Process Monitoring using a Combination of Data driven Techniques and Model based Data Validation

ARNAUD DUCHESNE[1]*, GEORGES HEYEN[2], PHILIPPE MACK[1], BORIS KALITVENTZEFF[3]

[1] Pepite s.a., Rue des Chasseurs Ardennais 4, B-4031 Angleur (BE)

[2] Laboratoire d'Analyse et de Synthèse des Systèmes Chimiques, Université de Liège, Sart-Tilman B6A, B-4000 Liège (Belgium)

[3] Belsim s.a., rue Georges Berotte 29A, B-4470 Saint-Georges-sur-Meuse(BE)

Process monitoring is made difficult when measurements are subjected to errors, since pertinent information is hidden in the measurement noise. To address this issue, one can use model based data validation, or rely on statistical techniques to analyze large historical data sets (data mining). An industrial case study is presented here, where a model based approach (data validation) is compared to data driven techniques.

Keywords: Data Validation and Reconciliation; Data Mining; Soft Sensors; Process Monitoring; Process Control

Current methods for process monitoring

Efficient process monitoring is a key issue in plant operation. However operators have to deal with measurement uncertainty, and take appropriate actions to address measurement errors.

Some sources of measurement errors depend on the sensors themselves:

intrinsic sensor precision is limited, especially for online equipments where robustness is usually felt more important than accuracy ;

sensor calibration is seldom performed as often as would be desired, since this is a costly and time consuming procedure, requiring competent manpower ;

signal converters and transmission add noise to the original measurement ;

synchronization of measurements may also pose a problem, especially for chemical analysis, where a significant delay exists between the sampling and the result availability.

Other errors arise from the sensor location or influence of external effects: for instance the measurement of gas temperature at the exit of a furnace can be influenced by radiation from hot wall in the furnace. Inhomogeneous flow can also cause sampling problems: a local measurement is not representative of an average bulk property.

A further source of error when calculating plant balances is the small fluctuations in the plant operating conditions and the fact that samples and measurements are not taken exactly at the same time. Using time averages for plant data partly reduces this problem.

Process state, including the value of KPI (key performance indicators), must be assessed with suitable precision to enable the optimization of operating conditions. Drifts in process efficiency have to be detected as early as possible, and faults have to be identified. Two strategies can be adopted for efficient process monitoring: one based on a first principle process model, used to reconcile measurements, or one based on feature extraction from a large historical data set.

Data validation [1, 2] uses sensor redundancy and a plant model to reduce measurement uncertainty and to calculate all non measured state variables of the system. Data validation is nowadays routinely performed for steady state processes and commercial software is available to implement it online [3, 4]. The data validation procedure comprises several steps.

The first one is the measurement collection. In well instrumented plants, this is nowadays performed routinely by automated equipment.

A second step is conditioning and filtering: all measurements are not available simultaneously, and synchronization might be required. Some data are acquired at higher frequency, and filtering or averaging can be justified.

A third step is to verify the process condition, and the adequacy of the model: for instance if a steady state model is to be used for data reconciliation, the time series of raw measurements should be analyzed to detect any significant transient behavior.

The fourth step is gross error detection: the data reconciliation procedure to be applied later is meant to correct small random errors, thus large systematic errors, that could result from a complete failure of a sensor, should be detected first. This is usually done by verifying that all raw data remain within upper and lower bounds. More advanced statistical techniques, such as principal component analysis, can also be applied at this stage. Ad hoc procedures are applied in case some measured value is found inadequate or missing: it can be replaced by a default value, or by the previous one that was available.

The fifth step checks the feasibility of data reconciliation. The model equations are analyzed, and the variables are sorted: measured variables are redundant (and thus can be validated) or just determined; unmeasured variables are determinable or not. When all variables are either measured or observable, the data reconciliation problem can be solved to provide an estimate for all state variables.

