Data Reconciliation in the process industries

Prof. Cesar de Prada
Dpt. Systems Engineering and Automatic Control
University of Valladolid, Spain
prada@autom.uva.es
Outline

✓ Presentation
✓ Measurements and information
✓ Data reconciliation
✓ Gross errors
✓ Examples:
  – Sugar factory
  – Petrol refinery
✓ Conclusions
Today’s process plants

- More technology
- More complex processes
- More instrumentation and systems
- More norms and regulations
- Reduced technical staff
- More data than ever
- Higher market pressures
From data to knowledge

- Huge amount of data available in real time or historians.
- Better instrumentation and new sensors
- With less trained people in the control room or the technical teams, supporting tools are required for process safety, process behaviour predictions, help in Abnormal Situation Management,…
- Models and simulations, decision support systems, etc., are recognized as elements to condense knowledge
- The focus is on software applications at the MES level
Models

- There is a lot of interest in the optimal (economic) operation of the processes
- Models play a key role in supporting the decision making process
- Advanced Control and Economic Optimization are the right tools
- Successful implementation requires suitable models and process information
- Few tools for estimating earnings and improvements
Data / Information

Complex decisions taken at different levels

From data to reliable and coherent information
Plant data

- Some measurements are not consistent or unreliable
- There are many unmeasured variables
- Model parameters need to be estimated
Inconsistencies
Inconsistences
Data reconciliation

✓ Use plant/lab measurements and knowledge stored in the models to:
  – Estimate the values of all variables and model parameters coherent with a process model and as close as possible to the measurements
  – Detect and correct inconsistencies in the measurements

✓ Formulated as an optimization problem
Data reconciliation

\[ \frac{dx}{dt} = f(x, u, \theta) \quad y = h(x, u, \theta) \]

A certain degree of redundancy in the measurements is required.

inputs

MODEL $\theta$

outputs

inputs

Measured inputs $u_e$

PROCESS

Measured outputs $y_e$
Redundancy

Mass balances

\[ F_1 = F_2 + F_3 \]
\[ F_1 X_1 = F_2 X_2 + F_3 X_3 \]

2 equations
6 variables

More than 4 measurements are required to avoid having a unique or multiple solutions

F flow
X composition
Data reconciliation

3 measurements, affected by noise, errors, etc.

Redundant variables

\[ F_1 = F_2 + F_3 \]

Estimated values must satisfy the model

Implicitly, we assume a gaussian distribution in the measurements

Probability of a measured value \( x_i \) around its true value (that verifies the model)

\[
p_i(x_{mi}) = \exp \left[ \frac{-(x_i - x_{mi})^2}{2\sigma^2} \right] \frac{1}{\sigma \sqrt{2\pi}}
\]
Data reconciliation

Criterion (ML): Maximize the probability that the measured value of each variable $x_m$ be equal to the true one, which verifies the model $x$ (to minimize its negative log)

$$\prod_{i=1}^{N} p_i(x_{mi}) = \prod_{t=1}^{N} \exp\left[\frac{-(x_i - x_{mi})^2}{2\sigma^2}\right]$$

Assuming independent variables

$$\min_{x_i, \theta} \left[-\log L(x_i)\right] = \min_{x_i, \theta} \sum_{i=1}^{N} \frac{(x_i - x_{mi})^2}{2\sigma^2} + N \log \sigma \sqrt{2\pi}$$
Data reconciliation

\[ F_1 + F_2 + F_3 = 0 \]

\[ \min \sum_{x} \left[ \frac{F_j - F_{mj}}{\sigma_j} \right]^2 \]

\[ f_i(F) = 0 \]

Compute variance from experimental data

True value

Measured value
Data reconciliation

\[
\min_{u, \theta} \sum_{i=1}^{N_{\text{measured}}} \alpha_i (y_i - y_{m,i})^2 + \beta_i (u_i - u_{m,i})^2
\]

\[
\frac{dx}{dt} = f(x, u, \theta) \quad y = h(x, u, \theta)
\]

\[
g(x, y, u, \theta) \leq 0
\]
Feasibility

\[
\min_{u, \theta, \varepsilon} \sum_{i} \alpha_i \frac{\text{meas}}{\sigma_i^2} (y_i - y_{m,i})^2 + \sum_{j} \beta_j \frac{\text{meas}}{\sigma_j^2} (u_j - u_{m,j})^2 + \sum_{k} \gamma \varepsilon_k^2
\]

\[
\frac{dx}{dt} = f(x, u, \theta) \quad y = h(x, u, \theta)
\]

\[
g(x, y, u, \theta) \leq \varepsilon \quad \varepsilon \geq 0
\]

