Chapter 2

Bibliographic Review

De omni re scibili, et quibusdam aliis

2.1 Introduction

The main objective of RTO systems is to operate the plant at every instant as near to the optimum operating conditions as possible. In order to achieve this objective, the RTO system contains basically the components described in figure 2.1.

The RTO loop can be considered as an extension of a feedback control system. It consists of subsystems for measurement validation, steady state detection, process model updating, model-based optimisation and command conditioning. Once the plant operation has reached steady state, plant data are collected and validated to avoid gross error in the process measurements, while the measurements themselves may be reconciled using material and energy balances to insure consistency of the data set used for model updating. Once validated, the measurements are used to estimate the model parameters to insure that the model correctly represents the plant at the current operating point. Then, the optimum controller set-points are calculated using the updated model, and are transferred to the control system after a check by the command conditioning subsystem.

Some RTO implementations may perform the figure 2.1 loop with a given frequency instead of performing steady state detection (Darby and White, 1988). In such a case, the period between two consecutive optimisations must be longer than the settling time of the process, to insure that units have returned to steady state operation. For illustration purposes, the settling time for an ethylene plant is about four hours, and for a sulphuric acid contact process is near to twelve hours (Darby and White, 1988). Consequently, the set-points for the process operating conditions may be optimised daily or hourly according to the time scale of the process and the economic incentives to make changes.

Qualitative RTO behaviour is illustrated in figure 2.2. From time t_{ini} to time t_{end} some disturbance occurs in the process (remain that the term disturbance in this thesis refers to an un-

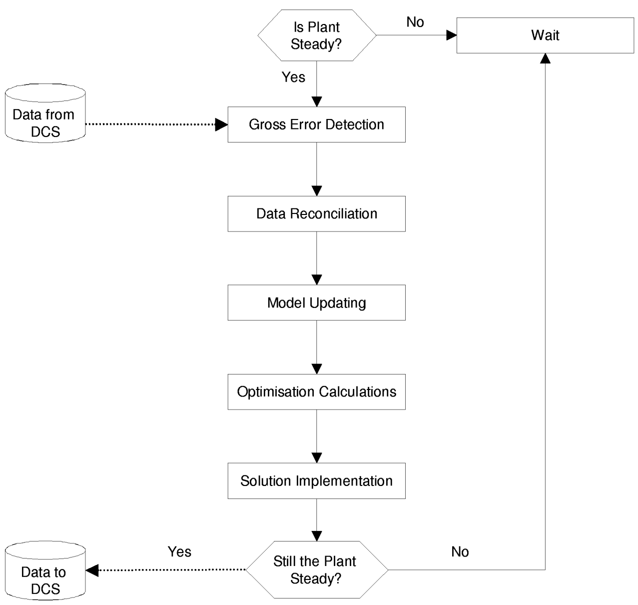

Figure 2.1: Schematic representation of the closed loop RTO system functionality (adapted from Georgiou et al. (1998))

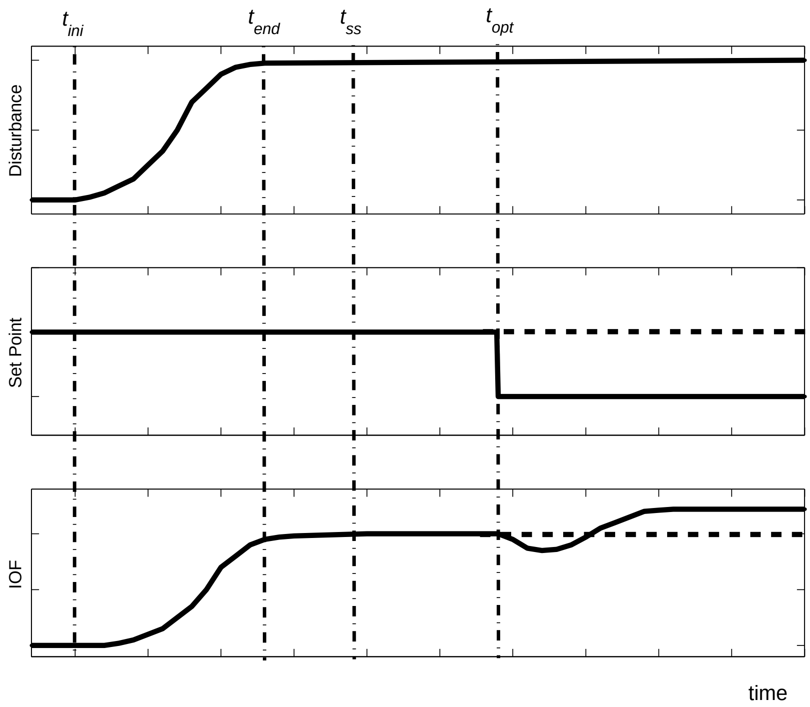

Figure 2.2: Qualitative results of applying an RTO system (solid lines) compared with no optimising system (dashed lines)

controlled variable, which affects the process economy, for instance, an improvement in feed composition). As a consequence, when no adjustment is made on the set-points, a change of the Instantaneous Objective Function (IOF, i.e. profit) value that one wants to optimise simultaneously takes place (dashed lines in 2.2). However, the RTO system reaction to disturbance only arrives at time t_{ss}, when the new steady state is detected and the optimisation procedure is started. When the optimisation is completed at time t_{opt}, the new set-points are implemented resulting in a better IOF after some stabilisation time.

Since its first implementation (Kuehn and Davidson, 1961; Crowther et al., 1961; Pendleton, 1961) RTO systems have included to some extension the functional elements indicated in figure 2.1. Following sections will summarise the progress on every functionality.

2.2 Steady state identification (SSID)

As indicated in the flowchart of the figure 2.1 the RTO loop begin with the steady state detection. Identifying steady state may be difficult because process variables are noisy and measurements do not settle at one value. One defines steady state to mean an acceptable constancy of the mean values of measurements over a given period of time. Therefore, tests for steady state are commonly based on the constancy of these mean values.

In probably the first method (Crow et al., 1955) a *F-test* was used, a ratio of variances as measured on the same set of data by two different methods. Data in the most recent window are averaged and the variance is first conventionally calculated as the mean-square-deviation from the average. The variance is alternately calculated from the sequential differences between data. If the process is at steady state, the two methods produce unbiased process variance estimates and their ratio is unity. However, due to limited sampling, the actual ratio of the variances will not be exactly unity but a statistic that averages near unity. Performing such kind of technique requires considerable data storage and computational effort (with respect to DCS capacity), as well as user expertise in choosing the window length.

The time series horizontal screening method has been also used in industry to detect the steady state. In this method, the measured values for key process variables are observed for a time period. If the measured values remain in a stable range with tolerant random noises, then the process is said operating at steady state. Another similar approach consists in performing a linear regression over a data window, and uses a *t-test* on the regression slope. If the slope is significantly different from zero, the process is almost certainly not at steady state. This method presents similar drawbacks to the prior one.