The sixth step is the solution of the data reconciliation problem. Each measurement is corrected as slightly as possible in such a way that the corrected measurements match all the constraints (or balances) of the process model. Unmeasured variables can be calculated from reconciled values using some model equations. The data reconciliation problem consists in identifying the state variables $\mathbf{x}$ verifying the set of constraints, and being close to the measured values in the least square sense, which results in the following objective function, for a nonlinear steady state model, and for cases where some variables z are not measured:

* *email: Ph.Mack@Pepite.be*

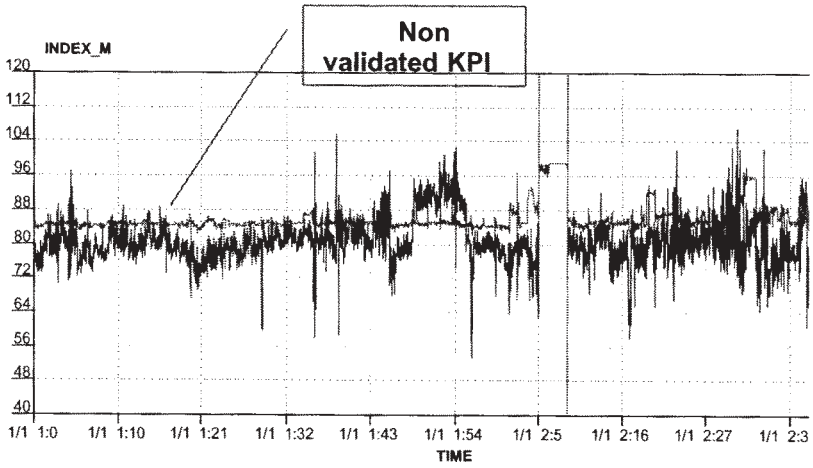

Fig. 1. noise reduction by using validation (raw values in red, validated results in blue)

$$\min_{x,z} (y-x)^T W (y-x) \quad s.t. \quad f(x,z) = 0 \tag{1}$$

where the model equations f are mass and component balance equations, energy balance, equilibrium conditions, link equations relating measured values to state variables (e.g. conversion from mass fractions to partial molar flow rates).

In the final step of the procedure, the software performs a result analysis. The magnitude of the correction for each measurement is compared to its standard deviation. Large corrections are flagged as suspected gross errors.

On the other hand, data mining uses large collections of historical data to seek the most favorable combination of operating parameters. Data Mining, also known as Knowledge Discovery in Databases (KDD), is an information extraction activity aiming at discovering new knowledge and facts from large databases. Data Mining uses a broad range of tools from statistics, automatic learning, pattern recognition, database technologies, visualization and artificial intelligence.

Manufacturing systems monitoring: sensors, PLCs, DCS, and SCADA systems allow the operators to monitor and control the manufacturing process in real time.

The whole measurements and actions (manual and automatic) made on the process are recorded and stored in historians that represent huge memories of the factory. To perform historical data analysis, experts are digging into the data off-line to detect flaws and improvement actions. This is also a unique opportunity to learn more quickly about the process and to detect hidden and complex relationships between all parameters. Given the increasing amount of these archives, the Data Mining solutions are more than welcomed to maximize the benefits of this data analysis task.

At the end of this life cycle, the knowledge synthesized by the Data Mining analysis is used by operators to bring improvements to their processes. This knowledge might also be recorded in a knowledge base used by an artificial expert system, e.g. in the form of a soft sensor.

Data clustering can reveal multiple ranges of operating conditions, and correlation analysis allows one to detect patterns in the data sets [5]. Both approaches provide help in process monitoring, but have complementary assets, as will be shown in the present case study.

An industrial case study

The case study focused on the steam system of a large industrial site (metallurgical plant, including coke furnaces, blast furnaces, steel plant, rolling mill, galvanization lines). Three steam generators are in operation (1x 120 ton/h, 130 bar, 530°C, 2x 100 T/h, 70 bar, 510°C). They mainly supply steam on site, but back pressure and condensing turbines also generate power. Multiple fuels differing in quality and cost can be burnt; some of them are by-products of the process (coke oven gas and low heating value blast furnace gas) and must be used in priority, while other imported fuels (natural gas and heavy fuel oil) come as supplements.

The goal of the study is to evaluate the energy efficiency of the steam generators, and to identify ways to increase the steam production, and consequently to raise the potential for electricity generation.

Methodology

Process data is collected automatically and values of the main process variables can be retrieved from the process information management system. Each of the 3 steam generators was first studied independently. Values for 70 process measurements were retrieved for a 5-month period, using 10 min averages.