Normalization: span, variance, instrument precision,…
Feasibility: slack variables incorporated
\(\alpha, \beta\): relative importance of the variables and eliminate variables affected with gross errors
Identificability, regularization,…
Gross errors increase the dispersion and distort the solution. The errors are spread through all variables.
Detecting gross errors

Two approaches:
- Gross errors detection and measurement removal
- Use of robust estimators

Analyze residuals with data without gross errors

Analyze residual of current data
PCA

Test for significant differences and, in particular, for the largest ones and locate the variables that most contribute to them
Gross errors

\[ \min_{u,\theta,\varepsilon} \sum_{i}^{\text{meas}} \frac{\alpha_i}{\sigma_i^2} (y_i - y_{m,i})^2 + \sum_{j}^{\text{meas}} \frac{\beta_j}{\sigma_j^2} (u_j - u_{m,j})^2 + \sum_{k}^{\text{feas}} \gamma_i \varepsilon_k^2 \]

\[ \frac{dx}{dt} = f(x, u, \theta) \quad y = h(x, u, \theta) \]

\[ g(x, y, u, \theta) \leq \varepsilon \quad \varepsilon \geq 0 \]

In practice, gross errors can be detected by a combination of rule base and cyclic solution of the optimization problem. After an initial removal of a set of measurements from the cost function using rules, the solution is checked against the variance of the signal and those variables with measurements outside the $3\sigma$ band, are removed again.
Data Reconciliation- Gross errors

Two approaches:

- Gross errors detection and measurement removal
- Use of robust estimators

![Diagram showing the reconciliation process]

- Inputs $u$
- Model
- Outputs $y$, $u$
- Measurements $y_m$, $u_m$
- Optimization algorithm
- Errors
- Reconciled values $\alpha, \beta$
- Check for gross errors

Cycle until no gross errors are detected
Robust Estimators

If the distribution of the measurement errors $\varepsilon_j$ is non-Gaussian, as may happen if gross errors are present, the LS estimation may give incorrect results as it is not robust against deviations from the assumed Gaussian distribution.

The robustness of a ML-estimator against deviations from non-Gaussianity is measured by the influence function, which is proportional to the first derivative of the estimator. The estimator is robust if the influence function is bounded as the residuals go to infinity.

In particular, the LS estimator is not robust as the derivative

$$\frac{d\varepsilon_j^2}{d\varepsilon_j} = 2\varepsilon_j$$

is not bounded.
Robust estimators

$$F_j = c^2 \left[ \frac{|\varepsilon_j|}{c} - \log\left(1 + \frac{|\varepsilon_j|}{c}\right) \right]$$

Robust estimators use different cost functions, such as the Fair function $F$, that fulfils the robustness property:

$$\frac{dF_j}{d\varepsilon_j} = \frac{\varepsilon_j}{1 + \frac{|\varepsilon_j|}{c}}$$

Robust data reconciliation formulation

$$\min_x \sum_{j \in M} c^2 \left[ \frac{|\varepsilon_j|}{c} - \log\left(1 + \frac{|\varepsilon_j|}{c}\right) \right]$$

$$\varepsilon_j = \frac{x_j - x_{mj}}{\sigma}$$

$$f_i(x) = 0$$
Redescending Función

\[
R_j = \begin{cases} 
0.5 \varepsilon_j^2 & 0 \leq |\varepsilon_j| \leq a \\
|\varepsilon_j| - 0.5a^2 & a \leq |\varepsilon_j| \leq b \\
ab - 0.5a^2 + 0.5a(c-b)(1 - \left(\frac{c-|\varepsilon_j|}{c-b}\right)^2) & b \leq |\varepsilon_j| \leq c \\
ab - 0.5a^2 + 0.5a(c-b) & c \leq |\varepsilon_j| 
\end{cases}
\]

Hampel’s redescending estimator

\[c > b + 2a\]
Welsch

\[ W_j = \frac{c^2}{2} \left[ 1 - \exp \left( -\left( \frac{\varepsilon_j}{c} \right)^2 \right) \right] \]

