DATA MINING TECHNIQUES APPLIED TO POWER PLANT PERFORMANCE MONITORING

D. Flynn, J. Ritchie and M. Cregan

The Queen's University of Belfast, N. Ireland

Abstract: The incorporation of distributed control systems in power plants has become increasingly standard. However, particularly under alarm conditions, ease of access to plant-wide signals complicates the task of monitoring plant operation. Data mining techniques are ideally suited to dealing with such data, which tends to be highly correlated and collinear. Typically, models identify relationships under normal conditions from historical data. Subsequently, these discovered relationships form the basis for detecting unusual deviations from trained behaviour. Principal component analysis and partial least squares techniques are demonstrated, and shown to take full advantage of readily available DCS data, and provide an informative monitoring method.

Keywords: performance monitoring, power station boilers, partial least squares

1. INTRODUCTION

Data mining is a generic term encompassing a wide variety of techniques which attempt to identify novel, and hopefully informative, patterns in data. Although, conceivably containing valuable information, data is generally archived, with its full potential never realized. Successful applications have been associated with industrial processes and scientific research fields, e.g. chemometrics and chemical engineering (Chen and Liao, 2002), industrial process control (Sebzalli, *et al*., 2000) and to a lesser extent power engineering (Rayudu, *et al*., 1997). Emphasis has been placed on creating online operator support systems for fault detection and diagnosis, and high-level interpretation of system operation and performance for engineers.

Ballylumford power station is the largest power station in N. Ireland, housing 6 thermal units, 3 x 120 MW and 3 x 200 MW. Recently, 500 MW and 106 MW CCGTs have been commissioned. The primary data source within the power station is the distributed control system (DCS) which records exceeding 15,000 analogue sensors. The frequency of measurement and distribution of sensors provide significant redundancy, which, as discussed in Section 2, can be exploited for fault detection and signal replacement using techniques such as principal component analysis. Furthermore, within these records there is potential information regarding factors affecting daily plant operation, but obscured by the sheer volume of data presented. This information may help improve plant operation by identifying variables influencing efficiency and plant performance, while enabling a comparison of different operator shifts and comparative performance across individual units. The monitoring of unit efficiency and emissions levels, as measures of plant performance, is discussed in Section 3, and partial least squares (PLS) is demonstrated as a viable solution. The PLS approach is then extended in Section 4 by incorporating a neural network for the inner mapping to enable modelling of plant behaviour over non-linear conditions.

2. PROCESS AND SENSOR FAULT MONITORING

With industrial processes, and the associated computer support systems, becoming ever more complex the challenge for an operator to detect and cope with real-time problems becomes ever more challenging. Introduction of a DCS can provide quantifiable improvements in both productivity and plant manoeuvrability. However, a significant side effect has been increased accessibility to a range of plant-wide signals. Consequently, a relatively common experience with any monitoring system is that of faulty sensors. The process operators must also distinguish genuine faults, where unusual measurements actually reflect

plant behaviour, and/or variations in plant performance arising from changes in operating conditions, product quality, etc. Fortunately, within a power station environment many sensor measurements are highly correlated due to parallel paths for the *steam* and *gas* circuits, a closed loop for the steam / water circuit, etc.

When a particular sensor becomes faulty it is desirable to first detect that there is a problem, identify the failing signal, and finally, disable the sensor or, if possible, reconstruct the readings. For a process or actuator fault a different strategy is required. With the fault identified and diagnosed by the operators, the plant control systems may be able to minimize the effects of the fault. Any degradation in performance should be considered, however, along with longer term implications for plant life and maintenance. Alternatively, entire subsystems, or the process itself, may be taken off-line for further assessment.

2.1 Principal component analysis

Perhaps, the simplest method of detecting anomalous behaviour is univariate statistical monitoring, whereby upper and lower bounds are defined for each signal. However, as but one example, a *stuck* sensor may present a faulty value which is actually within limits. Should the process be well defined and comprise only limited inputs / outputs, model-based approaches may be constructively applied, assuming fault mechanisms are known and have been incorporated into the model. However, this task can be time consuming, requiring comprehensive application knowledge.