The performance indices, like the thermal efficiency, are not measured directly, and must be evaluated from several measured variables. However the measurement uncertainty propagates to the estimates of the performance parameters, thus some noise reduction technique is needed to extract useful information. A steady state data reconciliation model was developed using Belsim-Vali software [4] and all data sets were processed in order to evaluate and validate several key performance indices, such as the boiler efficiency, the steam production, the fuel consumption, the oxygen content in the combustion gas.

The data base was also processed using data mining tools (PEPITo Data Mining toolbox, developed by Pepite [5, 6]). Several tools were exploited for data analysis: histograms, scatter plots, dendrograms, correlation analysis and principal component analysis [7]. Other tools were used later for modeling and knowledge discovery, like decision trees [8], artificial neural networks [9], and K-Means [10].

Data processing

The first attempt was to calculate the key performance indices using directly the raw measurements, but this provided little useful information, due to measurement uncertainty and noise. For instance, trying to calculate the energy efficiency directly from the measured values led to very noisy estimates, and sometimes unfeasible values (e.g. efficiency above 100%).This could be corrected using validated estimates (fig. 1). Adding validation results (e.g. validated efficiency) to the raw data sets provided additional dimensions to explore. Correlations between process variables and efficiency parameters were much clearer.

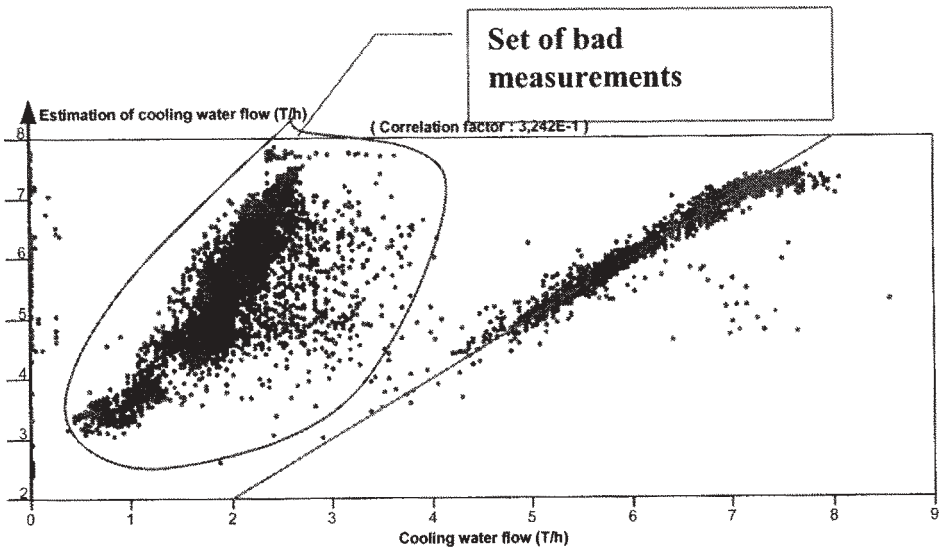

Fig. 2. Isolation of erroneous measurements

Data reconciliation allowed also detecting failing sensors: for instance, the measurement of oxygen concentration in the stack gas of one generator was systematically wrong. Temperature measurements located at the outlet of an air preheater were also flagged. These measurements were temporarily discarded, but data validation allowed obtaining estimates for those variables. Faulty equipment was also diagnosed: the efficiency of one pump was clearly below standard, and the equipment was replaced, which resulted in immediate savings.

In a few cases, the validation program could not provide a reliable answer, due to missing measurements for non redundant variables (temporary sensor failure), or due to convergence to a solution with large measurement corrections (thus with probable gross errors). These failures could be traced to operating conditions where the steady state assumption was not correct, and where operating parameters were modified suddenly (start up or shutdown of a boiler, change in fuel). The data mining toolbox allowed designing a filtering strategy that is able to detect most of the data sets where the validation program would fail (fig. 2); in this case, because of the poor status of the instrumentation system (occasional missing data) or because transients causing data inconsistent with steady state operation. Furthermore a neural network has been trained to provide estimates of the suspicious or missing measurements, thus allowing the validation program to return useful results in almost all cases.

