reliability: this metadata identifies the confidence in the derivation of the model viz: the goodness of the model assessed using training and testing data, and any constraints that need to be considered.

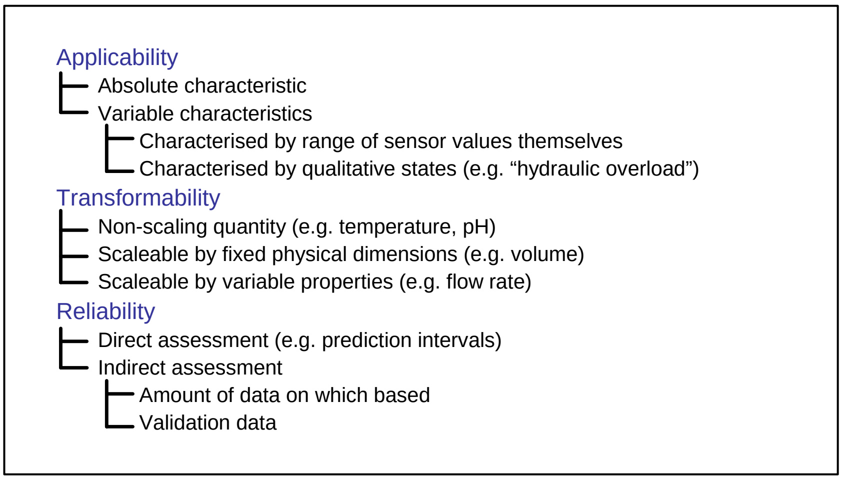
Applicability
Absolute characteristic
Variable characteristics
Characterised by range of sensor values themselves
Characterised by qualitative states (e.g. “hydraulic overload”)
Transformability
Non-scaling quantity (e.g. temperature, pH)
Scaleable by fixed physical dimensions (e.g. volume)
Scaleable by variable properties (e.g. flow rate)
Reliability
Direct assessment (e.g. prediction intervals)
Indirect assessment
Amount of data on which based
Validation data

*Figure 8***.** Metadata about datasets archived and available for data mining

Data Mining Resources

Figure 8 shows the metadata relating to datasets available as a Grid resource.

Applicability:

Absolute characteristic: this is a fixed feature of the system that never changes e.g. digester process type but not a potentially varying characteristic such as internal volume.

Variable characteristics: Generally each variable series is characterised by statistical summary data such as stationarity or variance. The series from the sensors are considered individually to determine the span of the variable and missing values. Qualitatively different behaviours of the digesters are characterised as states bound by ranges on subsets of the variables. Using these states an expert would be able to assess a priori whether they were not suitable for modelling other states e.g. data relating to hydraulic overload would not give a good indication of the behaviour of organically overloaded states.

Transformability:

Direct instrument readings sometimes need transforming to a consistent scale.

Non-scaling: these are quantities that typically have a direct scientific role such as temperature and pH.

Scalable by fixed physical dimensions: typically converts extensive to intensive e.g. using volume to scale bio-gas flow rates to m^3 per day, or to convert between time and frequency. Within this category we include time scale synchronisation where a variable is mapped to a different interval.

Scalable by a variable: maps to a new variable of interest e.g. using differences to remove a trend in a variable or produce a derived variable e.g. HRT is Volume/<inflow-rate> which is the length of time taken to feed into the digester the volume of liquid equal to the digester’s volume.

Reliability:

Direct assessment: These are methods where the prediction on the target data generates an estimate of the error. E.g. Prediction intervals can be obtained directly from neural net models of the unseen targets. Bootstrapping is widely used as an alternative approach for non-heteroskedastic data; it produces multiple models each on a variant of the training data.

Indirect assessment: These are methods where an estimate of the error is based on the quality of the model fit to its training and validation data. Eg information criteria and characteristics of residuals in linear regression.

5.3 Resource discovery

Having established a collection of resources with associated metadata, resource discovery proceeds by locating resources that satisfy the current needs of the user (at a TCC). Urgency and novelty of the digester state are factors that need to be taken into account when identifying potential resources such as data mining models that can be deployed. If a digester is in an alarming state which the Remote Experts have never seen before, then the experts would cast the net wider to look for resources that might help with the situation - accepting data mining models that are less reliable, for example, because at least they might offer some information of value. The broker would seek resources using such criteria[3]. Firstly it would need to match digester type and sensor set available in the archive; it would perform measurement unit conversion as appropriate. Then a suitable set of models would be selected with appropriate guidance. The system may even provide the Remote Expert with functionality that will advise on the urgency of the problem and whether it is novel.

5.4 Conclusions

TELEMAC is representative of a class of systems: networks of sensor arrays with significant heterogeneity and varying reliability. The sensors respond and report at different frequencies. Models need to be updated episodically over time as new data changes the characteristics being monitored. Expert knowledge can be deployed in different ways from advisory to automatic control. The knowledge base is used to infer behaviour of systems with different characteristics. The Grids architecture provides a knowledge, information and data architecture that enables a structured approach to developing this class of system.

Acknowledgements

This paper refers to work carried out under the IST project TELEMAC (Telemonitoring and Advanced Telecontrol of High-Yield Wastewater Treatment Plants). Especial thanks are due to J.P. Steyer (INRA), L. Lardon (INRA), O. Bernard (INRIA) and B. Le Dantec (ERCIM).

References

[1] TELEMAC: Telemonitoring and advanced telecontrol of high yield wastewater treatment plants, IST project no. IST-2000-28156, http://www.ercim.org/telemac.

[2] O. Bernard et al, An integrated system to remote monitor and control anaerobic wastewater treatment plants through the internet. *Water Science and Technology*, **52**(1–2), 457–464, 2005.

[3] K.G. Jeffery, Next generation GRIDs for Environmental Science, *Environmental Modelling and Software* xx, (2005) ppxxx ,2005

[4] H. Macarie, Overview of the application of anaerobic treatment to chemical and petrochemical wastewaters, *Water Science and Technology*, **42**(5–6):201–213, 2000.

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

[6] M. Dixon, J.R. Gallop, S.C. Lambert, and J.V.Healy, Experience with data mining for the anaerobic wastewater treatment process, *Environmental Modelling and Software*,pp xx, 2005, accepted.

[7] M. Dixon, J.R. Gallop, S.C. Lambert, L. Lardon, J.V. Healy, and J.P. Steyer. Data Mining to Support Anaerobic WWTP Monitoring, *IFAC Workshop on Modelling and Control for Participatory Planning and Managing Water Systems*, Proceedings CD. Venice 2004 http://epubs.cclrc.ac.uk/work-details?w=30122

[8] K.G. Jeffery, CRIS and Open Access, *Proc. World Library and Information Congress: IFLA 71*, Oslo, Norway. http://epubs.cclrc.ac.uk/work-details?w=34228, 2005

Appendix - Some heterogeneity issues

In addition to the usual problems of heterogeneity associated with data and their schemas such as consistency of names, scaling, units, applicability range there are some heterogeneities which affect data mining models from arrays of sensors. Data heterogeneity arises at two levels, from the diversity of sensors installed on what are essentially different instances of the same process, and from intrinsic differences between processes. For the first of these, unavoidable heterogeneities arise from the following:

1. different types of sensors measuring a given physical quantity by a different process;
2. different initialisation calibrations of the same sensor types;
3. complete failure of a sensor;
4. partial failure of the sensor through contamination, saturation, or drift.
5. different sampling frequencies and process time-constants

Also the anaerobic digesters themselves operate on different principles and are of different sizes. There are practical heterogeneities that arise from scaling variables; sometimes a key dimension is unknown or changing.

Given these limitations, there are implications for metadata representing the applicability and trustworthiness of results, and for the trade-off between the need for information and the possible unreliability of the information sources.