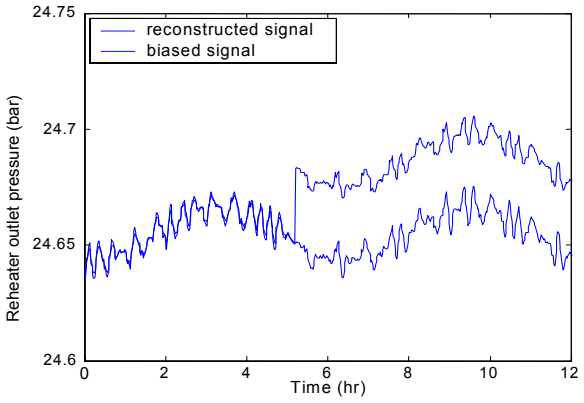

Fig. 1. Reconstructed reheater outlet pressure signal

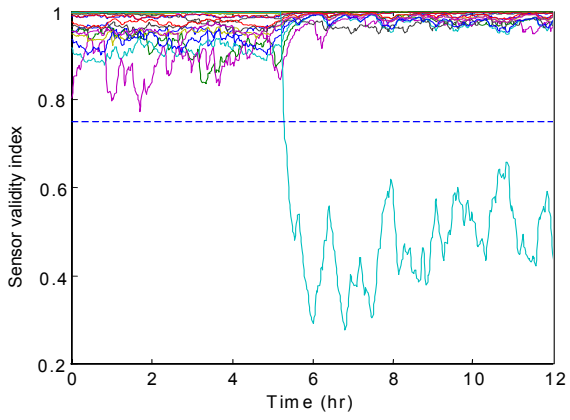

Fig. 2. Sensor validity index – sensor bias

In contrast, data mining techniques tend to be *data* rather than *knowledge* driven, and require much less refinement for particular processes. The approach adopted here is principal component analysis (PCA) which aims to reduce the dimensionality of a set of variables, while retaining as much of their variance as possible. This is achieved by identifying linearly independent latent variables (principal components). A reduced set of principal components may then be selected which capture the essential correlations and the majority of the observed variability. PCA has been used in a wide range of applications including batch monitoring from an industrial polymerisation reactor (MacGregor and Kourti, 1995) and sensor fault detection in a boiler process (Dunia, *et al*., 1996).

Since a system may have hundreds, if not thousands, of sensors, multi-block methods can be introduced, for convenience and practicality, by defining distinct linking subsystems (Nomikos and MacGregor, 1994). Such an approach is particularly convenient here for the 500 MW CCGT, consisting of a multi-shaft arrangement of 2 x 160 MW gas turbines (GTs), supplying a 180 MW steam turbine (ST). A single model can thus be created for each gas turbine. Further subdivision into condenser, HRSG and turbine stages can also be readily identified. The advantages of multi-block methods become apparent when fault identification is considered. Should a sensor fail, then only the model associated with that particular section of the plant will be affected, at least at first, making fault diagnosis that much more straightforward.

Once a (multi-block) model for normal operating conditions has been developed, it may be used to determine whether recorded plant measurements are consistent with historical values and neighbouring sensors. Calculation of the squared prediction error (SPE) and Hotelling's T^2 test can quickly help identify differences between the actual and reconstructed value of a variable (Sebzalli, *et al*., 2000). Both indicators are affected by noise, etc., but false alarms can be largely eliminated by simple filtering, and adjustment of the associated test threshold. Plotting of t scores can be combined with the above methods to distinguish between a failing sensor and a fault. When a process fault occurs, the individual points on the t score plots drift from the normal grouping into a separate cluster. The relative position of these clusters can assist in diagnosis (Kourti and MacGregor, 1995).

Having confirmed that there is a sensor fault the next step is to identify the failing sensor. Identification and reconstruction can be achieved by assuming each sensor has failed, and observing the reduction in SPE before and after reconstruction from the remaining signals. However, in certain situations the reduction in SPE can affect all inputs, making the faulty sensor unidentifiable. Instead, a sensor validity index (SVI) can be calculated for each variable (Dunia, *et al*., 1996). System transients and measurement noise can lead to false triggering so each signal is filtered and compared with a user-defined threshold.

2.2 PCA tests and results

Within Ballylumford power station data is archived from the DCS using PI universal data server. Given the large number of analogue sensors involved, spread across multiple units, data compression is essential. However, care must be taken that limited storage capabilities do not ultimately impinge on data quality. For each variable, an exception deviation is defined such that a value is only recorded when the signal changes sufficiently from that last recorded. The data is subsequently compressed by defining a compression deviation, whereby points are only archived when recorded values fall outside an error boundary parallelogram. Finally, the data is saved as scaled integer values. Consequently, depending on the chosen PI settings, only a few, or instead tens and hundreds of datapoints may be recorded over a given period.

Experience has shown that operators and engineers tend to focus on key unit parameters, and consequently for some sensors data is available at high resolution and quality, while for many others detail has been lost. For high-level process monitoring and fault identification and analysis this arrangement has worked well. However, one of the strengths of PCA, and later in this paper partial least squares, is its ability to integrate and assimilate information from multiple, correlated plant sensors. To an extent this ability has been compromised, although the main objective within a PCA model is to identify *steady-state* relationships between signals under *normal* conditions. The creation and validation of dynamic process models would undoubtedly be more challenging.

Training data for PCA analysis was obtained by selecting periods of interest in the PI data archive, with snapshots of all DCS process variables created at specified time intervals using linear interpolation between the archived datapoints. The PCA methods described earlier were then applied to a phase 2, 200 MW unit at the power station. Individual faults were detected using the SPE and T^2 indicators, and further clarified using the sensor validity index, with the reconstructed value substituting for the failing sensor. Otherwise, the t scores were examined to confirm that the fault was actually with the plant, and corrective maintenance or other actions scheduled.

Two PCA models were developed for normal operating conditions around generation outputs of 100 and 150 MW, with 2 principal components considered sufficient in both cases, following the PRESS statistic (Wold, 1978). Using a validated simulation of the plant (Lu and Hogg, 1995) for convenience, a 0.1% bias was introduced into the reheater outlet pressure signal after 5 hours of operation while generating at approximately 150 MW, Figure 1. Based on a 95% confidence limit, the SPE and T^2 tests promptly detect the fault after 30 and 10 minutes (not shown). Figure 2 shows the SVI fluctuations for each variable, with the reheater outlet pressure signal readily identified as being in error. The index for this signal falls in the range 0.3 – 0.6, against a threshold of 0.75. As the fault is with the sensor, a reconstructed (unbiased) measurement is substituted, Figure 1. Although not directly utilised for control, this signal forms an input to an on-line, advisory efficiency monitoring system on the plant, and thus invalid measurements may unduly influence operator actions.

With the plant now operated at a load of 100 MW, a positive drift is applied to the main steam pressure signal, regulated by a PI controller operating on fuel flow. Since the faulty signal is fed back for control there is minimal impact on the measured value, Figure 3. Instead, the controller, observing that the pressure signal is drifting upwards decreases fuel flow, and associated air flow, so that the faulty sensor indicates the correct value. In actuality, the main steam pressure is being progressively reduced, with knock-on effects for many sensors around the plant. Using the SPE and T^2 measures, the fault is detected after 50 and 30 minutes (not shown), and confirmed by the SVI plot for each sensor, Figure 4. Unlike Figure 2, the effects of this fault are more significant across the plant. It is now more challenging, although still straightforward, to identify the failing sensor. The steam pressure signal can be reconstructed by the PCA model and Figure 3 confirms that the pressure falls as a result of the fault.

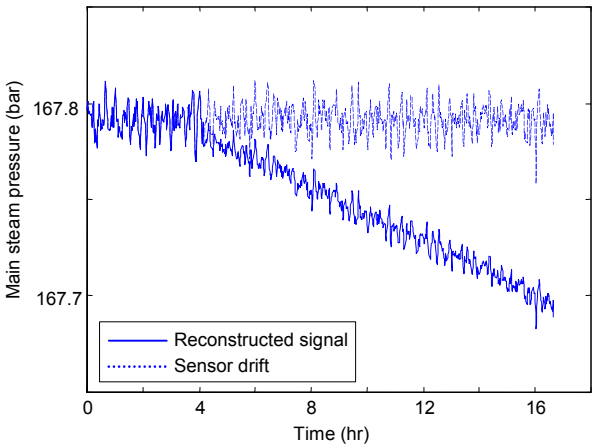

Fig. 3. Reconstructed main steam pressure signal

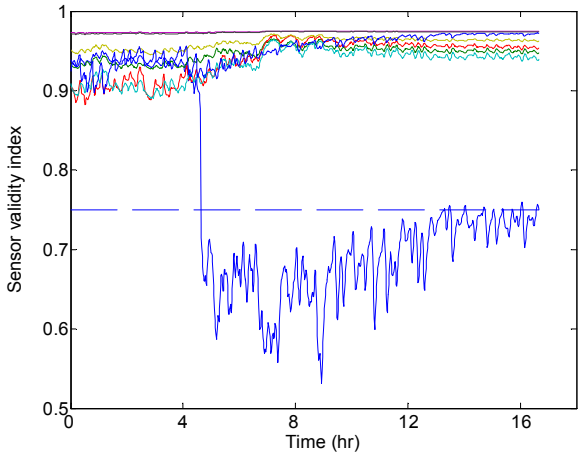

Fig. 4. Sensor validity index – sensor drift

3. OPTIMISATION AND DIAGNOSIS

Ballylumford power station operates on both heavy residual fuel oil and gas. Although initially designed for oil operation alone, the plant is now predominantly operated with gas. However, this is provided under an interruptible contract, requiring operation using oil on occasion. Not surprisingly, using two different fuels requires subtle changes in control strategies by the operators. Optimisation of unit efficiency is a multi-dimensional problem with factors such as fuel calorific value, burner configuration, daily cyclic variations of local sea temperature, etc. contributing to the end result. Issues such as condenser clogging are common to both types of operation (Ritchie and Flynn, 2003). However, soot buildup, for example, is much more significant for oil operation. Attemperator spraying is also affected: for gas operation, the flame ball is higher up and further back in the furnace, resulting in a differing heat distribution and more reheater spraying.

3.1 Partial least squares

With power generation becoming an increasingly competitive market place it is important that individual units operate at maximum possible efficiency while meeting contractual load obligations and monitoring emissions levels. From available plant records it is possible to investigate periods of operation identified by the operators as being representative of *good* plant performance. Potentially, this information can then be used to develop a *best case* model.

Partial least squares (PLS) is a robust, multivariate linear regression technique suitable for the analysis and modelling of noisy and highly correlated data. Using techniques previously applied in PCA, a reduced order model is developed which attempts to explain the variation in the process that is most predictive of the product quality variables. This procedure is enhanced, through linear regression, to provide a relationship between the process variables and the product quality variables. The technique has been successfully applied to various process control, and chemical engineering applications, such as monitoring of both a fluidised bed reactor and extractive distillation column (Kresta, *et al*., 1991) and steel casting (Zhang, *et al*., 2003).

3.2 PLS tests and results

PLS models were created using plant data gathered over the period of two weeks for a phase 1, 120 MW unit which was being operated for that period on oil. A hybrid model could be formed by gathering data for both oil and gas operation, but it was viewed as being much more informative to create distinct models for each fuel. Although the unit went through several load cycles during this period, the model was specifically trained for operation between 100 – 120 MW, as the unit was normally scheduled to generate within this range. Three quality variables were selected, namely unit thermal efficiency, NO_x and SO_x emissions, with distinct models created for each output variable.

Having developed distinct PLS models, it was of some interest to then determine how each input variable contributed to the % unexplained variance for the first component of each model. This is a measure of how individual variables affect the quality variable. Figures 5a and 5b show the % explained variance of the data block, for both efficiency and SO_x models. There are many similarities between the bar charts with, for example, unit output (1), primary steam flow (2), economiser feed inlet temperature (7), etc. being significant for both models. It is of greater interest, however, to identify differences between the charts - boiler flue gas oxygen (18) is significant for SO_x, while variables such as final outlet steam temperature A (8) and B (9), HP turbine exhaust temperature (29), etc. are much more significant for the efficiency model. These results highlight the most important variables to be monitored / adjusted when attempting to achieve different operational goals.

The monitoring capabilities of the PLS models could now be investigated on the plant. Figure 6 shows the normalised efficiency during a latter period, with the plant again running on oil. Superimposed on the graph is the PLS estimate of the plant's efficiency, with a clear distinction visible between the characteristics.

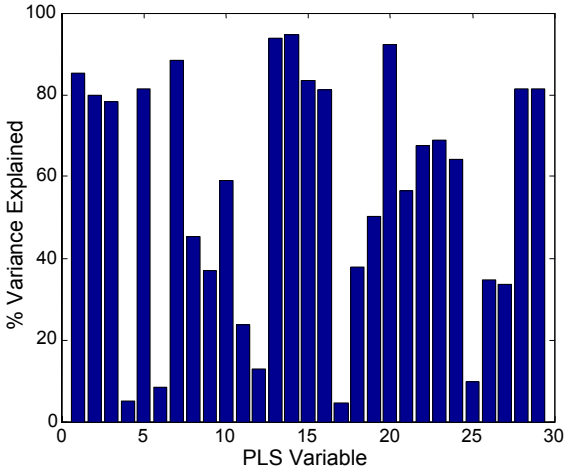

Fig. 5a. Bar chart – efficiency model

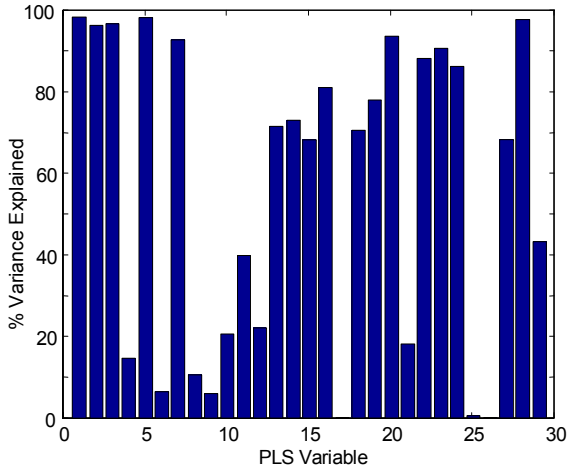

Fig. 5b. Bar chart – SO_x model

If the scores for the first two components of the PLS model are examined, Figure 7, region *A* represents the training data while region *B* represents the latter period. The t_1 score, in particular, is now significantly higher, suggesting that examination of how this score is formed (from the measured variables) should reveal the affected plant section. Figure 8 plots the change in t_1 score contribution for each PLS variable, revealing that condenser cooling water temperatures A (22) and B (23), and condensate temperature (24) are unusually high. It is known, from examination of the operator logs, that the unit was switched off on the next day, following which, as part of maintenance procedures, debris was removed from the condenser pipework.

4. NON-LINEAR PLS MODELLING

It was shown in the previous sections that linear models, both PCA and PLS, operate well over a limited range. However, all processes are inherently non-linear, with power generation being no exception. When applying PLS to a non-linear problem the minor latent variables cannot always be discarded, since they may actually contain significant information about the non-linearities. Intuitively, however, the non-linearities can be recognized using non-linear transformations of the original variables. More advanced methods have also been proposed including non-linear extensions to PCA (Li, *et al.*, 2000), and applying neural network, fuzzy logic, etc. methods to directly represent the non-linearities (Tan and Mavrovouniotis, 1995).

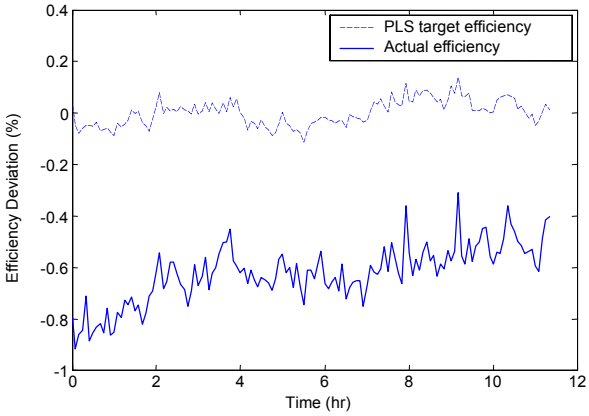

Fig. 6. PLS target efficiency for oil operation

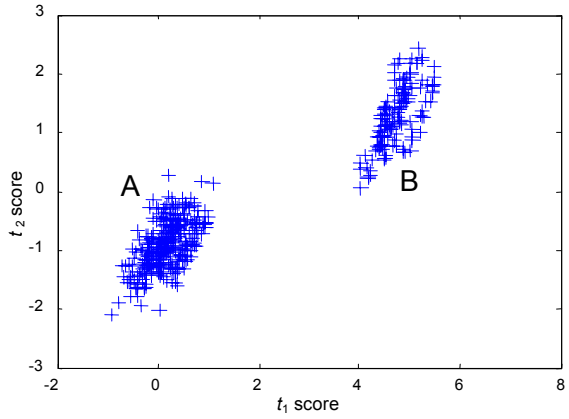

Fig. 7. t_2 v t_1 scores plot – condenser fouling

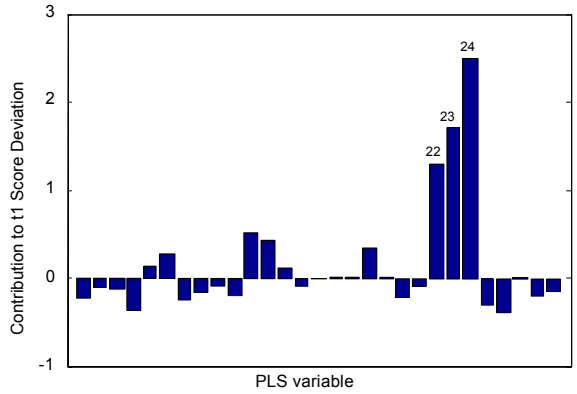

Fig. 8. Contribution to t_1 score – condenser fouling

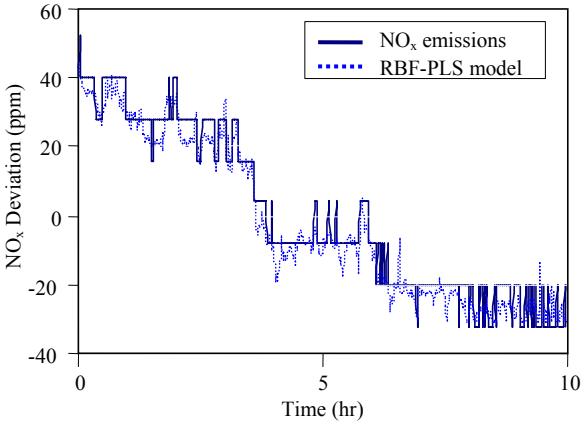

Fig. 9. RBF-PLS model – predictive performance

A more structured approach, considered here, is to introduce a non-linear function (a neural network is used for each component) linking the output scores *u* to the input scores *t*, without modifying the input and output variables (Baffi, *et al.*, 1999). Since the neural network is merely required to capture the relationship between *t* and *u*, a variety of neural structures can be arbitrarily applied. In this case a radial basis function (RBF) network has been chosen over other approaches. The advantage is that having chosen the number and position of basis function centres (using, for example, *k* means clustering, etc. and / or a priori experience) the remaining weights then appear as linear terms, which can normally be conveniently determined using least squares techniques. It should be noted, however, that while the (non-linear) model dimension is reduced, the dependency between latent variables is not as transparent as in linear methods.

4.1 RBF-PLS tests and results

RBF-PLS models were trained using efficiency, NO_x and SO_x emissions as quality variables, using data from a 200 MW unit firing on gas, across the unit's full operating range, 60 - 200 MW. Selection of centres and training of the RBF networks for each component was performed using the Matlab neural network toolbox. The NO_x model, for example, required 30 neurons for the first component, and 5 neurons for subsequent components, such that the first component explained 95% of the variance in the quality data.

Figure 9 confirms the predictive performance of the RBF-PLS model, using one component, with the unit load varying between 80 and 200 MW over a 10 hr test period. What is most clear from the figure is the poor quality of the NO_x data itself, as measurements are only available every 30 seconds, from alternate side of the boilers, and at a limited resolution.

5. CONCLUSIONS

The introduction of distributed control systems into many industrial processes has brought clear advantages in terms of productivity, plant manoeuvrability, etc. However, the vast amount of data which then becomes available is generally put to minimal use. This historical resource can, however, be exploited using data mining techniques. The application considered here was that of process monitoring at a thermal power station, taking advantage of data gathered from existing monitoring equipment. Traditional operator practice has been reactive, whereby actions are taken following the triggering of process alarms, often set over-responsive and mode insensitive – PCA methods have enabled a more proactive role, providing early warning of irregularities, and perhaps most importantly an increased awareness of data potential. Faults arising both with the plant and instrumentation were first investigated. A PCA model under normal operating conditions was created, which focused on identifying unusual deviations. PCA models were created for limited operating ranges and their ability to detect, and ultimately correct, sensor problems were discussed.

Monitoring of operating performance, and in particular unit efficiency and environmental emissions measures, can be greatly assisted through the availability of extensive historical records. Previously, only a restricted number of high-level variables (indicators) were regularly monitored, with the consequence that deviation in current unit performance could not easily be distinguished from ongoing plant problems, or seasonal effects. Distinct PLS models were developed to model thermal efficiency, NO_x and SO_x emissions. Subsequently, by running these models in parallel with the plant, operators could monitor how close to optimum the plant was performing. Furthermore, by tracking t score plots, and observing how individual signals contributed to the PLS scores it was shown that discrepancies in plant performance could be pinpointed to particular plant items. The PLS models were trained over a limited operating range, corresponding to the unit's normal loading range. In order to capture global (non-linear) behaviour, neural strategies were proposed, with the inner mapping between the t and u scores for each component replaced by a RBF network 'curve fit'. Subsequently, reduced order models were developed covering the operational unit range.

Future work will focus on applying PLS techniques to the two CCGTs within the power station. The ability of the multi-shaft CCGT to operate in several configurations will impact on operational procedures, and unit performance. Furthermore, in addition to the factors raised in Section 3, CCGT operation is particularly sensitive to variation in ambient conditions – a 5° C change in air temperature or a 25 mbar variation in atmospheric pressure will typically cause a 3% variation in maximum power output (Lalor and O'Malley, 2003). However, even measuring ambient temperature is not necessarily straightforward as sunny conditions can cause large fluctuations in measured values – multiple sensors are currently employed. Performance targets are thus likely to be more dynamic than for the conventional boiler plant.

REFERENCES

Baffi, G., E.B. Martin and A.J. Morris (1999). Non-linear PLS revisited (the neural network PLS algorithm). *Computers and Chemical Engineering*, **23**, 1293-1307.

Chen, J. and C. Liao (2002). Dynamic process fault monitoring based on neural network and PCA. *J. Process Control*, **12**, 277-289.

Dunia, R., S. Qin, T.F. Edgar and T.J. McAvoy (1996). Identification of faulty sensors using principal component analysis. *AIChE Journal*, **42**, 2797-2812.

Kourti, T. and T.F. MacGregor (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, **28**, 3-21.

Kresta, J., J. MacGregor and T. Marlin (1991). Multivariate statistical monitoring of process operating performance. *Canadian J. Chemical Engineering*, **69**, 35-47.

Lalor, G. and M. O'Malley (2003). Frequency control on an island power system with increasing proportions of CCGTs, *IEEE Powertech*, Bologna.

Li, W., H.H. Yue, S. Valle-Cervantes and S.J. Qin (2000). Recursive PCA for adaptive process monitoring. *J. Process Control*, **10**, 471-486.

Lu, S. and B.W. Hogg (1995). Integrated environment for power plant performance analysis and control design. *IFAC Control of Power Plants and Power Systems*, Cancun, Mexico, 37-42.

MacGregor, J.F. and T. Kourti (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, **3**, 403-414.

Nomikos, P. and J.F. MacGregor (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, **40**, 1361-1375.

Rayudu, R., S. Samarasinghe, D. Kulasiri and J. Ypsilantis (1997). 'Model-based learning for fault diagnosis in power transmission networks', *Engineering Intelligent Systems*, **5**, 63-73.

Ritchie, J. and D. Flynn (2003). Partial least squares for power plant performance monitoring. *IFAC Control of Power Plants and Power Systems*, Seoul, Korea, 744-749.

Sebzalli, Y.M., R.F. Li, F.Z. Chen and X.Z. Wang (2000). Knowledge discovery from process operational data for assessment and monitoring of operator's performance. *Computers and Chemical Engineering*, **24**, 409-414.

Tan, S. and L. Mavrovouniotis (1995). Reducing data dimensionality through optimizing neural network inputs, *AIChE Journal*, **41**, 1471-1480.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397-405.

Zhang, Y., M. Dudzic and V. Vaculik (2003). Integrated monitoring solution to start-up and run-time operations for continuous casting. *Annual Rev. Control*, **27**, 141-149.