95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant \( c = 2.9846 \)
Data reconciliation

- Static
  - Steady state detector
  - Data averaged over a period of time

- Dynamic
  - Dynamic optimization problem

- Batch
  - Open field
Beet sugar factory

Washing & Cutting → cossettes → Difussion → Raw Juice → Depuration → Depurated Juice → Evaporation → Syrup → Sugar house → sugar → Sugar Drying

Drying

Boilers

Fuel/Gas

Steam

Vapor

Water

Lime

Lodos

CATTLE FOOD

Process Data: Non measured variables, efficiency, steam consumption

Information

KPIs

Plant

Mathematical treatment
Sugar plant DR software

✓ Main elements:

✓ Periodic characterization of the plant status, using a steady model of the sugar plant.

✓ On line connection with the plan Distributed Control System (DCS) to obtain, the measured variables necessary for the balances and model identification.

✓ Data reconciliation, correcting measured variables in a way that the model is adjusted and calculating at the same time that unknown variables and model parameters.

✓ As a by-product of the data reconciliation, key performance indicators are estimated from calculated values in the reconciliation.
Models

- Static
- Mass energy balances
- Flows, pressures
- Equations and properties of the application domain
- Formulated in the EcosimPro environment
- Measurements averaged for a period of time
- Rules to eliminate bad measurements
Data reconciliation

\[
\min_{u, \theta, \varepsilon} \sum_{i}^{\text{meas}} \frac{\alpha_i}{\sigma_i^2} (y_i - y_{m,i})^2 + \sum_{j}^{\text{meas}} \frac{\beta_j}{\sigma_j^2} (u_j - u_{m,j})^2 + \sum_{k}^{\text{feas}} \gamma_i \varepsilon_k^2
\]

\[y = h(x, u, \theta)\quad g(x, y, u, \theta) \leq \varepsilon\quad \varepsilon \geq 0\]

Solved with a sequential approach

Optimizer

EcosimPro Simulation

gives \(J(u, x(t)), g\)
Implementation in EcosimPro

- EcosimPro dynamic/static model
- Optimization assistant
- Open loop optimization
- C++ class
- External application
- Simulated process
- Results in simulation

 dll + OPC
Real time application
SCADA implementation
DR system / SCADA

CONSUMO VAPOR FÁBRICA - RENDIMIENTO ENERGÉTICO

RESUMEN VAPOR

CAMBIADORES

PRODUCTO

PRINCIPAL
DR system SCADA
Information

- Detection of inconsistent measures. Help in fault detection.
- KPI: Evaluation of energetic behaviour indexes, efficiency, comparison between process heat transfer coefficients versus theoretical coefficients.
- Estimation of all unmeasured variables, some of them relevant for the energy evaluation such as steam consumptions.
- Keeping track of the time evolution of key variables during the sugar beet campaign, helping managers in locating malfunctions in the process or equipment fouling and planning maintenance.
Inconsistencies
Key Performance Indicators KPI

Heat exchanger coefficients

Boiler efficiency
Hydrogen network
Arquitecture

HMI (Excel)

Data treatment

GAMS

User

SCADA

Data and results
Data treatment

key role played by the data treatment in the success of the application in the refinery. If data from the SCADA system are not analyzed and filter previously to their use in the numerical methods, there are no chances to obtain good results. This layer is composed of a set of rules that detect faults and information inconsistencies in the raw data and decides which options are the most adequate ones. For instance, detecting when a flow is actually zero, a plant is stopped, a measurement is out of range, etc. It has been developed for specific cases combining physical knowledge and heuristic rules.

As a result of these rules, the system adjust the model parameters and optimization weights, so that, e.g. a measurement can be eliminated from the data reconciliation cost function. Mayor changes take place when a plant is not operating. To deal with these cases, the network is formulated as a superstructure that allows to remove groups of equations depending on the value of binary variables that represent the state of the plants.
DR

Pureza reciclo

Pureza BP

Flujo reciclo

Flujo BP

Aporte CBP

Conc. S salida

18 days
Conclusions

✓ Data reconciliation is a model based approach to obtain coherent information from the plant.
✓ It allows to compute KPI to follow the time evolution of the process operation.
✓ Formulated as an optimization problem.
✓ Open problems:
  – Gross error detection
  – Speed, batch, non-independent variables,…