As an example, 6433 data sets have been processed by data validation, resulting in precise estimates of the thermal efficiency of one steam generator. 1335 validation results were used as a training set in order to tune a neural network able to reproduce validated efficiency using raw measured values. The other data set were used to validate the predictive capability of the neural net.

Because the training has associated validated and raw measurement values, the neural network reproduces not only the relationship between process variables and efficiency, but it also involves the correction of the measurement bias (fig.3). It involves two 10-neurones hidden layers and handles 38 process inputs. This model is able to predict validated efficiency with a standard deviation of 0.085%, even when the validation c^2 test detects the presence of gross errors. This estimate is now displayed in real time in the control room (thus much faster than the validated value, that is available every 15 minutes), and provides a useful reference to the operator, who has some

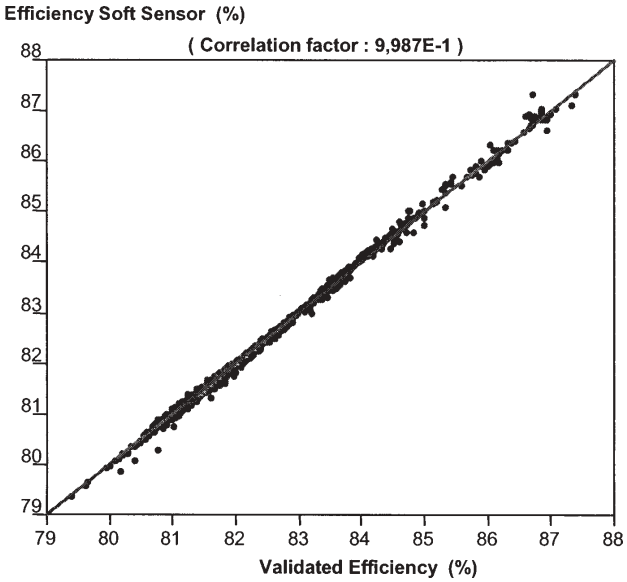

Fig. 3. Efficiency predicted by neural net, compared to validated estimates

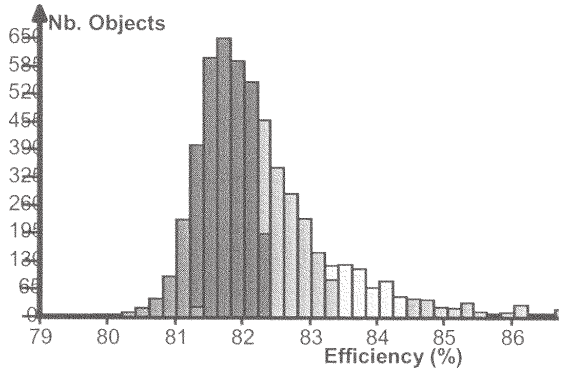

Fig. 4. Histogram: efficiency variation for one steam generator

immediate feedback when process parameters are modified.

The use of such a tool does not replace at all data reconciliation: in fact the neural network has to be trained periodically with updated reconciled values, in order to integrate changes in the process conditions, such as calibration or replacement of sensors. Furthermore the validation results are more complete.

Extrapolating our findings, we suggest that the synergy of both techniques allows to display most wanted key performance indices in real time and to access more