Narasimhan et al. (1986) presented a two-stage composite statistical test to detect departures from steady state. Random errors with zero means were assumed, and two successive time periods assumed at steady state are examined. So that the first test establishes whether the unknown covariance matrices were equal, and the second tests whether the means of the two periods were equals (using the Hotelling's T^2 test). In other work, Narasimhan et al. (1987) applied the theory of evidence to the detection of changes in steady states. Other simple methods, like band test and filters based test are described in Contino (1987).

More recently, Loar (1994) proposed an approach that uses a Statistical Process Control (SPC) moving average chart. In another method Alekman (1994) compares the average calculated from a recent history to a “standard” based on an earlier history and then apply the *t-statistic*. An improved possibility consists in calculating the standard deviation of the process measurements over a moving window and compare it to a threshold value (Jubien and Bihary, 1994). Again, storage and operation on the data window are a computational burden.

Finally, Cao and Rhinehart (1995), developed a method based, as that of Crow et al. (1955), in a primitive *F-test* type of statistic. To reduce computational effort exponentially weighted moving average and variances are calculated in place of the conventional average or variance. In their method, a filtered value provides an estimate of the data mean replacing the average. The only four equations for this method require no logic, low storage and low computational operation calculations. An *R* coefficient, and its critical values (Cao and Rhinehart, 1997) allows a very effective way to perform SSID. Brown and Rhinerhart (2000) extended such procedure to the multivariable case.

2.3 Data improvement

Measured process data inherently contain inaccurate information because the measurements are obtained with imperfect techniques. When imperfect information is used for estimation of process variables and process control, the state of the system is be misrepresented and the resulting control performance is poor, leading to sub-optimal and even unsafe process operation.

In order to obtain a representative value of a process variable using past measurements, filtering techniques are applied. There are two main approaches for filtering purposes: statistic-based and model based. The former approach rely on the statistical analysis of each process measurement independently of the other process measurements. Statistical filtering are used in the pre-processing task as a data conditioned process. Indeed, some standard filter (such as Exponential Smoothing and Moving Average) are already incorporated in many Distributed Control System (DCS). Furthermore in the pre-processing procedure the outlier measurements are eliminated by setting permissible lower and upper bound of process variables. The second approach, make use of the available knowledge about the process in the form of a model, with the aim of not only filter the signals but also make them consistent. Data reconciliation belongs to this latter approach.

The relationship between a measurement of a variable and its true value can be represented by:

$$y = x + e \tag{2.1}$$

where:

y: vector of measured values.

x: vector of true values.

e: vector of error values.

Being the errors (e) typically classified into the following types:

- *Random errors*: they are zero-mean and are often assumed to be normally distributed (Gaussian), since their are the result of the simultaneous effect of several causes.

- *Non-random errors*: they are usually caused by large, short-term, non-random events. They can be subdivided into measurement-related errors, such as malfunctioning sensors, and process-related errors, such as process leaks. *Gross errors* occur when measurement devices provide consistently erroneous values, either high or low. In this case, the expected value of e is not zero. This behaviour may be owed for instance to incorrect calibration of measurements device, sensor degradation and damage to the electronics.

With the aim of identify and eliminate both type of errors, process data are improved using spatial, or functional, redundancies in the pair measurements-process model. Data are spatially redundant if they exceed the minimum data required to completely define the process model at any instant. In short, the system is over-determined and requires a solution by least squares fitting. Similarly, data can be improved using temporal redundancies. Measurements are temporally redundant if past measurements values are available and can be used for estimation purposes. Dynamic models composed of algebraic and differential equations provide both, spatial and temporal redundancy.

A simplified view of data validation techniques can be divided into three basic steps, as shown in figure 2.3. The first step, variable classification (figure 2.4) involves determining which variables are observable or unobservable and which are redundant or under-determined. Several authors have published algorithms for variable classification (Stanley and Mah, 1981; Kretsovalis and Mah, 1988a,b; Crowe, 1989; Mah, 1990; Romagnoli and Sanchez, 1999). Those that are undeterminable are not available for improvement. Next, all gross errors are identified and removed, and after that data reconciliation concentrates on removing the remaining small, random measurement errors from the data.

A key assumption frequently made during the reconciliation step is that the errors are normally distributed, but gross errors severely violate that assumption. If a measurement containing a gross error was allowed into the reconciliation scheme, the resulting estimates of the values of the variable would contain a portion of the gross error distributed among some or perhaps all the estimates (“smearing”). In practice, gross error detection and elimination are usually performed iteratively along with the final step, data reconciliation. Following sub-sections summarise the main advances in these areas.

2.3.1 Data reconciliation (DR)

Data reconciliation is the procedure of optimally adjusting measured data so that the adjusted values obey the conservation laws and other constraints. Unfortunately, in presence of biases, all

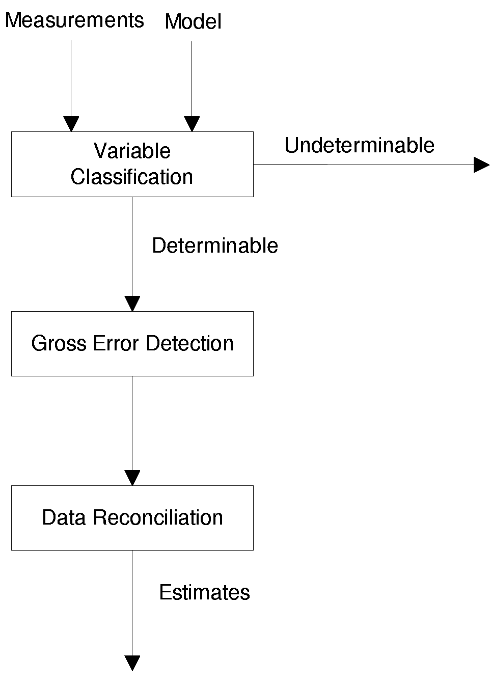

Figure 2.3: Steps for data validation

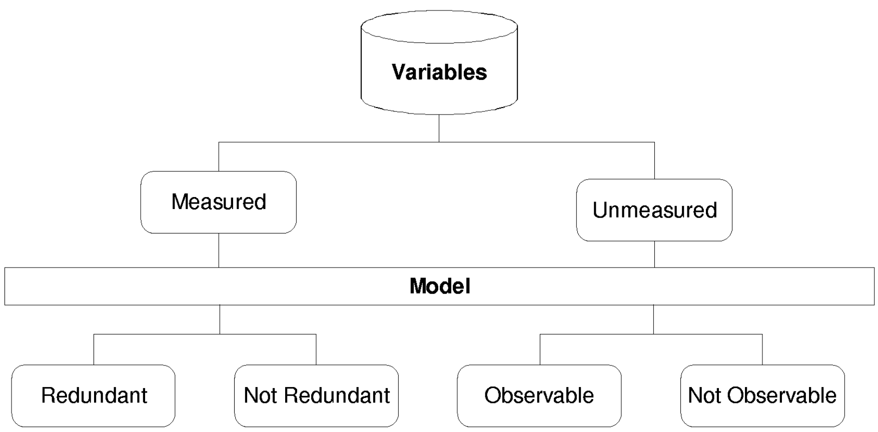

Figure 2.4: Variable classification

the adjustment are greatly affected, by these "gross errors" and would not in general be reliable indicators of the state of the process.

In Chemical Engineering, Kuehn and Davidson (1961) were the first to publish an analysis of data reconciliation. They study both, the linear (mass flows) and bilinear case (enthalpy). The problem of steady state data reconciliation can be defined in the following way (considering for the sake of simplicity the linear case of material flows reconciliation):[1]

$$min\ f = \frac{1}{2}(x-y)^T Q^{-1}(x-y) \tag{2.2}$$

subject to:

$$Ax - c = 0 \tag{2.3}$$

where:

f:	the standard quadratic error objective function.
Q:	variance-covariance matrix (usually diagonal).
y:	vector of the measurements.
x:	vector of the reconciled estimates.
A:	matrix of linear constraints (the mode).
c:	vector of right-hand side terms in linear constraints.

The solution for such problem can be analytically obtained:

$$x^* = \left[I - QA^T(AQA)^{-1}A\right]y + QA^T\left(AQA^T\right)^{-1}c \tag{2.4}$$

In other words, the idea is to correct the measurements in order to make them consistent with a good model, commonly the conservation law. The term of variance-covariance matrix in the objective function allows greater adjustment over the measurement with more variance. In the general case, where there are some measured concentrations in streams whose total flows are not measured, the constraints (mass balances, equation 2.3) become bilinear, and thus care should be taken during the optimisation, in order to guaranty global optimality (or apply specialised procedures).

Another early application of data reconciliation to an industrial process was also reported by Reilly and Carpani (1963), who formulated the collective *chi-square test* of all the data and the univariate test for constraints, based on the normal distribution. Václavek (1969) presented the first comprehensive analysis of data reconciliation. He also first addressed the question of how to

[1] In the general sense, and according to the Bayesian theory, the data reconciliation problem can be stated as a maximum likelihood problem where the probability of the plant process variables of taken the obtained value is maximized, given the process measurements. Indeed, the least squares case is just the particular case corresponding to normally distributed errors.

eliminate unmeasured quantities by combining units connected by an unmeasured stream and by deleting units which had an unmeasured feed or product stream incident to them (the Reduced Balance Scheme).

Crowe et al. (1983) proposed a projection matrix technique to decompose the data reconciliation problem that has linear constraints and unmeasured variables into the solution of two sub-problems. First, the unmeasured variables in constraints are removed by multiplying a matrix (projection matrix) and the variables in constraints are all measured, then the solution of this sub-problem is obtained. Then the solution of the unmeasured variables can be determined through the original constraints and the reconciled values of the measured variables. Crowe (1986) extended the projection matrix technique to the case of non-liner constraints using an iterative algorithm with successive linearisations.

Several researchers (Tjoa and Biegler, 1992; Robertson et al., 1996) have claimed the advantages of using non-linear programming techniques over such traditional data reconciliation methods, as successive linearisation for steady state or dynamic processes. Through the inclusion of variable bounds and a more robust treatment of non-linear algebraic constraints, improved reconciliation performance should be realised. Extended discussions about such aspects can be found in Mah (1990); Madron (1992); Romagnoli and Sanchez (1999) and Crowe (1996).

Extended Kalman filtering has been a popular method used in the literature to solve the dynamic data reconciliation problem (Muske and Edgar, 1998). As an alternative, the non-linear dynamic data reconciliation problem with a weighted least squares objective function can be expressed as a moving horizon problem (Liebman et al., 1992) similar to that used for model predictive control. The non-linear objective function (usually quadratic) is:

$$\begin{array}{ll} min & f\left[y(t),x(t)\right] \\ x(t) & \end{array} \tag{2.5}$$

which is subject to the dynamic model:

$$h\left[\frac{dx(t)}{dt},x(t)\right]=0 \tag{2.6}$$

and inequality constraints:

$$g\left[x(t)\right]\leq 0 \tag{2.7}$$

This problem can be solved using a combined optimization and constraint model solution strategy (Muske and Edgar, 1998) by converting the differential equations to algebraic constraints using orthogonal collocation, some other model discretisation approach (Albuquerque and Biegler, 1995) or artificial variables. For instance, Bagajewicz and Jiang (1997) used a control vector parametrisation approach, where the variables profiles are approximated by a polynomial of a given order.

2.3.2 Gross error detection (GED)[2]

Ideally, measurements ought to produce residuals of the balances that are randomly distributed with expected values of zero. However, gross errors resulting from different sources like instrument malfunctions, biases in off-line analysers or unrepresentative sampling will adversely affect the performance of data reconciliation. Therefore and ideally, the challenge tasks around gross errors are:

- Detect their existence.
- Identify their location.
- Identify their type.
- Determine their size.

And then, the gross errors should be either corrected or the measurement eliminated.

The trivial way to detect gross errors is by monitoring the adjustment values resulting from successive data reconciliation. The adjustment values must follow a normal distribution. Therefore, a classical SPC monitoring should detect deviation from normal distribution behaviour and therefore the presence of a non-random error. Since a non-desirable delay is inherent to such strategy, more elaborated approaches have been proposed. They provide the most effective way to detect the gross errors in plant data, and are based on the statistical principles of hypothesis testing.

For illustration proposes, consider a linear system, with gross errors caused by sensor biases. The proposed model for the measurements is:

$$\begin{aligned} y &= x + e \\ e &= e_r + e_g \cdot \delta_i \end{aligned} \tag{2.8}$$

where x, y and e have the same meaning than in equation 2.1. The error involves the random one (e_r) and the gross error. The variable e_g is an unknown, being δ_i a unit vector (one in position i, and 0 elsewhere). Besides the steady state process model is used (linearised according to equation 2.3). Then, two auxiliary variables defined according to the following equations arise:

$$r_c = A \cdot y - c \tag{2.9}$$

$$r_m = y - x = e \tag{2.10}$$

where:

r_c: the constraint residuals vector.

[2] The results for gross error detection/estimation have been reviewed in detail by Mah (1990); Crowe (1996) and Bagajewicz (2000), between others.

r_m: the measurements residual vector.

It is typically assumed that there are not unmeasured variables since they can be eliminated using the Crowe's projection matrix. Then, several statistical test have been constructed for the detection of gross errors. Some of them are based in the distribution of constraint residuals r_c, and others are based in the distribution of the measurement adjustments r_m. Note that the evaluation of r_c does not require solving previously the associated data reconciliation problem. Most of them are of limited applicability because their assumptions (linear constraints, many measurements must be available, residuals normally distributed, etc.). However, the main contributions are listed in the following:

- *The Global or Collective Test* (Reilly and Carpani, 1963; Ripps, 1965): based on the fact that the objective function of the data reconciliation problem, at the minimum is distributed as a chi-square variable if the measurements are normally distributed about their true values. The magnitude of the data reconciliation objective function is then compared to the tabulated chi-square value for a chosen confidence level (e.g. 95%) and for degrees of freedom equal to the number of remaining balances. Almasy and Sztano (1975) examined the case when the variances of the measurements are unknown and derived a collective test (*Fisher F-test*).

- *The Constraint and Nodal Test* (Reilly and Carpani, 1963; Mah et al., 1976): based on the constraints residual values divided over the standard deviation of the residual.

- *The Measurement Test* (Mah and Tamhane, 1982): based on the measurement adjustment divided by its standard deviation.

- *The Maximum Power Test* (MP, Almasy and Sztano (1975)): based on the normal distribution of the measurements which has the greatest probability of correctly detecting a single gross error in the measurement, when only one is present.

- *The Generalised Likelihood Ratio Test* (GLR, originally Willsky and Jones (1976) and then Narasimhan and Mah (1987)): which is equivalent to the MP test because the GLR test value is the square of the MP test value, and has the capability of directly identifying the gross error location and distinguishing between different types of gross errors such as leaks and biases.

- *The UnBiased Estimation Technique* (UBET, Rollins and Davis (1992, 1993)). This method is restricted to normally distributed errors, steady state and linear constraints. Firstly, a global test is conducted, and then UBET is used to detect the number and location of gross errors by trial and error, using two test statistics (*F* and *Bonferoni tests*). The approach has been extended to the bilinear case.

- *The Akaike's Information Criterion* (AIC, Yamamura et al. (1988)), a function of the likelihood function and the number of errors. This criterion divides the measured variables in two types: with only random errors and those with also gross errors. The gross errors are identified by comparing the values of AIC function for all possible combinations.

- *The Principal Component Analysis Test* (PCA, Tong and Crowe (1997)). PCA is an effective tool for multivariate data analysis. In this technique, a set of correlated variables is transformed into a new set of uncorrelated variables, the Principal Components. The nodal test can be applied over a new vector, showing sharper results.

For multiple identification of gross errors, two strategies are available: the serial elimination and the serial compensation. In the first one, in every iteration the measurement identified as containing a gross error is deleted, and the GED test is applied to the remaining measurements (see Ripps (1965); Nogita (1972)). In the other approach, the measurement identified as containing gross error is corrected using an estimation and the test is subsequently applied to the compensated set of measurements in order to detect more gross errors (Narasimhan and Mah, 1987).

It is important to note that no test for gross error has a guarantee of consistently finding all of them. However, gross errors are generally persistent so that repeated data reconciliation of plant data would show a non-random pattern of a test statistic.

Available commercial software for data reconciliation will typically reject a variable value containing a gross error (identified by any of the methods previously mentioned). As explained before, this leaves the system with a smaller degree of redundancy and the precision of the reconciled variables deteriorates. Unfortunately, the correction of a value with gross error requires a proper location of such error, what is impossible under certain circumstances (Bagajewicz, 2000).

2.4 Modelling

As important as the data quality are the models quality. Depending on the need, the process and economic models can be very elaborated or a simple value-added equation. A precise plant model is necessary to simulate the process for on-line optimisation. It serves as constraints for data validation and parameter estimation to relate individual measurement together for error rectification and for economic optimisation to determine the best operation conditions of the plant. Chemical processes can be simulated by an open form equation based model or a closed form sequential modular model. The open form model has the advantage of computation speed. The close form model can be easily developed and modified using flowsheeting programs, and usually allows a closer representation of the process. Therefore, the choice between one or another, must be done on the basis of the process and optimisation complexity. Another possible but less frequently used approach consists in using artificial intelligence techniques, such as Artificial Neural Networks (ANN).

If the model became too complex for performing on-line optimisation, there are mainly two alternatives:

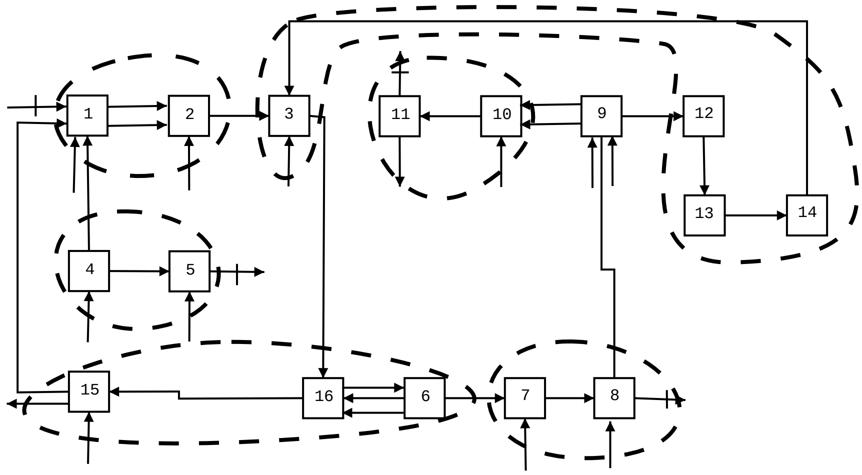

Figure 2.5: Model simplification partitioning

- To split the optimisation problem hierarchically, and sequentially, thus developing and implementing individual on-line optimisation systems to smaller plant sections.

- To use simplified models and update them periodically according to an upper level detailed model (Riggs and Hetzel, 2000).

For illustration purposes, figure 2.5 shows a sub-division of a macrosystem into smaller subsystems, using typical techniques of tearing and partitioning (e.g. Rudd and Watson (1968)).

It should be mentioned, however, that it appears that the development of simulation software will soon provide a convenient graphical user interface environment similar to sequential modular simulation for developing open form equation based models. Furthermore, hybrid forms of models are becoming a possible alternative (CAPE Open standards, Braunschweig et al. (1999)). Besides, the advances in hardware and parallel computing techniques will likely favour the use of every day more complete and reliable models.

Both, economic and process models require a deep understanding of the system in order to successfully implement and achieve the benefits of on-line optimisation. Special efforts must be focussed in these issues at the design stage of an on-line optimisation system, including collecting a huge quantity of information from the plant. In any case, despite the off-line modelling aspects, some research results are available to address also an on-line improvement of the process model what is summarised next.

2.4.1 Model updating

Selected model parameters are updated using plant data to reduce plat-model mismatch. In a RTO application, the models can be highly non-linear and the updated parameters appear implicitly in the models. For an unconstrained problem, the model can be expressed in terms of the process variables, x, fixed parameters (not updated), p_α, and updated parameters, p_β given in the following equation:

$$f(x, p_\alpha, p_\beta) = 0 \tag{2.11}$$

The vector of parameters, p_β, can be estimated from a set of measurements, y, using an appropriate formulation of the parameter estimation problem. The effect of measurement noise can be reduced by a low pass filter that is typically an averaging filter. Indeed, data reconciliation reduce significantly the data variability. Thus, filtered measurements are used for parameter estimation (section 2.3, page 23).

There are two types of models for parameter estimation according to Britt and Luecke (1973). One type is the explicit model, in which measurements are divided into two sets of measured variables, independent variables and dependent variables. In this type of model, independent variables are measured with a much greater accuracy than dependent variables. The dependent variables can be expressed as an explicit function of independent variables and the parameters. For this type of model, parameters can be estimated by minimising the sum of squared errors of dependent variables (least squares method) or maximising the likelihood function, a probability distribution function of the measurement errors of dependent variables (maximum likelihood method). This is an unconstrained optimisation problem, and linear regression method is one of examples for this type of estimation.

The other type of model is implicit or error-in-variables model (e.g. Kim et al. (1990)). There are errors in all measurements and the variables cannot be partitioned into dependent and independent variables as in the explicit model. The constraints of process models are implicit. Therefore, the optimisation problem of parameter estimation must be formulated as constrained optimisation problem.

An often used formulation in parameter estimation is the simultaneous data reconciliation and parameter estimation (Britt and Luecke, 1973; MacDonal and Howat, 1988; Kim et al., 1990, 1991; Chen, 1998; Chen et al., 1998), and its application in industrial RTO systems has been reported (Hardin et al., 1995). In this formulation, the vector of variables, x, are partitioned into measured and unmeasured variables, x_m and x_u, respectively, and the difference between each measured variable and the sensor value is signified by a variable (e), with:

$$y = x_m + e \tag{2.12}$$

The parameters, p_β, measurement adjustments, e, and unmeasured variables, x_u, are evaluated so that the objective function is minimised and the equality constraints in equation 2.11 are satisfied exactly. The usual objective function is the sum of squared adjustments weighted by the covariance matrix of the measurements, Q. So that the updating formulation is stated as follows:

$$\min_{p_\beta, e, x_u} \quad e^T Q^{-1} e \tag{2.13}$$

subject to:

$$f(y-e, x_u, p_\alpha, p_\beta) = 0 \tag{2.14}$$

Some efforts have been also devoted to the selection of parameters to be updated on-line, p_β. The straightforward approach, which basically consists in a sensitivity study, has been proposed by Krishnan et al. (1992). In this way, a sub-set of the parameters in the model is selected on the basis that there is anticipated to be significant variability or uncertainty in their values, and that changes in their values cause significant variations in the optimal operating conditions determined from the model. In this way, only those parameters that significantly affect the optimum are used for on-line updating.

On the other hand, Forbes et al. (1994) proposed a systematic way to evaluate the models adequacy requirements for their use in on-line optimisation, which follows the inverse reasoning of the prior proposal. Although the mathematical explanation is cumbersome, their conclusion is actually quite simple: a model is adequate for RTO if is able to give correct optimum values rather than giving accurate outputs for any condition (similarly to inner-outer algorithms, e.g. Biegler et al. (1985)). The associated difficulty is that the model must be able to be “adequate” in that sense, for all possible scenarios. Consequently, one drawback of such approach is that a reference model (one that is good in the conventional sense) is required in order to perform such adequacy test, which in case of being available could be directly used for on-line optimisation. In case that the model is too complex to perform on-line calculations, the proposed analysis became too complex as well for applying the suggested procedure (with the exception that can be done in the design stage).

In order to remove the effect of plant-model mismatch, it has been also proposed to integrate the parameter estimation and optimisation, rather than using the sequential approach favoured in most applications (Haimes and Wismer, 1972; Roberts, 1979; Roberts and Williams, 1981; Cheng and Zafiriou, 2000). The idea is similar to a simplified stochastic problem formulation, however, the approach requires generating on-line sensitivity studies over the plant, which is hardly acceptable.

Recently, Yip and Marlin (2002), have shown the advantages of using multiple data sets for updating the stead state plant model and to use prior knowledge to categorise the parameters as fast or slow changing. In a previous work the same authors (Yip and Marlin, 2000) propose the use of on-line experiments (using design of experiments techniques) over the plant for improve the quality of the model updating results.

2.5 Optimisation

Once that data and models are reliable, there is an optimisation problem to solve. Most of on-line optimisation problems are non-linear programs (Pibouleau et al., 1999). The three widely used optimisation (*indirect*) algorithms for solving non-linear programming problems are Successive

Linear Programming (SLP), Successive Quadratic Programming (SQP) and the Generalised Reduced Gradient (GRG) methods (for more information see e.g. Edgar et al. (2001)).

- *Successive Linear Programming* linearlises the objective function and constraints around a feasible starting point and solves a sequence of linear programming problems to arrive at a local optimum.

- *Successive Quadratic Programming* uses a quadratic approximation to the objective function and a linear approximation to the constraints and solves a sequence of quadratic programming problems to arrive at a local optimum. Quadratic programming uses the Kuhn-Tucker conditions to convert the quadratic programming problem to a set of linear equations which can be solved by linear programming. Thus, successive quadratic programming implicitly solves a sequence of linear programming problems as well. To avoid evaluating the Hessian matrix of second partial derivatives of the objective function, a quasi-Newton update formula such as BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) is used which only requires gradient values.

- *The Generalised Reduced Gradient* also linearlises the objective function and constraint equations about a starting point, and it manipulates these equations to form a reduced gradient line to provide a direction to perform a series of line searches to arrive at a local optimum.

All of the methods use the same information, values of the first partial derivatives of the objective function and constraints, but each one uses this information in a different way. It should be noted, in addition, that industrial implementations may use also *direct search procedures*, this is, methods which do not use information about derivatives (i.e. the Simplex-Amoeba method, the Complex Box, and so on). Such kind of methods are based on just objective function and constraint evaluations, and not on derivatives. Besides, there is a variety of improvement strategies using geometric forms. Although they are significantly slower than the *indirect* methods, they are robust and reliable, thus suitable for on-line use, in addition to the fact that they can be more easily coupled with modular-sequential process simulators.

2.6 Results analysis and uncertainty aspects

After performing optimisation, the optimal values of the decision variables are implemented as set-points to the control system. Although the operator may have the chance of review the proposed set-points, the rate of change of their values is normally bounded because operational and safety reasons (they must not exceed the capabilities of the control system). But despite so, often the optimum for a process operation is largely or fully constrained by inequality limitations of various kinds. Thus, the set-points returned by the on-line optimisation system will likely lead to plant operation on these constraints. What if the parameter estimates used in the model are

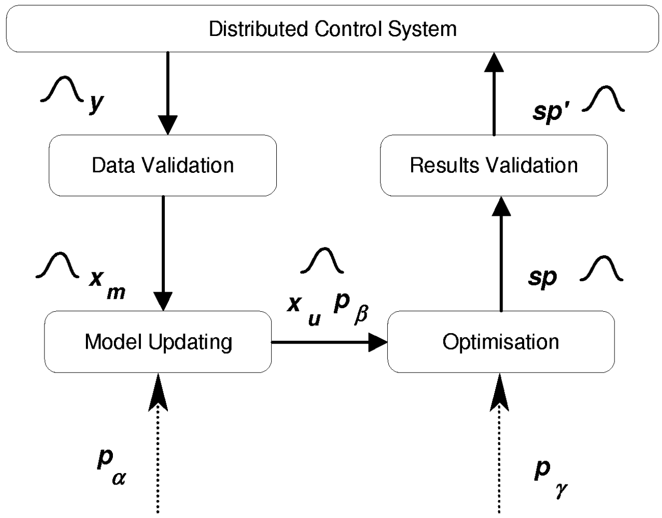

Figure 2.6: Variability and uncertainty propagation through the RTO functional blocks

not accurate?... Such fact has motivated the study of very related aspects in the on-line optimisation framework: the uncertainty and variability propagation, which are the subject of following subsections.

2.6.1 Uncertainty

As mentioned in section 1.4.2 (page 12), some on-line optimisation systems may have poor long-term service factor due, at least partially, to inadequate optimisation robustness. Although most RTO systems attempt to improve model accuracy through model updating, that is to say, update some of the model parameters, these ones, as any information flowing into the loop (figure 2.6), have associated some degree of uncertainty:

- *Measurements*: Every sensor provides a value (y) with some degree of precision.
- *Parameters*: Model parameters, both, the fixed ones (p_α) and the updated online (p_β), will carry also an associated degree of precision. These model parameters include typically:
  - *Economic parameters*: prices, costs (p_γ).
  - *Process Model parameters*: catalyst activity, fouling coefficients, equipment's efficiencies and so on (p_π).

  – *Process variables*: that are taken as fixed input in the process model (usually, not measured variables, with low variability, i.e vessel pressures) and that commonly do not lead to a significant loss of accuracy.

As a consequence of their simultaneous influence and the uncertainty propagation in the loop, the optimisation results when using mean/estimated values for the previous sources of uncertainty have as well associated some degree of "confidence", which will be function of the mentioned uncertainty sources. This fact may degrade the process economical performance (e.g. it is not the same to optimise using averaged input values, that obtaining the optimum which offers the best expected value for the uncertain input). Additionally, and what is more important, it may lead to the violation of some operational constraint.

Regarding the constraints satisfaction, it should be noted however, that some type of constraints are related to averaged conditions rather than instantaneous ones, and their punctual violation does not lead to significant problems (e.g. the composition of an output stream that is fed to a storage tank). Although there are control functions devoted to that, special care must be taken with a second type of constraints related to punctual conditions that are usually related to safety conditions.

De Hennin et al. (1994) have shown how to back off from the constraints to allow for three effects, and still keep the operation feasible:

- slow disturbance in the constraints, which would make the implemented optimum infeasible before the next update;

- errors and uncertainties in parameter estimates;

- errors in unestimated parameters.

However, in a more general sense, the problem of dealing with uncertainty has been addressed by using one of these typical ways:

- *Solution sensitivity and stability analysis:* It can be measured the impact of changes in the problem uncertain data on the optimal solution. Furthermore, the objective function may include a term penalising such sensitivity (Backer et al., 1994).

- *Stochastic optimisation*: It explicitly deals with the uncertainty in the data, using recourse functions or chance constraints. Commonly, the original stochastic problem is transformed into a new one, deterministic, which somehow is able to deal with the uncertainty by for instance using the expected value for the objective function. For realistic cases with a few decision variables, and some uncertain parameters, simplification procedures are likely to be required (i. e, delayed sampling approach, as in Hou et al. (2000, 2001); or the linearisations and bias updating used by Zhang et al. (2002)).

2.6.2 Results analysis

As illustrated in figure 2.6 the output of an optimiser may be affected by the measurement noise (high frequency variability), producing variability in the predicted optimal set-points by the optimisation system under fairly constant plant conditions.

Although the original guidelines are given in Box and Draper (1969), and is formally proposed for RTO systems in Koninckx (1988), the problem has been recently addressed by Marlin and Forbes. Many of their works deal with this subject, being the more representatives those of Miletic and Marlin (1998) and Zhang et al. (2001). The idea is quite simple: to study the variability propagation trough the information flow using steady state sensitivity and then apply a significance test to evaluate if the solution (set-points) has changed because of the random variability or not. This test is then used to accept or reject the proposed set-points changes. In the work of Miletic and Marlin (1996), the analysis is done for the unconstrained optimisation case (in the reduced space of independent variables), applying the Hotelling's T^2 significance test over such decision variables respect to the current conditions. The work is later extended by Zhang et al. (2001) to consider also the inequality-constrained case, applying in a second stage the T^2 significance test over the dual variables associated to the inequality constraints.

However, it is important regarding this issue that appears more convenient (practically and also conceptually), to apply such kind of tests before the optimisation rather than after. Besides, such study is performed over a number of optimisation runs using a steady state model to compare their proposal against the "always change" policy using a very high noise in the measured variables while the system dynamics have been completely ignored. Moreover, a key aspect remains obscure: although the moves may degrade the short-term profit, the plant response is providing significant information for model updating purposes, and thus for potentially increase of the long-term profit.

2.7 Design and performance evaluation of whole RTO systems

Unfortunately, the functional modularity of on-line optimisation system has lead to an apparently independence of many research areas, where the structural behaviour of the whole systems receives little and may be, not enough attention. However, as any other complex system, a successful on-line optimisation project requires a set of structural decision (Perkins, 1998):

- Structure of the regulatory control system, in terms of set-points, measured and manipulated variables.
- The variables selected to transmit the results from the on-line optimisation system to the plant.
- The model parameters to be estimated from plant data in an on-line way.

- The plant measurements used to estimate model parameters.
- The level of model complexity to be used in the RTO system.

In this way, when making choices in each of these areas from the range of feasible possibilities, the aim should be to achieve the best possible performance of the closed-loop system, by mitigating the effects of inevitable errors introduced though measurements noise and plant-model mismatch on optimised plant performance.

Currently, there are two systematic approaches to RTO system design and performance evaluation, with special emphasis over this latter. Both approaches are strongly based in the concepts explained in the previous section (page 34):

- De Hennin et al. (1994), also Loeblein and Perkins (1996, 1998) proposed a performance characterisation for RTO systems based on the average deviation from optimum and the required back-off from active inequality constraints, which help insure only feasible operating points are produced by the optimisation system. In this approach, better RTO designs will produce a smaller back-off and average deviation from optimum. The idea has been later extended to consider the process dynamics, assuming a lower layer of Model Predictive Control, which is usually the case.
- In a parallel way, Forbes and Marlin (1994b,a) used the "design cost" criterion, which is defined as the average long-term lost profit due to imperfections in the design of an RTO system. It is based in the propagation of noise from sensors to the predicted optimum, which degrades therefore the long-term performance of the system. Then, any RTO design decision should be made to minimise this design cost, and the performance of an RTO system can be determined in terms of its design cost. Zhang and Forbes (2000) have extended the idea to include both, the transient and long-term behaviour of a closed-loop RTO system. In their context, the term transient refers to the intermediate bounded solutions given by the RTO system since the true dynamics transient of the process have been ignored (Narraway et al., 1991).

2.8 Overall strategy

There is a common denominator in all the literature reviewed in the previous sections: the functional scheme of figure 2.1. Several authors have then devoted their efforts in proposing changes in that scheme with the aim of avoiding the weak points mentioned in section 1.4.2 (page 12). The more significants are:

- Besl et al. (1998), implemented a system that does not wait until steady state but is periodically optimising, performing data reconciliation only when the steady state is achieved. Unfortunately, this approach cannot always be implemented because the optimisation quality strongly depends on data and model quality. Considering that data and model are fitted only under steady state, this approach could lead to optimise an unsuitable model with unsuitable data. Furthermore, for non-smooth disturbances, it could mean aggressive and profitless changing of set-point values.

- Otherwise, Cheng and Zafiriou (2000), have presented an interesting approach, similar to that of Roberts (1979); Roberts and Williams (1981) where model updating is implicitly made during optimisation, and the optimisation algorithm is directly applied over the plant in successive steps. This approach has some disadvantages as well. Firstly, the updated model is not available for off-line studies. Secondly, several set-point changes are required (over the plant) for gradient evaluation, which can lead to undesirable plant behaviour. Moreover, after each set-point change, the steady state has to be reached in order to allow an acceptable gradient approximation, which is clearly time consuming, and limit its applicability to small settling time processes with low disturbance frequency. Finally, in that work, the methodology is tested over a steady state model, which is a very restrictive approach.

- An interesting contribution came from Multivariable Predictive Control (MPC) research area, which incorporates economic objectives explicitly into the controller in order to achieve an adequate transition between the current and the desired operating point. One way of incorporating economy into the controller is to add to the standard quadratic regulatory objective a simplified economic objective as proposed by several authors (e.g. Tvrzská and Odloak (1998); Becerra et al. (1998); Nath and Alzein (2000)). However, the fact that all the control/optimisation requirements are translated into a single scalar performance index imposes a trade-off between control and optimisation, making the choice of relative weights quite difficult. Furthermore, such combination will typically lead to non-linearities, and thus, its on-line solution may be hardly compromised.

- Another alternative, widely used in industry (e.g. Cutler and Perry (1983); Qin and Badgwell (1997); Sorensen and Cutler (1998); Rao and Rawlings (1999); Ying and Joseph (1999)) is to send the RTO results to a local Linear Programming (LP) or Quadratic Programming (QP) steady state controller coupled to the MPC (termed LP(QP)-MPC). Such a cascade control scheme continuously computes and updates the set-points used by the lower level MPC algorithm, producing an evolutionary transition with an excellent control performance (figure 2.7). However, besides the complexity of such systems, significant problems occur when both the LP(QP)-MPC and RTO layers have economic and performance objectives which may not match (Mizoguchi et al., 1995). In addition, it should be also considered that for certain processes the implementation of MPC could be not justified or even to be not the appropriate choice (Qin and Badgwell, 1997).

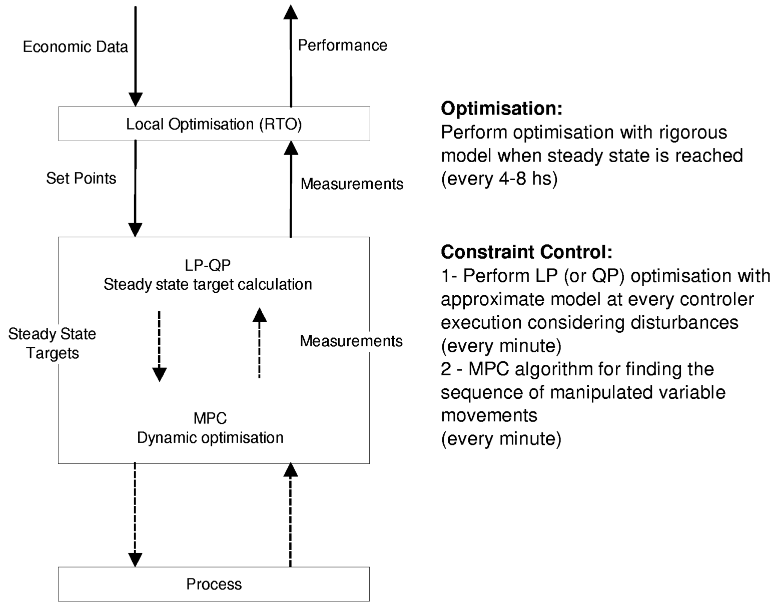

Figure 2.7: RTO-MPC hierarchy

2.9 Processes with decaying performance

The efficiency of equipment units in chemical process plants decreases with time because of fouling, formation of by-products, catalysts deactivation etc. Therefore, periodical equipment maintenance and/or cleaning are required to restore original operating conditions that allow maintaining the desired high productivity rates. In such cases a trade-off decision must be made between the cost associated to equipment shutdown and the resulting benefit of productivity improvement (Reg Bott, 2001). The resulting decision-making problem consists in determining *when* such actions should be carried out, thus leading to a trade-off between the cost of the action itself (resources required, production break, etc.) and the expected benefit it generates (productivity increase, reduction of operational costs, etc.).

Moreover, in order to keep a smooth production output, plant equipment may be run in parallel. Typical examples of this situation are catalytic reactors, evaporators, filters etc. The problem with catalytic reactors is catalyst progressive deactivation. In the case of continuous evaporation units, efficiency decreases because of different factors like fouling, deposit of insoluble materials, etc. Semi-batch filtering efficiency is affected by the continuous increase of the pressure drop through the cake. The existence of parallel lines introduces additional constraints related to the mass balance, besides possible bounds in the average processing rates.

Regarding the scenario contemplating a single equipment item, several methods for the optimisation of the required cleaning schedules have been reported. Epstein (1979) presented an analytical method for the calculation of the optimal evaporator cycle with scale formation. Similarly, some authors (e.g. Casado (1990); O'Donnell et al. (2001)) used a detailed cost model to calculated the optimal cleaning cycle of a heat exchanger under fouling by exploring the major operating trade-off. In addition, Sheikh et al. (1996) presented a reliability-based cleaning strategy by incorporating uncertainty, in a linear fouling model. In the same context, some different stochastic fouling models have been also considered (Zubari et al., 1997).

In addition, for the case of catalytic reactors, many detailed models coming from the reaction engineering research area have been useful for state off-line optimal control problems for key variables like temperature and compositions, including as well the cycle length as optimisation variables. For instance, Borio et al. (1992) addressed the simulation and optimisation of a fixed bed reactor operating in cocking-regeneration cycles. In other work, Borio and Schbib (1995), study a set of catalytic reactors used for de-hydrogenation of butene into butadiene. Similarly, Castilla et al. (1992), studied an isothermal fixed-bed reactor devoted to the transformation of methanol into hydrocarbons and Taskar and Riggs (1997) studied a semi-regenerative catalytic naphtha reformer.

Several approaches have been proposed to cope with parallel equipment, which usually contemplate standard planning and scheduling problems formulated as mathematical programs, generally MILP and MINLP. Most of the times, these formulations are adaptations of already available models for the planning and scheduling of batch plants.

Muller-Steinhagen (1998) proposed an integrated approach for developing alternative fouling mitigation strategies based on both, experimental and modelling works. Jain and Grossmann (1998), proposed a mixed non-linear programming model using an exponential decay with time, to find a cyclic schedule for feed processing. Smaïli et al. (1999) presented also a MINLP for addressing the problem of cleaning synchronisation of heat exchanger networks (HENs) with application to sugar cane industry. In a subsequent work, Smaïli et al. (2001), developed an MINLP model for optimising the cleaning schedule in large continuously operating heat exchanger networks under fouling. Georgiadis and Papageorgiou (2000b,a) presented general mathematical frameworks for optimal scheduling and energy management in HENs under fouling. Mixed-integer linear programming models in discrete time domain were developed, which were solved to global optimality. Similar works were proposed by Pinto and Grossmann (1994); Alle and Pinto (2001). Alle et al. (2002) extended these previous scheduling models for multi-product, multi-stage continuous plants based on continuous time representation to include cleaning considerations.

Some works addressed the problem from a general perspective (e.g. Dedopoulos and Shah (1995); Sanmartí et al. (1997); Pistikopoulos et al. (2001)) including the simultaneous study of both preventive maintenance (also corrective in some of the works) and production planning of multi-purpose batch plants. In this context, it should be mentioned the contribution of Vassiliadis and Pistikopoulos (2001) that also considers parametric uncertainty.

Unfortunately, the implementation of the decisions obtained using mathematical programming models may result in non-optimal operation because of plant-model mismatch and the inherent variability of plant operating conditions. Furthermore, little (if any) attention has been paid to the on-line management of the discrete decisions involved in processes with decaying performance. Specifically, it is usually ignored that the model updating module of an on-line optimisation system, or more generally, the distributed control system (DCS), may provide significant information about the process degree of performance and its variation with time.

2.10 Industrial applications

There have been several industrial applications of on-line optimisation reported recently, mostly in refineries and chemical plants where the models are highly reliable. The improvements in plant operation and economics range from a 1% to 20 % increase in profit (e.g., Lauks et al. (1992); Van Wijk and Pope (1992); Hardin et al. (1995); Bayles (1996); Basak et al. (2002)).

Ayala (1997) reviewed 31 Aspen CLRTO systems in operation and 21 in progress of which about 50% were refinery applications, 30% were ethylene plants and 20% were chemical processes. Although many details are commonly hidden in the original papers, following list gives a short summary of some of the available literature related to industrial implementations, which is far from being exhaustive.

Some reported industrial implementations of RTO systems		
Authors-Plant		**NOTES**
Crowther et al. (1961)	Amoco's CDU (140.000 bbl/day), American Oil Co.	The first detailed implementation of on-line optimisation system. Benefits of approximately 0.05 $/bbl.
Lauks et al. (1992)	Refinery unit plus Ethylene Plant at ÖMV Deutschland GmbH. Burghausen, Germany.	The program runs on a DG-AVIION 4200 Unix system. About 100 constraints and 37 decision variables. It takes 70 minutes approximately to run the optimisation. The improvements on profit reported are of 1-3 %
Van Wijk and Pope (1992)	Shell Refinery: DCU, FCC, CR and GAS	Use of Shell Multivariable Optimising Control Package and Online Optimisation Package. About 9% increase in production.
Hardin et al. (1995)	Conoco's Lake Charles, L.A. Refinery	Aspen Tech's RT-Opt was used for modelling and optimisation, and the results were implemented through Set Point's Idcom Controller yielding an additional 0.03 $/bbl
continue ...		

Some reported industrial implementations of RTO systems		
Authors-Plant		**NOTES**
Pedersen et al. (1995)	20000 bpd hydrocraker complex at Sunoco Inc., Sarinia Refinery (Canada)	Commercial equation solving and optimisation packages used, along with interface software to solve flowsheeting problems. Applied with advanced DMC controllers, resulting in huge incremental proffit
Yoon et al. (1996)	Hyudai Petrochemicals Complex at Daesan, Korea, with 350000 tons of ethylene and 175000 tons of propylene /year	Equation-based model with 60000 equations. SQP used for optimisation; 2.45 % reduction in energy and feedstock consumption reported.
Chiari et al. (1997)	Agip Refinery in Sanazzaro dei Burgondi, Italy	Hydrogen and Sulphur plants. The optimisation package (ORO) has been installed on a IBM risc 6000 computer and it runs by using on-line hourly average plant data.
Georgiou et al. (1998)	Steady-State implementation of CLRTO at Mobil Chemical's Beaumont, TX, ethylene plant	Accurate furnace models having material balance closure checks developed for optimisation of the unit. The system was also used for off-line planning and performance monitoring
Tvrzská and Odloak (1998); Zanin et al. (2000)	LPG in FCC converter. Sao José dos Campos, Brazil.	Non conventional integration of RTO and advanced controllers.
Besl et al. (1998)	Penex process for light naphtha isomerisation at RVI's Ingolstad Refinery, Germany	Aspen Tech's RT-OPT open form equation-based optimisation software used
Georgiou et al. (1999)	CLTRO of FCC and alkylation complex in Mobil's Joliet Refinery	Library Model from AspenTech used for optimisation, implemented through DMC Plus type controller for maximum profitability
Geddes and Kubera (2000)	CLRTO in olefin production	Production scheduling LP model coupled with RTO to improve overall olefin plant operations significantly
continue ...		

Some reported industrial implementations of RTO systems		
Authors-Plant		**NOTES**
Hou et al. (2001)	CDU: Lanzhou Petroleum Refinery, China	Stochastic approach to avoid narrow quality. \$500.000/year improvement
Tresmondi et al. (2001)	Rhodia Brasil, Phenol Plant	RTO integrated to MPC and online analysers. Reduction of 4.7 % of by-products.

2.11 Commercial vendors

Many advanced control and modelling technology companies commercialises software packages that apply at least partially, the techniques explained in this chapter. Some of the most important advanced control companies and their packages are following included:

- Simulation Sciences: ROMeO
- ChemShare: Mirror Model
- Aspen Technology RT-Opt
- Hyprotech RTO+
- Setpoints: OPTCOM
- Treiber Controls: OPS
- Profimatics: On-Opt
- Dynamic Matrix Control (DMC) Corporation CLRTO

It should be noted, however, that advanced control and on-line optimisation are growing areas, and thus such companies are forming partnerships that capitalise on their individual capabilities. Many cases of co-operative agreements have done between the traditional advances control companies, with significant degree of expertise in model predictive control (e.g. SetPoint, DMC, Honeywell, etc.) and those with long tradition in chemical process simulation (SimSci, Aspentech, Hysys, etc.). These changes were caused by an industry demanding for the integration of on-line optimisation and advanced control and it is recognised that they do not share details of methodology to maintain a competitive advantage.