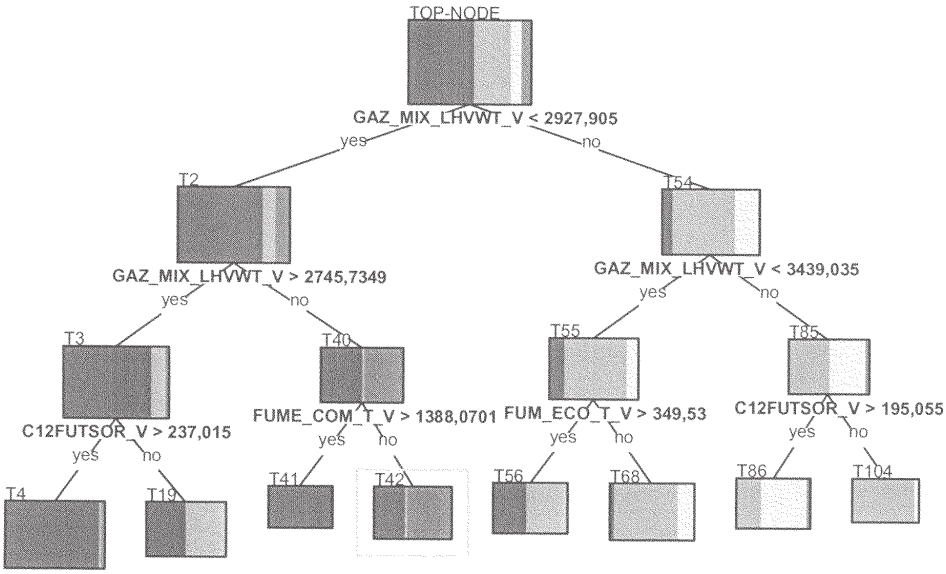

Fig. 5. decision tree to classify operating conditions according to efficiency

numerous quality information data to optimize operation. Let us mention the possibility to access in real time the validated parameters to be compared to the set points in Advanced Process Control systems.

The next step was to analyze the variability in the operating conditions, in order to try to identify those leading to the best efficiency. The range of efficiency variation is approximately 10%, as shown in figure 4.

The root causes for efficiency variations were explored by building a decision tree in order to classify all data sets. Figure 5 shows that just a few variables are needed to explain most of the variability. The most significant parameters appear to be:

1.the mixed gas lower heating value (50%)
2.the combustion chamber temperature(10%)
3.the boiler feed water flow rate (2%)

Results and discussions

This analysis provides clues on ways to improve the operation. The main decision has been to improve the control of excess air. The second one is to take advantage of design differences between the boilers, to select the right combination of boilers to operate according to the composition of the gas mix available. Coke oven gas is richer in hydrogen than blast furnace waste gas, and produces a flame that radiates better. This results in a difference in the internal temperature profiles, and a small but significant difference in efficiency.

Conclusions and perspectives for future work

This case study shows clearly that data driven techniques and model based validation can operate in parallel and benefit from each other. Synergistic effects have been demonstrated: data validation is able to reduce the uncertainty on measured process variables and calculated values of performance indicators. Working with reconciled data helps data mining in the identification of efficient operating conditions, and in the detection of abnormal process states.

Future developments are going on. They focus on the inclusion of the regression models in a decision tool, that should help the operator in optimizing the load distribution among all the available steam generators, in order to maximize the energy efficiency for a given power demand and a given gas mix availability.

Acknowledgements: This project was supported by the Walloon Region (F.I.R.S.T. Entreprise Program, Grant 5050)

References

1. N. ARORA, L. T. BIEGLER, G. HEYEN, Data Reconciliation Framework, in B. Braunschweig and R. Gani (eds) Software Architectures and Tools for Computer Aided Process Engineering, Elsevier, 2002
2. G. HEYEN, B. KALITVENTZEFF, Process monitoring and Data Reconciliation, in L. Puigjaner and G. Heyen (eds), Computer Aided Process Engineering, Wiley-VCH, 2006
3. B. KALITVENTZEFF, G. HEYEN, M. MATEUS TAVARES, Data Validation, a Technology for intelligent Manufacturing, in L. Puigjaner and G. Heyen (eds), Computer Aided Process Engineering,Wiley-VCH, 2006
4.*** http://www.belsim.com/Vali.aspx , accessed April 18,2007
5. *** http://www.pepite.be/en/produits/PEPITo , accessed April 18,2007
6. *** PEPITo Data Mining software UserGuide v1.5 (c) PEPITe SA, 2006
7. J. B. MACQUEEN, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297 - 1967
8. LOUIS WEHENKEL. Decision tree pruning using an additive information quality measure, Uncertainty in Intelligent Systems, Elsevier-North Holland, 1993, p. 397
9. C. M. BISHOP, Neural Network for Pattern Recognition. Clarendon Press, Oxford, 1995
10. FUKUNAGA, KEINOSUKE, Introduction to Statistical Pattern Recognition, Elsevier, 1990

Manuscript received: