Statistical Process Adjustment: a brief retrospective, current status, and some opportunities for further work

E. del Castillo

Department of Industrial & Manufacturing Engineering
The Pennsylvania State University, University Park, PA 16802, USA

November 2005

Abstract

Industrial Statisticians frequently face problems in their practice where adjustment of a manufacturing process is necessary. In this paper, a view of the origins and recent work in the area of Statistical Process Adjustment (SPA) is provided. A discussion of some topics open for further research is also given including new problems in semiconductor manufacturing process control. The goal of the paper is to help display the SPA field as a research area with its own identity and content and promote further interest in its development and application in industry.

This is a fuller version of a paper to be published in *Statistica Neerlandica*.

Key Words: Engineering Process Control, Time Series Control, EWMA controllers, Setup Adjustment, Bayesian Control, Deadband Adjustment.

1 Introduction

Located at the intersection of Control Theory, Time Series Analysis, and Statistical Process Control, the Statistical Process Adjustment (SPA) field is a set of Statistical Techniques aimed at modelling, and hence, forecasting and controlling a dynamic process. Two distinctive characteristics of SPA are a) that the process responses relate to quality characteristics of a product (or of a process producing it), and b) the implementation of the adjustments is not fully automatic since SPA corresponds to a higher-level supervisory controller, i.e., a controller of lower-level controllers which in turn operate on a production process. Property a) differs from many control theory applications where some physical variable is of interest, but the aim is not necessarily *quality* control, and b) emphasizes the hierarchical nature with which the adjustments are implemented, on whole complex processes or machines made of several different components, but modeled as a single processing stage. Given the complexity of the machine or process, only a statistical, i.e., data-based modeling is feasible. This is in contrast to first principles models frequently used in control theory.

A key question we would like to address in this paper is: is SPA an area with enough intellectual content and practical relevance to justify its study within Statistical methodology? We pose this question because two widespread conceptions found among Statisticians and Engineers:

1. process adjustments are, for the most part, unnecessary in practice. This believe is based mainly on statements in Deming's writings and in particular in relation to his "funnel experiment";

2. process adjustments are of course necessary, but practically all the relevant problems have been solved by control theorists. While control theory is a fertile and active area of research, all the problems in SPA have by now been solved. Most of the work on SPA is simply a repetition of previous control theory work.

Believe 1 is found mainly among Statisticians and has been discussed in the literature at length based on the funnel experiment (e.g., MacGregor 1990); believe 2 is found among Engineers, and has not been discussed much, if at all, in the literature. It is the purpose of this paper to show that both viewpoints are misconceptions, but we will place more emphasis on arguing against the second viewpoint. In order to do this, we will review how the SPA field originated, what were the main initial problems, how it has evolved in general terms, and perhaps more importantly, what recent work relevant in industrial practice has been conducted. No effort to provide a complete literature review was made. For bibliographic references up to 2001 see Del Castillo (2002a).

It is hoped that the problems described here will provide renewed impetus to the area. The paper closes with a discussion of relevant, practically-important problems which are still open for solution. The objective is to provide, by example, enough evidence for an unqualified "yes" answer to the key question posed above, and provide some hints for further research.

2 Origins of SPA

Since its origins in the early 60's, work in what we now can consider the SPA field was developed by both Control Engineers (working in quality control applications) and by Statisticians, who, for the most part, had a background in Chemical Engineering. This is more than a simple anecdote, since it determined what type of processes and corresponding problems were initially studied in this field. SPA originated from work apparently done independently by Box and Jenkins (1962, 1963) on adaptive optimization and control and minimum mean square error (MMSE) control, and by Åström (1963) on "minimum variance" (equivalent to MMSE) control (Åström was interested in implementing Kalman's ideas on Adaptive Control based on operating data). While there were some interesting papers on process control (as opposed to control charting) written by Statisticians in the late 50's and early 60's (e.g. Barnard, 1959), the work by Box and Jenkins was the most influential in the Statistics literature.

The MMSE control problem relates to finding a rule (a "controller") that tells us how to vary a controllable factor x_t such that the MSE of a dynamic response y_t (which we assume to be deviations from target) is minimized in the following transfer function model:

$$y_t = \frac{B'(\mathcal{B})}{A'(\mathcal{B})} x_{t-k} + \frac{C'(\mathcal{B})}{D'(\mathcal{B})} \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2). \tag{1}$$

Here, B', A', C' and D' are polynomials in the backshift operator $\mathcal{B}$ (all polynomials start with a one except B' which starts with some arbitrary constant b_0) and k is the delay, i.e., the number of whole discrete time periods between the controllable factor is changed and its effect on the response appears to be observed for the first time. It is assumed all time series have equidistant observations in time. Even control engineers to this day call this model the *Box-Jenkins model*. Contrary to other different ways of writing a transfer function model, (1) has a natural "signal plus noise" interpretation, since if $x_t = 0$ for all t, then $y_t = \frac{C'}{D'}\varepsilon_t$ is an ARIMA model that represents the uncontrolled output (here polynomial D' may have one or more roots on the unit circle, allowing to model *homogeneous non-stationarity*, see Box, Jenkins and Reinsel, 1994). This also occurs if x_t =constant since then the first term on the right is simply a constant. Model (1) also has the advantage of avoiding multiple common terms when fitting (Box et al., 1994).

Despite the nice interpretation and model-fitting advantages, it is perhaps easier to derive an MMSE controller with fewer polynomials around. The ARMAX form of a transfer function model, used by Åström, is

$$A(\mathcal{B})y_t = B(\mathcal{B})x_{t-k} + C(\mathcal{B})\varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2) \tag{2}$$

with the obvious relations between the two models being $A = A'D'$, $B = B'D'$, and $C = C'A'$. Let n be the maximum order of the polynomials A, B and C. The optimal MMSE

feedback controller, found by both Åström and Box and Jenkins is:

$$x_t = -\frac{G(\mathcal{B})}{B(\mathcal{B})F(\mathcal{B})} y_t \tag{3}$$

where the G and F polynomials are of orders $n-1$ and $k-1$, respectively, and are obtained by equating coefficients of like powers of $\mathcal{B}$ in $C = AF + \mathcal{B}^k G$ (here, G starts with a constant but F starts with a one). The *adjustments* are then given by $x_t - x_{t-1}$, a quantity usually denoted in the time series literature by ∇x_t. Under the actions of this controller, the controlled output then obeys the model

$$y_t = F(\mathcal{B})\varepsilon_t, \tag{4}$$

which, as it can be seen, is a moving average process of order $k-1$. Thus, controlled processes will not always be uncorrelated, even when controlled via a MMSE controller.

It was immediately obvious to both Åström and Box and Jenkins that (3) may be a very expensive control rule since it calls for considerable manipulation of x_t; if there is a cost associated with the adjustments this rule cannot be optimal. Noticing this, Box and Jenkins further studied two related problems:

1. the case when the adjustment cost is fixed, thus the objective is to minimize

$$J_1 = E\left[\sum_{i=1}^{\infty} y_i^2 + c\delta(x_{t-1})\right]$$

   where $\delta(x) = 1$ if $x \neq 0$ and $\delta(x) = 0$ otherwise. This gave birth to the so-called *deadband adjustment policies* (Box and Jenkins, 1963), which we discuss in Section 4 below;

2. the case when the variance of the input (x_t) is constrained. Thus, we min $E[\sum_{i=1}^{\infty} y_i^2]$ subject to $\mathrm{Var}(x_t) \leq c$. The Box-Jenkins constrained variance controllers are relatively complicated to obtain and result in complex controllers. Constraining the variance is equivalent to assuming a quadratic cost in the control factor itself, thus we could minimize

$$J_2 = E\left[\sum_{i=1}^{\infty} y_i^2 + cx_t^2\right]$$

   (alternatively, the adjustments ∇x_t may be used instead). If the errors are normally distributed, this is a well-known problem which is solved with Linear-Quadratic-Gaussian (LQG) control theory, see Åström (1970).

From this initial work, an explosion of related work took off. Åström and his colleagues and students (notably L. Ljung) went ahead and founded the Swedish school of Adaptive Control. Adaptive Control is by now a mature discipline within Control Theory (Åström and Wittenmark, 1989). Adaptive controllers continuously re-estimate the parameters of a given model, thus their properties are difficult to analyze. Although such recursive estimators are known to *burst* if the inputs (the x's) do not vary enough, in practice several safeguards that monitor the "health" of the estimator (not unlike SPC schemes) are applied to provide persistent excitation without bursting (see Ljung, 1999) (the regression equivalent of the lack of excitation problem is an $X'X$ matrix very ill-conditioned due to similar rows in X).

For their part, Box and Jenkins and their students (notably J. MacGregor) continue to develop SPA in the 70's and beyond by studying problems with a clear Statistical content. We will review some of this work. But first it is pertinent to address a not so-well-known misconception created by Deming's funnel experiment.

3 When are process adjustments necessary?

There is a considerable good understanding of when adjustments are necessary and why, see e.g., MacGregor (1988) and Del Castillo (2002a, chapter 1), who discuss this issue based on Deming's funnel experiment. We would like to point out a common misconception made by some authors who prefer not to contradict directly Deming's remarks, as expressed, e.g., in his *Out of the Crisis* book (Deming, 1982). It is sometimes argued that Deming's remarks about not to adjust are actually correct provided the process mean is not moving. It is argued that if the process mean changes with time, then adjustments are needed, otherwise, they are not. This is incorrect. *A moving mean is neither a necessary nor a sufficient condition for adjustments to be required.* To show why, consider the funnel experiment in which, the analogous univariate process would obey Shewhart's model (the marbles evidently obey a bivariate process):

$$y_t = \mu + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2), \quad t = 1, 2, ...$$

where y_t is the observed deviation from target and μ is the mean deviation from target. Deming assumes the funnel to be on target, i.e., $\mu = 0$. There are two important aspects to notice: first, the process starts on target and remains there unless an adjustment is made. Second, the observations $\{y_t\}_{t=1}^{\infty}$ form an i.i.d. (normal if the ε_t's are normal) sequence.

In a very important but somewhat neglected paper (despite being reprinted in 1983), F. Grubbs (1954) assumed the second condition above but assumed that, at startup of the process, $|\mu| = d \neq 0$, where d is a setup error. If the only cost of interest is the cost incurred when running the process off-target, then it is evident that corrective action is necessary. How to do this in such a way that $E[\sum_{t=1}^{N} y_t^2]$ is minimized is called the *Setup Adjustment Problem*, with many variants that have been studied considerably in the last 7 years (Del

Castillo, 1998, Trietsch, 1998, 2000, Pan and Del Castillo, 2003, 2004, Colosimo et al., 2004, Lian et al., 2005).

Note how the process mean is *constant*, there is no autocorrelation, and still process adjustments are needed. This shows by counterexample that it is not *necessary* that the process mean moves for adjustments to be required. A moving mean is neither a *sufficient* condition for adjustments to be required: as a counterexample consider the case of a process with a moderate drift in the mean such that all product will be within specifications (or not too far from target to cause a substantial cost) for the duration of the production run, and suppose the cost of adjustments is relatively very large. Then, it follows that adjustments are not justified.

In conclusion, the need for process adjustments depends on the process model and the cost structure. Evidently, if all conditions behind Deming's funnel experiment hold, then simple process monitoring is optimal from a MMSE point of view.

4 Some recent work in SPA

In this section we highlight some recently studied problems that illustrate the SPA field.

4.1 The Setup Adjustment Problem

Solutions to the setup adjustment problem and to many of its variants, some to be described shortly, are very important for the control of *discrete part* manufacturing processes. In this type of processes, the operation of setting up a machine for production of a new lot may induce offsets or shifts in the values of the quality characteristics of the parts relative to their targets. No disturbance other than the setup offset and white noise are assumed. If the unknown offset is a constant d, the deviation from target at time t can be expressed as (Del Castillo et al., 2003):

$$y_t = \mu_t + v_t, \quad v_t \overset{iid}{\sim} (0, \sigma_v^2) \tag{5}$$

$$\mu_t = \mu_{t-1} + \nabla x_{t-1}, \quad t = 2, 3, ... \tag{6}$$

$$\mu_1 = d + x_0 \tag{7}$$

For this simple but practical model, Grubbs (1954) solved the problem

$$\min \ \mathrm{Var}(\mu_{n+1}) \quad \text{s.t.} \quad E[\mu_{n+1}] = 0.$$

The solution is given by the "harmonic" rule (so called given the series $\{1/t\}$, see Trietsch, 1998):

$$\nabla x_t = -\frac{1}{t} y_t \tag{8}$$

for all t. If several lots are produced, the setup error can be considered random over lots. Grubbs then proposed to model the offset as $d \overset{iid}{\sim} N(0, \sigma_d^2)$, in conjunction with (5-7). Thus σ_d^2 is the between batch variance and σ_v^2 is the within batch variance. The objective now is to

$$\min \quad E\left[\sum_{i=1}^{n} \mu_t^2\right].$$

The solution is given by Grubbs' "extended" rule:

$$\nabla x_t = -\frac{1}{t + \frac{\sigma_v^2}{\sigma_d^2}} y_t. \tag{9}$$

Del Castillo et al. (2003) show how (9) results from using a simple Kalman filter to estimate the "state" μ_t and adjusting by $\nabla x_t = -\widehat{\mu}_t$. This formulation allows to apply Linear Quadratic Gaussian theory to extend the basic setup adjustment problem to multiple input-multiple output (MIMO) problems, problems with errors in the adjustments, and problems with quadratic adjustment costs. It also allows to make interesting connections between the setup adjustment problem and other Statistical techniques such as Stochastic Approximation and Recursive Least Squares. These extensions and connections were not possible using Grubbs' more complex approach to the problem.

While the basic setup adjustment problem can be solved with well-established control theory techniques, many more important variations are problems that cannot be solved making use of existing control techniques and require new methodology.

From a Statistical perspective, the most interesting and practical variation of this problem is when the process parameters μ_d, σ_d^2, and σ_v^2 are unknown. Then the problem is an Adaptive Control problem, as it involves controlling a process with unknown parameters. The structure of the problem, however, has not been addressed by Control theorists, as far as we know, since here the variances need to be estimated on -line.

Recent work in setup adjustment under unknown paraments is Bayesian and based on Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) techniques. In Colosimo et al. (2004), the assumed model is analogous to (5-7):

$$\begin{aligned} y_{ij} &= \theta_{ij} + v_{ij} && (10)\\ \theta_{ij} &= \theta_{i(j-1)} + \nabla x_{i(j-1)} && (11)\\ \theta_{i0} &\sim (\mu, \sigma_\theta^2), \quad v_{ij} \overset{iid}{\sim} (0, \sigma_v^2) && (12) \end{aligned}$$

where $i = 1, ...I$ denotes lots and $j = 1, ..., J$ denotes parts within a lot. The objective is to minimize $E\left[\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2\right]$. The resulting MCMC controller adjusts between lots and within lots. The between lot adjustments are

$$\nabla x_{i0} = x_{i0} - x_{i(-1)} = x_{i0} = \begin{cases} -\widehat{\mu}|D_{(i-1)J}, & i \geq 3 \\ 0, & i = 1, 2 \end{cases} \tag{13}$$

where $\widehat{\mu}|D_{(i-1)J}$ is the mean of the posterior distribution $p(\mu|D_{(i-1)J})$ and $D_{(i-1)J} = \{y_{11} - x_{10}, y_{12} - x_{11}, ..., y_{(i-1)J} - x_{(i-1)(J-1)}\}$ is all data observed before lot i starts (note how the model can be written as $y_{ij} - x_{i(j-1)} = \theta_{i0} + v_{ij}$). Adjustments within lots are

$$\nabla x_{ij} = -\widehat{\theta}_{ij}|D_{ij}, \quad j = 1, 2, ..., J-1$$

The Bayesian MCMC controller learns how to "anticipate" the offsets, providing a performance that eventually mimics a feedforward controller. Unnecessary adjustments that may inflate the overall process variance are reduced via a conditional first adjustment rule, in which (13) is implemented only when a credibility interval for μ excludes zero. See Lian et al. (2005b) for details.

Recent work on setup adjustment includes the case when parameters are known with sufficient accuracy. If this is the case, then the sum of the total cost of running the process off target and of adjusting can be minimized by defining a *schedule* of adjustments, much in the sense of a maintenance plan. See Trietsch (2000) and Pan and Del Castillo (2004). Other recent work includes the case of an asymmetric off-target cost, a common situation in discrete part manufacturing. One approach is to let the process converge to target from the side of least cost. Stochastic approximation techniques can then be used for this purpose. See Colosimo et al. (2005) for more details.

Other relevant variations of the setup adjustment problem, in particular, integration with process monitoring, the case when there a re fixed adjustment costs –resulting in "deadband" policies– and the use of SMC techniques are described in the next sections.

4.2 "Deadband" adjustment policies

As mentioned before, Box and Jenkins showed that a fixed adjustment cost implies a "deadband" controller. They consider a process with a pure delay function and IMA(1,1) noise, namely,

$$y_t = gx_{t-1} + \frac{1-\theta\mathcal{B}}{1-\mathcal{B}}\varepsilon_t. \tag{14}$$

In this "machine tool" problem, as called by Box and Jenkins, the objective is to minimize J_1 (using ∇x_{t-1} instead of x_{t-1}). The problem leads to a dynamic programming formulation, although Box and Jenkins instead minimized the long-run average cost per time unit. The solution is

$$\nabla x_t = -\frac{1}{g}\widehat{y}_{t+1|t}, \qquad \widehat{y}_{t+1|t} = (1-\theta)y_t + \theta\widehat{y}_{t|t-1} \tag{15}$$

which is applied whenever the one-step ahead predictor falls outside of a "deadband", i.e., whenever $\widehat{y}_{t+1|t} \notin (-L, L)$. The predictor is an exponentially weighted moving average (EWMA) of the observations. The deadband limit L is a function of c, σ_ε^2, and θ. Tables for L are in Box and Luceño (1997). More discussion on the long-run average cost approach

to solving deadband adjustment problems can be found in Luceño and Gonzalez (1999) and Luceño (2003).

Crowder (1992) solves the machine tool problem for a finite horizon, namely

$$J_3 = E\left[\sum_{i=1}^{n} y_i^2 + c\delta(\nabla x_{i-1})\right].$$

The dynamic programming problem, solved by Crowder, yields a deadband solution analogous to the infinite-horizon solution but with deadband limits L_t that "funnel out" as the end of the production run approaches. The implication is that if the process will end soon, an adjustment at that point brings less future benefits than an adjustment early in the production of the lot. Jensen and Vardeman (1993) consider the same finite-horizon problem as in Crowder (1992), but studied the case when adjustment errors can occur randomly. They show that even if no fixed adjustment cost exists, adjustment errors imply a deadband policy.

The work on deadband adjustment thus far summarized is based on knowing all process parameters. For the model assumed by Crowder (equivalent to that used by Box and Jenkins, equation 14):

$$\begin{aligned} y_i &= \theta_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2) \\ \theta_i &= \theta_{i-1} + \nabla x_{i-1} + v_i, \quad v_i \overset{iid}{\sim} N(0, \sigma_v^2) \end{aligned}$$

the parameters σ_v^2 and σ_ε^2 are assumed known by these authors, thus a Kalman filter (essentially the EWMA used in equation 15) can be used to estimate the state θ_i.

Recent work by Lian and Del Castillo (2005) considers the machine tool problem when the process variances are unknown. A Bayesian approach based on Sequential Monte Carlo methods (see Section 5) was implemented to solve this problem. The procedure updates the posterior of θ_i, σ_v^2, and σ_ε^2 at each point in time i, from which the adjustment

$$\nabla x_i = -E[\theta_i | \text{data up to } i] = \widehat{\theta}_i$$

is implemented provided $|\theta_i| > L_i$, where L_i changes dynamically based on the posterior distribution of the parameters. Interestingly, at the beginning of a lot the deadband limits are wide due to the uncertainty in the parameters, then they narrow as more data is collected and finally they funnel out towards the end of the lot due to the end of horizon effect.

Recently, Crowder and Eshleman (2001) propose an alternative to the Bayesian SMC approach just described based on using Maximum Likelihood Estimation for the variances from a set of *n open loop* runs (i.e., data obtained when the controller is disconnected),

and then plug-in those estimates in the usual Kalman filter estimate of the state. They reported small sample properties of the estimators, concluding that at least between 25 to 50 observations are necessary to obtain reliable estimates. It would be interesting to see how can the MLE method be modified for closed-loop data, and how it behaves compared to the Bayesian approach. It seems likely that for non-informative priors, the performance should be quite close, with the advantage of the Bayesian method of being able to incorporate prior information about the parameters in case it exists, which would improve convergence of the estimation process.

It is important at this point to emphasize that these and other variations of deadband control are not considered in the control theory literature.

4.3 PI and EWMA control

A discrete-time proportional integral derivative (PID) controller has the form

$$x_t = K_P y_t + K_I \sum_{i=1}^{t} y_i + K_D \nabla Y_t$$

or, in the more common *incremental form*:

$$\nabla x_t = K_p \nabla Y_t + K_I Y_t + K_D \nabla^2 Y_t.$$

The last term, with action proportional to the difference of the response, is frequently not used in practice (Del Castillo, 2002a, Chapter 6). This results in PI controllers, which have received considerable attention in the Statistics community due to the work by Box and Luceño (1997). These authors show how to implement a PI controller graphically (an idea first shown by Box and Jenkins, 1963), and show how PI controllers are quite robust with respect to variations in the assumed process model. They convincingly show that the inflation in variance due to adjusting with a PI controller a process that requires no adjustment, –like Deming's funnel– is quite moderate.

The robustness of PI controllers is widely acknowledged and known by process engineers in practice. Process engineers working in industry know well the adagio that says that *all it takes most of the times is a good integral controller.*

The robustness of PI controllers comes from its integral action, $K_I \sum_{i=1}^{t} y_i$. A "P" controller, one in which the controller linearly tracks the response, is a quite poor controller in general, since, for example, it does not provide offset-free control. In contrast, and this should be of interest to persons familiar with Statistical Process Control (SPC) charts, *an integral controller will compensate against shifts in the mean of a stationary process.* This will make detection of a shift in a PI-controlled process difficult (see Jiang and Tsui, 2002,

and the discussion below on SPC-EPC integration). The time to recover of the process will be a function of the shift magnitude and of the integral parameter K_I. In principle, if the adjustments are unconstrained in size, and integral controller will eventually compensate against shifts of *any* size. The design of a PI controller consists in selecting K_P and K_I (see Box and Luceño, 1997, Tsung et al., 1998). Constrained input variance PI controllers are discussed by Box and Luceño, 1997. An input variance constrained PI controller that tunes K_P and K_I on-line (i.e., it is self-tuning) was developed by Del Castillo (2000). The approach solves for the Lagrange multiplier of the constraint $\text{Var}(\nabla x_t) = c$, and uses this multiplier in the Clarke et al. (1975) controller, which utilizes it to constrain the input variance (see Del Castillo, 2002a). Since it is not based on a recursive estimator and utilizes Box and Luceño suggested settings for the parameters as initial values, the bursting behavior typical of adaptive controllers is avoided.

A particular type of integral controller, very popular in semiconductor manufacturing where it is described in "run to run" control applications (Moyne et al. 2000), is the so-called EWMA controller. It is easy to show that for a process like

$$y_t = gx_{t-1} + N_t \tag{16}$$

and integral controller is equivalent to setting

$$x_t = -\frac{1}{\widehat{g}}\widehat{N}_{t+1|t}$$

where $\widehat{N}_{t+1|t}$ is an EWMA predictor of the disturbance N_t (following semiconductor manufacturing literature, we show $\widehat{g}$, an off-line estimate of the process gain, as g will be unknown in practice). Being an integral controller, an EWMA controller will compensate for shifts, but it will *not* compensate for drift, which will result in an offset in the quality characteristic (this is a well-known result in the Forecasting literature related to exponential smoothing). Since some real semiconductor manufacturing processes exhibit severe drift due to wear-out phenomena, a double EWMA controller (similar to double exponential smoothing) has been proposed (Butler and Stefani, 1994):

$$x_t = -\frac{a_t + b_t}{\widehat{g}}$$

where

$$\begin{aligned} a_t &= \lambda_1(y_t - \widehat{g}x_{t-1}) + (1-\lambda_1)a_{t-1} \\ b_t &= \lambda_2(y_t - \widehat{g}x_{t-1} - a_{t-1}) + (1-\lambda_2)b_{t-1} \end{aligned}$$

are two linked EWMA equations such that $a_t + b_t = \widehat{N}_{t+1|t}$.

A minimal condition for a good controller is that it must be stable. Stability has not always been considered in the SPA literature, as some of the discussants of the Box and Jenkins (1962) paper pointed out. Stability conditions of EWMA and DEWMA controllers has been a matter of study in the last 10 years (see, e.g., Ingolfsson and Sachs, 1993, Guo et al., 2000, Del Castillo, 1999). For a large variety of disturbances, an EWMA controller is stable if and only if $|1 - \lambda\xi| < 1$, where $\xi = g/\widehat{g}$. Stability conditions for DEWMA controllers with unit delay were derived by Del Castillo (1999) and later simplified by Tseng et al. (2002). They show how if N_t is an ARIMA(p,d,q) with drift model ($d \leq 2$), a sufficient condition for asymptotic stability is that $g/\widehat{g} < 3/4$. MIMO double EWMA controllers have been studied by Del Castillo and Rajagopal (2002).

The work thus far described on PI and EWMA control is just an application of existing methodology in Control Theory. To the eyes of control theorists, this work looks as straightforward approaches, compared to the complexities of current control theory research. Let us now turn to some new methodological developments that built on the previously cited work. In the last section of this paper we will also delineate some further problems related to EWMA control that arise in practice and require new methodologies.

Some interesting recent work by Hamby et al. (1998) introduces the concept of "probability of stability" and "probability of performance" in the design and analysis of EWMA controllers. These authors noted how in run to run applications, the gain g is usually fitted off-line based on designed experiments. In their paper, the gain is actually a vector $\boldsymbol{\theta}$, as they analyzed the multiple input, single output case (MISO) model:

$$y_t = \theta_0 + \boldsymbol{\theta}'\boldsymbol{x}_{t-k} + N_t.$$

Using the variance of the gain estimate they propose to evaluate

$$P(\mathbb{S}) = \int_{\mathbb{S}} p(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}$$

where $\mathbb{S} = \{\boldsymbol{\theta} : \text{system is stable}\}$ is the set of parameter values that make the process stable given the off-line estimate, and "data" is all available data obtained off-line. Thus, for a single EWMA controller, $\mathbb{S} = \{g : |1 - \lambda g/\widehat{g}| < 1\}$. They also introduced the closely related concept of "probability of performance", defined as

$$P(\mathbb{P}) = \int_{\mathbb{P}} p(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}$$

where $\mathbb{P} = \{\boldsymbol{\theta} : J(\theta) < \gamma\}$ is the set of parameter values such that a performance index J is less than some given value γ. The authors give a formulation to compute $P(\mathbb{P})$ for the case J is the mean squared deviation (MSD) of the controlled output assuming some given model, for which there are expressions available (Del Castillo, 2001). By maximizing

either $P(\mathbb{S})$ or $P(\mathbb{P})$ with respect to the EWMA weight λ, Hamby et al. (1998) provide a way to obtain controller settings that maximize either the probability of a stable system or the probability of having an MSD less than γ. Tseng et al. (2005) consider computing the sample size required in an off-line DOE that will guarantee some given probability of stability in a double EWMA controller, i.e., they suggest to compute the sample size n such that

$$P\left(\frac{g}{\widehat{g}} < \frac{3}{4}\right) \geq \gamma$$

holds. A recent and related paper is by Apley and Kim (2004) who also consider the MISO system as in Hamby et al. (1998). They suggested to compute the probability distribution of the inflation in variance that is, the distribution of $(\sigma_y^2 - \sigma_\varepsilon^2)/\sigma_\varepsilon^2$, where σ_y^2 depends on the true process parameters and controller parameters and σ_ε^2 is the uncontrollable variation. The distribution is taken with respect to the posterior distribution of the parameters, using a Bayesian approach. Apley and Kim (2004) went on to propose a *cautious controller*, which instead of minimizing the unconditional MSE, $J_4 = E[y_t^2]$, minimizes the conditional expectation:

$$J_5 = E[y_t^2|\text{off-line data}].$$

The conditioning is only on the data from an off-line experiment used to determine the gains and the expectation is with respect to the joint posterior of the gains and disturbance parameter estimates (an IMA(1,1) was assumed, so it has one other parameter, θ) and with respect to the variance of the errors (Apley (2005) extends these results to a general ARIMA disturbance).

Apley and Kim (2004) show that $P(\mathbb{S})$ for the rule that results from minimizing J_5 is always higher than $P(\mathbb{S})$ obtained using an MMSE controller, assuming the disturbance follows an invertible IMA(1,1) model. The resulting controller is "cautious" because it considers the uncertainty of the parameters, but is not adaptive as no updating of the parameter estimates is proposed (i.e., model fitting is done only once, off-line). This avoids the complexities of the analysis of adaptive control schemes, in particular with respect to stability analysis (Åström and Wittenmark, 1995). It does not imply, however, that adaptive controllers are not advantageous in practice, since they have considerable value –although are not a panacea– mainly for short-run processes where little is known *a priori*. As mentioned earlier, bursting behavior can be monitored and controlled by a variety of schemes that safeguard the adaptive controller. Although simple applications of Adaptive Control procedures for Quality Control may not be academically challenging, it seems strange to object to their use in practice on the grounds that it is too *easy* to do so. Interestingly, the early papers on SPA promoted the idea of adaptation for control (Box and Jenkins, 1962), and that work actually predates the development of that field of control theory.

A last comment in this section relates to the idea of robustness used in the SPA literature and that in the highly technical *Robust Control* literature (see, e.g., Morari and Zafirou,

1989). The aim is the same, to develop controllers that are insensitive to uncertainties in the assumed model. As mentioned by Apley and Kim (2004), Robust Control, and in particular, H_∞ optimization is a mature field that has dominated most of Control Theory research in the last couple of decades. Its central precept is that if one can place deterministic bounds on the unknown parameters of a process, then a worst case performance index which considers variations of the parameters within such bounds can be optimized and a robust controller design obtained.

In the type of manufacturing quality control applications where SPA has evolved, such view of robustness is not satisfactory since parameters are usually estimated from production data of complex industrial processes (this is echoed by Åström and Wittenmark, 1989, when proposing Adaptive Control techniques). In such environment, it will be hard or impossible to place definite bounds on the variation of the parameters. These noisy, data rich environments imply that a probabilistic measure of uncertainty will generally be possible and preferable. The means by which probabilistic measures can be developed, as in the last two paragraphs above, is Bayesian inference, to which we return in Section 5.

4.4 "SPC-EPC" integration

Called "Fault Detection" and "Advanced Process Control" in semiconductor manufacturing circles, the integration of SPC tools and "Engineering Process Control" methods to operate on the same process is a problem that naturally falls within the SPA field. There have been two fundamentally different approaches for doing this integration of techniques:

1. The SPC mechanism acts in conjunction with an MMSE, PI, or other known controller which is active all the time. This approach was stated conceptually by Vander Weil et al. (1992), Faltin et al. (1993) and Tucker et al. (1993) who coined the term "Algorithmic Statistical Process Control". In this case, monitoring is typically conducted on the output of a controlled process, although approaches have been proposed to monitor both x_t and y_t jointly (Tsung and Shi, 1999). Note how the cause of a SPC signal can be assignable to a faulty feedback controller. In this way, the SPC scheme helps to monitor both the health of the process and of the EPC scheme. This has connections with the considerable body of work on SPC for autocorrelated data (literature too numerous to cite here), since the output of a controlled process is typically correlated in time (e.g., consider the closed-loop equation 4). This, in turn, relates to the analysis of the response patterns or "signatures" of a dynamic system to specific upsets (see, e.g., Yang and Makis (2000), Tsung and Tsui, 2003).

2. The SPC mechanism acts as a trigger of the EPC mechanism. This is the approach of authors such as Sachs and Ingolfsson (1995) and of Guo et al. (2000) in the area of "run to run" control (an early reference of this approach is Bishop, 1965). Usually, a step-like disturbance is assumed to occur with some probability. An SPC-like scheme is

used to detect the shift, but then an EPC scheme is used to correct for it. This typically makes sense in discrete-part manufacturing in which easy to vary controllable factors exist to compensate for such disturbance. The typically investigation and removal of the underlying cause of the shift can still be carried on in the usual SPC manner, as a record of such event is logged by the SPC scheme. To illustrate the decisions this problem entails, suppose again we assume the process is described by the model $y_t = \mu_t + \varepsilon_t$, where:

$$\mu_t = \begin{cases} \mu_{t-1} & \text{with probability } p \\ \sim (0, \tau^2) & \text{with probability } 1-p \end{cases} \tag{17}$$

so now the process can shift to a new random level (mean) at *any* time period t with some probability p, which we assume to be low (in accordance to what one would expect in a discrete manufacturing process). Let t_0 be an actual shift time. Then this SPC-EPC integration approach involves solving three problems:

(a) detection of the occurrence of the shift, i.e., estimate t_0.;

(b) estimation of the new process level μ_{t_0};

(c) given $\widehat{\mu}_{t_0}$, adjust the process to return to in-control level (0), i.e., find $x_{t_0}, x_{t_0+1}, x_{t_0+2}, ...$

Note how this links together three large fields within Industrial Statistics: Changepoint detection, Statistical Inference, and Process Control.

An instance of recent work along the first line of reasoning described above is by Jiang and Tsui (2002), who studied the Average Run Length properties of SPC charts designed to monitor the type of autocorrelated processes which result from adjusting a process with a MMSE or PI controller. The case of an MMSE controller is particularly tractable, given the closed-loop equation (4), thus essentially the problem is one of monitoring a MA($k-1$) process. They concluded that for PI-controlled process, detecting the presence of a shift is difficult by monitoring the output y_t. They suggested instead to monitor the level of the controllable factor (x_t). This can actually be generalized to any controller that has integral action: the integral action will compensate for the shift disturbance, thus only a transient "spike" in y_t will appear. The more aggressive the integral action is, i.e., the larger K_I is, the shorter this window of opportunity to detect will be. In some industrial processes, e.g., semiconductor manufacturing, aggressive I control is common, so this is a relevant problem in practice. Because of this "masking" of the assignable causes that impede their removal through the usual –but not modelled– process improvement steps that SPC recommends (called "technical feedback" by Box and Jenkins, 1962), some authors have argued against process adjustments. This is typically not an option, particularly if the process drifts, i.e., if it is open-loop unstable.

One way around this situation is precisely the second approach to SPC-EPC integration delineated above. Pan and Del Castillo (2003) studied the SPC-EPC integration problem for a process described by (17). These authors evaluated several combinations of SPC detection, estimation, and adjustment mechanisms. From extensive simulation studies, it was concluded that the best of the integrated approaches tried from a mean squared deviation point of view was a an integrated CUSUM/harmonic adjustment approach. This works by using a CUSUM to detect a shift and to estimate its magnitude. Once the CUSUM signals, a harmonic controller (8) starts to operate from that point on, i.e.,

$$x_t = \begin{cases} 0 & \text{for } t < t_0 \text{ and } t > t_0 + 5; \\ -\widehat{\mu}_{t_0} & \text{for } t_0; \\ x_{t-1} - \frac{1}{t-t_0} Y_t & \text{for } t = t_0 + 1, t_0 + 2, ..., t_0 + 5 \end{cases}$$

For a process like Shewhart's, more than five sequential adjustments were observed to be unnecessary. The mean squared deviation from target is minimum with this combined approach.

Despite the simplicity and excellent performance of this approach under the stated assumptions, which happen to be true in many real discrete-manufacturing processes, it is a method not known in the Industrial Statistics community. We therefore elaborate on its use in what follows. Figure 1 illustrates the CUSUM-harmonic rule integrated approach. A shift of magnitude $d = 2$ was simulated at time $t = 7$. In the example data shown, the CUSUM chart detects the shift at time $t_0 = 10$ and triggers the harmonic rule, setting $x_{t_0} = -\widehat{\mu}_{t_0} = -11.84$ and $\nabla x_t = -1/(t - t_0)$ for $t > t_0$. Note how the process returns to target, and how $x_t \to -d$, providing an estimate of the shift, which is a Stochastic Approximation estimate of it (see Pan and Del Castillo (2003) for more details).

Thus, in this alternate SPC/EPC integration approach, an EPC mechanism is invoked only when it is necessary. This alternative resembles a deadband controller and the "machine tool" problem, but the assumed disturbances and motivations are different. The machine tool problem assumes an IMA(1,1) disturbance, which is well-known to be optimally forecasted through the EWMA in eq. (15). There, the fixed cost of adjusting implies the deadband structure of the solution; in the integrated SPC-EPC approaches the SPC acts as a deadband since no other disturbance is supposed to exist between shift detection times. Interestingly, as p, the probability of a shift in any time point, increases, then the corresponding stochastic process for y_t increasingly resembles an IMA(1,1) process (which in itself can be thought of as a random walk observed with error). This implies that when p is large, simply using an EWMA controller based on (15) without a deadband will work better from a mean squared deviation point of view than the type of integrated CUSUM/harmonic rule described here. This was also noted by Chen and Elsayed (2002), who studied how to tune an EWMA controller $x_t = -a_t/g$ where a_t is an EWMA of the y_t's, when the disturbances follow the step-like process described by (17). From our description above, the IMA(1,1) will be an increasingly better model and a controller based on it will be increasingly closer to optimal

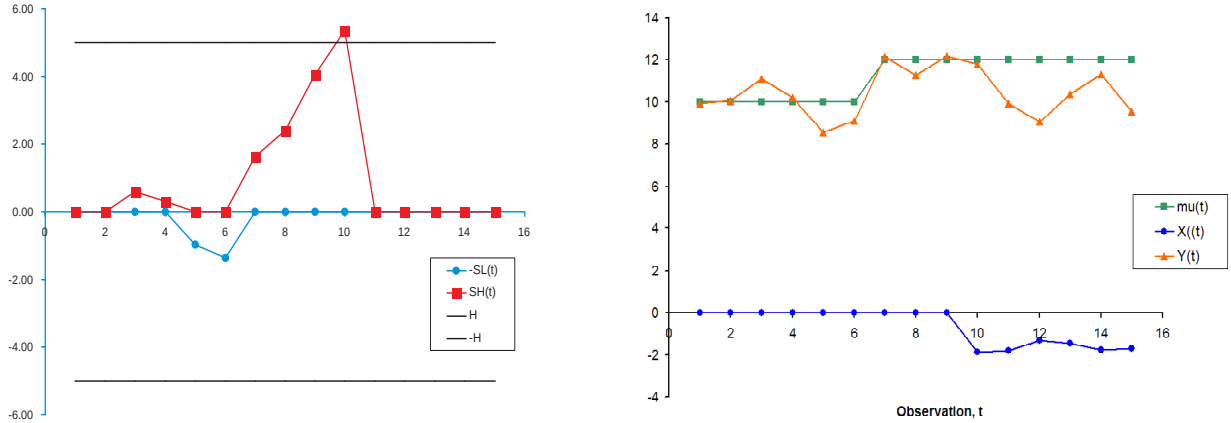

Figure 1: CUSUM-harmonic rule integrated approach. Left: CUSUM chart; right: observed and mean quality characteristic (top) and controllable factor values (bottom).

from a MSE point of view as p increases. Box and Luceño (1997, Chapter 5), showed that the IMA(1,1) model is in fact a good approximation to the random jump model even for relative low values of p. Despite this fact, if one always uses an EWMA controller based on an IMA(1,1) model, monitoring for the eventual elimination of assignable causes will be a task harder to accomplish.

Note how the "machine tool" deadband does not include a monitoring scheme, thus assignable casue *removal* is not possible using a deadband scheme (despite its resemblance of an SPC chart).

4.5 Closed loop identification

In control theory, identification is the combined efforts of finding the structure of a model and of the parameters of a given structure based on input-output data. For a comprehensive presentation of systems ID from a Control Theoretic point of view, see Ljung (1999). Here we wish to mention one important problem that arises frequently in quality control practice.

Classical methods for identification in single input, single output (SISO) processes are presented by Box et al. (1994). These methods are based on the cross-correlation between the $\{x_t\}$ and the $\{y_t\}$ series (perhaps after "prewhitening" them). This is because *if the input series* $\{x_t\}$ *is uncorrelated with the noise series* $\{\varepsilon_t\}$, the $x - y$ crosscorrelations are proportional to the impulse response weights w_j in

$$y_t = w_0 x_t + w_1 x_{t-1} + w_2 x_{t-2} + ...$$

This allows to identify the true structure of a Box-Jenkins transfer function model, from which various numerical methods based on ML estimation can be applied to find the parameter estimates in (1). If the process was controlled while the time series data were collected – i.e., if the feedback loop was closed– then $\{x_t\}$ and $\{\varepsilon_t\}$ will be correlated, the cross-correlation may not contain useful information and no identification (ID) can take place. Identification means to arrive at a model structure which is equivalent to the true model structure, with parameter estimates that converge to the true parameters asymptotically. Knowing under which conditions identifiability problems occur and how to avoid them is called the closed loop ID problem.

Closed loop ID is important in practice since many processes are open-loop unstable, hence there is an incentive to control and identify at the same time, rather than running open loop experiments which may be very expensive, typically with the ultimate goal of improving the controller used while the data was collected.

Akaike (1967) was one of the first to look at this problem. He shows that if the process model is $y_t = G(\mathcal{B})x_t + H(\mathcal{B})\varepsilon_t$ and the controller is $x_t = C(\mathcal{B})y_t$, then identification based on the cross-spectrum (and hence, on the cross-correlation of the input and output) will yield the inverse of the controller, i.e., $\widehat{G} = 1/C$. He suggests (see also Box and MacGregor (1974,1976) and MacGgregor and Fogal (1995)) to add an external *dither* signal that breaks the linear relation between the $\{x_t\}$ series and the $\{\varepsilon_t\}$ series, i.e.:

$$x_t = C(\mathcal{B})y_t + d_t.$$

Akaike shows that the estimated transfer function will then be a weighted average of the true transfer function and the transfer function of the inverse of the controller, with the weight given to the former being proportional to the signal to noise ratio σ_d^2/σ_N^2 ($\sigma_N^2 = \mathrm{Var}(H(\mathcal{B})\varepsilon_t)$).

The disadvantage of the dither approach is clear. As σ_d^2/σ_N^2 increases, the process becomes easier to identify, but it becomes less controlled. Many authors have searched for alternative approaches. Söderström et al. (1976) derive identifiability conditions when $r(>1)$ feedback controllers operate in parallel, and the control action alternates between each of the controllers after some runs. The resulting controller is non-linear, and the linear dependency of the input on the noise series is broken. This explains indirectly why adaptive controllers can often avoid identifiability problems, since in them the parameter estimates change continuously, so they are non-linear controllers.

Söderstrom et al. (1975) derive necessary and sufficient conditions for system identifiability for an ARMAX model of the form:

$$A(\mathcal{B})y_t = B(\mathcal{B})x_{t-k} + C(\mathcal{B})\varepsilon_t$$

controlled with $F(\mathcal{B})x_t = G(\mathcal{B})y_t$. The necessary and sufficient conditions, under the assumption that the fitted model contains the true system structure as a particular case, are

that

$$\max(n_F + n_A, k + n_G + n_B) - n_P \geq n_A + n_B \tag{18}$$

where $P(\mathcal{B})$ is the factor common to the C and $AF - BG\mathcal{B}^k$ polynomials, and n_M denotes the order of polynomial $M(\mathcal{B})$. This expression says that the number of linearly independent equations, after cancelling any common terms, must be at least equal to the number of unknowns. Box and MacGregor (1976) applied this result to Box-Jenkins models controlled with a MMSE controller. It is interesting to note that the SI condition (18) will tend to hold as the order of the controller increases. This gives evidence against using simple controllers when collecting data for closed loop ID. It will also tend to hold for processes with large input-output delay, but such processes are inherently harder to control. In addition, the results obtained by Box MacGregor (1976) indicates that MMSE-controlled processes will be harder to identify than non-MMSE controlled processes, and in some cases, the former will be completely not identifiable.

Since identifiability in this context is not well-known among Statisticians, we now illustrate what does lack of identifiability entails in practice. Consider a process described by the model:

$$Y_t = 20 + \frac{15\mathcal{B}^2}{1 - 0.8\mathcal{B}} x_t + \frac{5.0}{1 - \mathcal{B}} + \frac{1 - 0.3\mathcal{B}}{1 - \mathcal{B}} \varepsilon_t \tag{19}$$

which has 1^{st} order dynamics, 2 unit time delay, and an IMA(1,1) with drift noise disturbance. This is the true description of the process. If a PI controller with parameters $K_P = -0.01$ and $K_I = 0.015$ is applied to this process, the closed-loop equation is:

$$(1 - 1.8\mathcal{B} + 0.725\mathcal{B}^2 + 0.15\mathcal{B}^3)Y_t = 1.0 + (1 - 1.1\mathcal{B} + 0.24\mathcal{B}^2)\varepsilon_t \tag{20}$$

which is an ARMA(3,2) process. This is the closed loop equation of the true process description. Let us assume this is the model we actually fit, so in this analysis we neglect sampling variability. This will show that the identifiability conditions are mathematical, not statistical, i.e., they do not depend on the data. Suppose we are unaware of the true process model, so we fit a Box-Jenkins model (1) with orders $n_{A'} = 2$, $n'_B = 5$, and a noise disturbance (second term in equation (1)) equal to an ARIMA(2,d,2) with drift model, i.e.:

$$N_t = \frac{C'(\mathcal{B})}{(1 - \mathcal{B})^d \phi(\mathcal{B})} \varepsilon_t + \frac{\delta}{1 - \mathcal{B}}.$$

If a PI controller of the form

$$x_t = \frac{c_1 - c_2\mathcal{B}}{1 - \mathcal{B}} y_t$$

(where $c_1 = K_P + K_I$ and $c_2 = -K_P$ in this alternative parametrization) is applied, the closed loop equation is:

$$\phi(\mathcal{B})[A'(\mathcal{B})(1 - \mathcal{B}) - B'(\mathcal{B})(c_1\mathcal{B} + c_2)]y_t = (1 - \mathcal{B})^{1-d}\phi(\mathcal{B})A'(\mathcal{B})\delta + (1 - \mathcal{B})^{1-d}C'(\mathcal{B})A'(\mathcal{B})\varepsilon_t. \tag{21}$$

Sol	a_1	a_2	b_1	b_2	b_3	b_4	b_5	ϕ_1	ϕ_2	θ_1	θ_2	δ	Obj
1	0.80	0.00	0.07	-14.87	-0.05	-0.01	0.00	0.00	0.00	0.30	0.00	5.00	3.76E-12
2	1.10	-0.24	7.36	-0.25	-3.82	0.52	-0.05	-0.34	-0.03	0.00	0.00	5.22	2.83E-10
3	0.80	0.00	-0.40	-15.73	0.30	0.06	0.00	0.00	0.00	0.30	0.00	4.99	1.72E-09
4	0.81	-0.01	3.42	-8.73	-2.67	-0.20	-0.08	-0.03	0.03	0.29	0.00	5.09	2.36E-09
5	**0.81**	**-0.01**	**2.61**	**-10.20**	**-1.96**	**-0.20**	**-0.04**	**-0.03**	**0.03**	**0.29**	**0.00**	**5.07**	**1.83E-08**

Table 1: Five best solutions obtained from minimizing the sum of squared errors of the system of equations obtained from (20) and (21). Note that the true system (bold) does not correspond to the lowest objective function value, a consequence of lack of identifiability.

To see if the process is identifiable under the actions of the PI controller, we can solve the system of equations that result from equating coefficients of like terms in the polynomials of $\mathcal{B}$ found in (21) and (20). One way to do this is to minimize the sum of squared errors with respect to the parameters $b_1, ..., b_5$, a_1, a_2, ϕ_1, ϕ_2, and θ_1, θ_2 (d can be inferred by looking at the unit roots of the MA polynomial; in this case there is no unit root in (20) so we set $d = 1$ in comparison to (21)). If *for the global optimum of the sum of squared errors function* the values of these parameters equal the true values in (19), we will have system identifiability.

Table 1 shows the five best solutions obtained from minimizing the sum of squared error function from a set of 2000 random starting solutions. As it can be seen, the true solution (in bold) does not correspond to the best objective function value found. Thus, when identifiability conditions (18) do not hold, this implies that the global minimum of the sum of squared errors function will not coincide with the true description of the system. This behavior can only get worse if the model is fit from noisy data, so evidently this result applies in general.

It is important to note that even though the SI conditions may be satisfied in practice, the precision of the parameter estimates based on finite samples may be very poor. The addition of a dither signal will improve the precisions (MacGregor and Fogal, 1995). Other approaches that incorporate additional information are possible. For example, stationarity and invertibility conditions can be added to enhance the identifiability properties (Pan and Del Castillo, 2001, Del Castillo, 2002b).

5 Some areas for further research

This section provides a discussion of two SPA areas where further work would be of benefit: Bayesian process adjustment methods using SMC techniques and context-based EWMA control for application in semiconductor manufacturing.

5.1 Bayesian methods in process adjustment

We have already referred to recent SPA methods that are Bayesian, such as setup adjustment using MCMC techniques (Colosimo et al., 2004), deadband schemes for process adjustments (Lian and Del Castillo, 2005), and cautious control (Apley and Kim, 2004). In this section we further comment on the potential of modern Bayesian statistical techniques in SPA and some areas open for research. Well-known control theory techniques have connections with Bayesian techniques or can be interpreted in a Bayesian way, two examples being Kalman filtering for state estimation (with known parameters) and Adaptive Control. The main potential for new Bayesian SPA methods, much along the type of problems discussed earlier, is on the adjustment of short-run processes with unknown parameters.

Breakthroughs in numerical integration developed over the last 15 years can now be routinely utilized for posterior inference when non-conjugate priors are desired. In particular, MCMC methods (Gelman et al., 2003) have been developed intensively and proved to provide solutions to previously untractable problems.

For a problem in which data arrives sequentially in time, however, MCMC methods may not be the best choice. In MCMC, Markov Chains iterations yielding the target posterior distribution are repeated from scratch every time a single new observation y_{t+1} is obtained, without reusing the posterior distribution previously obtained a period before, i.e., at period t. An alternative to MCMC is Sequential Monte Carlo (SMC) methods (see Figure 2). SMC methods also rely on Monte Carlo algorithms for the solution of Bayesian inference problems. In SMC, posterior distributions of "particles" $\theta^{(i)}$ (values of the parameter) are created numerically from calculating associated weights w_i. These weights are recomputed after each observation is obtained based on the likelihood of the corresponding particle given the new datum and the previous set of weights, keeping in this way information from the previous step. The weights are then used to provide posterior estimates of any function of the parameter of interest at time $t+1$. A major advantage of SMC techniques is that they are considerably faster than MCMC, allowing for on-line control. A brief sketch of the computations required to approximate the expectation of some function of an unknown parameter θ at step i is as follows:

1. Draw $\theta^{(j)} \sim \pi(\theta)$, $j = 1, 2, ..., M$, and set $w_j = 1/M$;
2. At each step i we do the following:
   (a) Update $w_j \leftarrow w_i \times L(\theta^{(j)}|y_i)$; normalize weights such that $\sum w_j = 1$;
   (b) If sample degenerates, perform a "rejuvenation" step and resample $\theta^{(j)}$, $j = 1, ..., M$ using importance sampling based on the w_j's;
3. Compute $E[f(\theta)] \approx \sum_{i=1}^{M} w_i \times f(\theta^{(i)})$, increase i and goto 2 unless end of data.

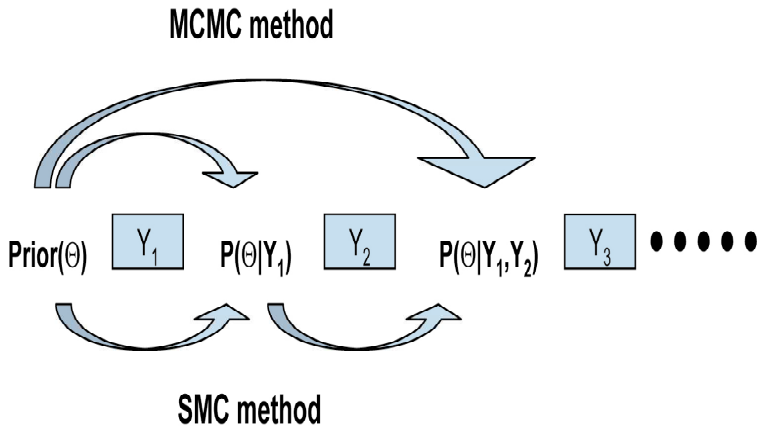

Figure 2: The update of a posterior distribution in MCMC and SMC.

Here $L(\theta^{(j)}|y_i)$ is the likelihood function of the jth particle given the latest observation. If, e.g., interest is in the sample mean, then $f \equiv 1$. The *rejuvenation step* is executed if the sample of particles is too poor. This will tend to happen when $\pi(\theta)$ is a non-informative prior. In such case, many particles will be unlikely given the data, so their weights w_i will be zero after a few iterations. The distribution of the w_i's will contain only a few non-zero weights, and will provide biased estimates. A rejuvenation step (Balakrishnan and Madigan, 2004) smooths the posterior distribution of the particles. Then, importance resampling of the parameters θ is performed using the updated weights. See Doucet et al. (2001) for more details.

Lian et al. (2005) apply the SMC method to the setup adjustment problem for unknown parameters. They show how SMC gives results equivalent to MCMC but a fraction of the computing time. The bayesian deadband adjustment scheme mentioned earlier (Lian and Del Castillo, 2005) also utilizes SMC. There is a wide range of other relevant control problems with unknown parameters that could be approached with SMC techniques. This includes adaptive filtering problems and, in general, State-Space models. The SMC procedure provides posterior parameter distributions of any relevant parameters which in turn can be used to minimize a variety of cost functions. The solutions so obtained will in general be suboptimal since the certainty equivalence principle (which indicates when using plug-in estimates leads to optimal solutions, see Del Castillo, 2002a, Appendix 8B) applies only in restrictive cases. Nevertheless, the solutions obtained may still have excellent performance considering that a "dual control" optimal solution is computational prohibitive in most cases (see Åström and Wittenmark, 1989). In a given application of SMC to process adjustment, additional work is needed to quantify its performance over some known lower bound or reference point of performance.

A second area we would like to highlight where Bayesian inference can play an important role is in Closed-loop identification. If lack of identifiability is a problem due to not having enough information about the process parameters, it seem natural to use a Bayesian approach in which any prior information available can be incorporated. How to determine such prior(s) and what type of additional pieces of information one should be able to model with the priors are questions open for future research. MCMC methods used for open and closed loop identification have recently been mentioned by Ninness et al. (2002) and Thil and Gilson (2004), respectively.

5.2 Context-based EWMA control

EWMA (and PI) controllers have been studied extensively from a statistical point of view in the last decade. This body of work concerns mostly the application of a single EWMA controller to one manufacturing process, usually in a run-to-run control setting (i.e., the within-run variation is not modelled). In this section we would like to point out some complexities of realistic EWMA control as used for run to run control in the semiconductor industry, created by the complexity of the operation, but approachable due to the data-rich environment in which they are applied.

As pointed out by Braun et al. (2003), run to run control should deal with the context information associated with each run. Thus, for example, we may know that a particular run at time t will start in tool i by operator j in order to perform task k. The usual model in EWMA control simply considers offsets (or "biases") $b_t (= N_t$ in (16)), that is:

$$y_t = g x_{t-1} + b_t. \tag{22}$$

The availability of context information implies the quality characteristic can be modelled instead as

$$y_{ijk,t} = g_{ijk} x_{ijk,t-1} + b_{ijk,t}. \tag{23}$$

In equation (22), the estimate of the offset (the predictor of the disturbance) is simply

$$\widehat{b}_{t+1|t} = \lambda(y_t - \widehat{g} x_{t-1}) + (1-\lambda)\widehat{b}_{t-1}$$

where $\widehat{g}$ is an off-line gain estimate. The controller is simply $x_t = -\widehat{b}_t/\widehat{g}$, assuming y denotes deviations from target.

With the context information, managing all the EWMA controllers poses an interesting challenge. At one extreme of simplicity, one could have *complete segregation* of controllers (Braun et al. 2003). Thus,

$$\widehat{b}_{ijk,t+1|t} = \lambda_{ijk}(y_{ijk,t} - \widehat{g}_{ijk} x_{ijk,t-1}) + (1-\lambda_{ijk})\widehat{b}_{ijk,t|t-1}.$$

Evidently, if there are I tools, J operators, and K different tasks, this will imply managing up to IJK EWMA predictors, each with its corresponding weight parameter λ_{ijk}.

Another possibility is to assume that there are different offset effects due to tool, operation, and tasks that act additively:

$$\widehat{b}_{ijk,t+1|t} = \widehat{\alpha}_{i,t+1|t} + \widehat{\beta}_{j,t+1|t} + \widehat{\gamma}_{k,t+1|t} \tag{24}$$

where α is the tool offset effect, β is the operator offset effect, and γ is the task offset effect. This case will require $I+J+K$ EWMA predictors to handle, typically with different weights. For example, a tool may drift, while operator and task effects will probably remain stable over time. This would suggest that $\lambda_i > \max(\lambda_j, \lambda_k)$ is indicated. Braun et al. (2003) suggest to use

$$\widehat{\alpha}_{i,t+1|t} = \lambda_i(y_{ijk,t} - \widehat{g}_{ijk}x_{ijk,t-1} - \widehat{\beta}_{j,t-1} - \widehat{\gamma}_{k,t-1}) + (1-\lambda_i)\widehat{\alpha}_{i,t-1} \tag{25}$$

and similarly for $\widehat{\beta}_{j,t+1|t}$ and $\widehat{\gamma}_{k,t+1|t}$. Models with non-linear effects (interactions) are also possible, in which, for example:

$$\widehat{b}_{ijk,t+1|t} = \widehat{\alpha}_{i,t+1|t} + \widehat{\beta}_{j,t+1|t} + \widehat{\gamma}_{k,t+1|t} + \widehat{(\beta\gamma)}_{jk,t+1|t} \tag{26}$$

where

$$\widehat{(\beta\gamma)}_{jk,t+1|t} = \lambda_{jk}(y_{ijk,t} - \widehat{g}_{ijk}x_{ijk,t-1} - \widehat{\alpha}_{i,t|t-1} - \widehat{\beta}_{j,t|t-1} - \widehat{\gamma}_{k,t|t-1}) + (1-\lambda_{jk})\widehat{(\beta\gamma)}_{jk,t|t-1}$$

and so on.

The similarities between (24) or (26) and an effects model used in ANOVA are obvious. Complete segregation corresponds to an analysis in which the effect of each cell ijk is tracked individually with an EWMA. If considerable historical information is available in each cell, then this may work well, although there is the problem of handling that many EWMA predictors. We see that (24) and (26) correspond to effects models in which the effects are computed in a non-standard way via ad-hoc expressions such as (25), using EWMA's (to allow for more rapid response instead of using non-weighted averages), assuming no constant term, and without the usual restrictions $\sum \alpha_i = 0, \sum \beta_j = 0, \sum \gamma_k = 0$, etc., added in ANOVA to obtain estimable effects (Milliken and Johnson, 1984). Broun et al. also tried recursive least squares to estimate all the offset effects. Different initial values of the parameter vector and the covariance matrix will lead to different solutions of the system of equations. It is not clear how to choose these initial values to achieve best control, apart from simulation experiments.

An open question in this context-dependent control scenario is how to select all the EWMA weights. This depends on the particular model assumed (either complete segregation, additive, or non-linear effects) and implies there is a combinatorial problem (finding the best effects model) coupled with a continuous optimization problem (finding the λ's).

The objective of this mixed discrete-continuous problem is to minimize the mean square deviations of all quality characteristics involved.

Related to this problem is to find overall system stationarity conditions for a set of controlled processes, given context information. This will involve both the EWMA weights and the gains and their estimates.

Another challenge refers to measurement delays. Usually, the order in which runs are measured is not the same as the order in which they were produced, implying that variable measurement delays are present. Developing tuning methods for EWMA controllers in the presence of uncertainty in the input-output delay is an open problem for further work.

6 Conclusion

In this paper, a view of the origins, present status, and a discussion of some areas for further research on Statistical Process Adjustment methods was given. The goal was to provide convincing examples that would demonstrate the intellectual and practical value of this field of Industrial Statistics, and to promote interest for further research.

Acknowledgements. While writing this paper the author benefitted from discussions with Drs. Mani Janakiram and Ramkumar Rajagopal (Intel Corporation), who he thanks. Thanks also to O. Arda Vanli (Penn State) for Table 1 and to Zilong Lian (Penn State) for Figure 2.

7 Bibliography

Akaike, H., (1967). "Some Problems in the Application of the Cross-Spectral Method" in *Advanced Seminar on Spectral Analysis of Time Series*, Harris, New York: Wiley.

Apley, D.W. (2004). "A cautious minimum variance controller with ARIMA distrubances". *IIE Transactions*, 36, pp. 417-432.

Apley, D.W., and Kim, J. (2004). "Cautious control of Industrial Process Variability With Uncertain Input and Disturbance Model Parameters", *Technometrics*, 46(2), pp. 188-199.

Åström, K.J., (1970). *Introduction to Stochastic Control Theory.* Academic Press, San Diego, Cal.

Åström, K.J., and Wittenmark, B. (1989). *Adaptive Control.* Reading, Mass: Addison Wesley.

Balakrishnan S., Madigan D. (2004). "A One-Pass Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets". Technical Paper, http://www.stat.rutgers.edu/~madigan/papers/.

Barnard, G.A. (1959). "Control Charts and Stochastic Processes". *Journal of the Royal Statistical Society, B*, 21 (2), pp. 239-271.

Bishop, A.B., (1965). "Automation of the Quality Control Function", *Industrial Quality Control*, 21, pp. 509-514.

Box, G.E.P. and Jenkins, G.M. (1962). "Some statistical aspects of adaptive optimization and control", *Journal of the Royal Statistical Society, B*, 24(2), pp. 297-343.

Box, G.E.P. and Jenkins, G.M. (1963). "Further contributions to adaptive quality control: simultaneous estimation of dynamics: no-zero costs", *ISI Bulletin*, 34th session, Ottawa, Canada, pp. 943-974.

Box G.E.P., G.M. Jenkins, and Reinsel, G. (1994). *Time Series Analysis, Forecasting, and Control* 3rd. ed., Englewood Cliffs: Prentice Hall (1994).

Box, G.E.P., and Luceño, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*. John Wiley & Sons, New York, NY.

Box, G.E.P., and MacGregor, J.F. (1974). "The Analysis of Closed-Loop Dynamic Stochastic Systems", *Technometrics*, 16, 3, pp. 391-398.

Box, G.E.P., and MacGregor, J.F. (1976). "Parameter Estimation With Closed-Loop Operating Data", *Technometrics*, 18, 4, pp. 371-380.

Braun, M.W., Jenkins, S.T., and Patel, N.S., (2003). "A Comparison of Supervisory Control Algorithms for Tool/Process Disturbance Tracking", *Proceedings of the American Control Conference*, Denver, CO, pp. 2626-2631.

Butler, S.W. and Stefani, J.A. (1994). "Supervisory Run-to-Run Control of a Polysilicon gate Etch Using In Situ Ellipsometry," *IEEE Transactions on Semiconductor Manufacturing*, 7, 2, 193201.

Chen, A., and Elsayed, E.A., (2002). "Design and Performance Analysis of the Exponentially Weighted Moving Average Mean Estimate for Processes Subject to Random Step Changes," *Technometrics*, 44(4), pp. 379-389.

Clarke, D.W., and Gawthrop, P.J. (1975). "Self-Tuning Controller", *Proceedings of the In-*

stitution of Electrical Engineers, 122, 9, pp. 929-934.

Colosimo, B. M., Pan, R., and Del Castillo, E. (2004). "A Sequential Markov Chain Monte Carlo Approach to Setup Process Adjustment Over a Set of Lots", *Journal Applied Statistics*, 31 (5), pp. 499-520.

Colosimo, B.M., Pan, R., and Del Castillo, E. (2005). "On-line Process Adjustment for Asymmetric Cost Functions,", *International Journal of Production Research*, 43(18), pp. 3837-3854.

Crowder, S.V. (1992). "An SPC Model for Short Production Runs: Minimizing Expected Cost," *Technometrics*, *34*, 64-73.

Crowder, S.V., and Eshleman, L., (2001). "Small Sample Properties of an Adaptive Filter Applied to Low Volume SPC", *Journal of Quality Technology*, 33, 1, pp. 29-38.

Del Castillo, E. (1998). "A Note on two Process Adjustment Models," *Quality & Reliability Engineering International,* 14, 23-28.

Del Castillo, E., (1999). "Long-run and Transient Analysis of a Double EWMA Feedback Controller", *IIE Transactions*, 31, 12, pp. 1157-1169.

Del Castillo, E. (2000). "A Variance Constrained PI Controller That Tunes Itself," *IIE Transactions*, 32, 6, pp. 479-491.

Del Castillo, E. (2001). "Some Properties of EWMA Feedback Quality Adjustment Schemes for Drifting Processes", *Journal of Quality Technology*, 33(2), pp. 153-166.

Del Castillo, E. (2002a). *Statistical Process Adjustment for Quality Control*, New York: John Wiley & Sons (*Probability and Statistics Series*).

Del Castillo, E. (2002b). "Closed-loop Disturbance Identification and Controller Tuning for Discrete Manufacturing Processes," *Technometrics*, 44, 2, pp. 134-141.

Del Castillo, E., Pan, R., and Colosimo, B.M. (2003). "A Unifying View of Some Process Adjustment Methods", *Journal of Quality Technology*, 35, 3, pp. 286-293.

Del Castillo, E., and Rajagopal, R., (2002). "A Multivariate Double EWMA Process Adjustment Scheme for Drifting Processes", *IIE Transactions*, 34 (12), pp. 1055-1068.

Deming,W.E., (1982). *Out of the Crisis.* Cambridge, MA: MIT Center for Advanced Engineering Study.

Doucet A., de Freitas N., Gordon N. (2001). *Sequential Monte Carlo Methods in Practice*, Springer-Verlag New York, Inc.

Faltin, F.W., Hahn, G.J., Tucker, W.T., and Vander Weil, S.A. (1993). "Algorithmic statistical process control: some practical observations," *International Statistical Institute*, 61, pp. 67-80.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., (2003). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC.

Grubbs, F.E., (1954). "An Optimum Procedure for Setting Machines or Adjusting Processes," *Industrial Quality Control*, July 1954, reprinted in *Journal of Quality Technology*, 1983, 15, 4, pp. 186-189.

Guo, R.S., Chen, A., and Chen, J.J., (2000). "An Enhanced EWMA Controller for Processes Subject to Random Disturbances", in Moyne, J., Del Castillo, E., and Hurwitz, A., eds., (2000), *Run to Run Process Control in Semiconductor Manufacturing*, CRC Press, Boca Raton, FL.

Ingolfsson A. and E. Sachs (1993). "Stability and Sensitivity of an EWMA Controller," *Journal of Quality Technology*, 25, 4, pp. 271-287

Jensen, K.L., and Vardeman, S.B., (1993). "Optimal Adjustment in the Presence of Deterministic Process Drift and Random Adjustment Error". *Technometrics*, 35, pp. 376-389.

Jiang, E., and Tsui, K-L., (2002), "SPC Monitoring of MMSE- and PI-Controlled Processes", *Journal of Quality Technology*, 34(4), pp. 384-398.

Lian, Z., Colosimo, B.M., and Del Castillo, E., (2005a). "Setup Adjustment of Multiple Lots Using a Sequential Monte Carlo Method", accepted in *Technometrics*.

Lian, Z., Colosimo, B.M., and Del Castillo, E. (2005b). "Setup Error Adjustment: sensitivity analysis and a new MCMC control rule", to appear in *Quality & Reliability Engineering International*.

Lian, Z. and Del Castillo, E. (2004). "Setup Adjustment Under Unknown Process Parameters and Fixed Adjustment Cost" accepted in *Journal of Statistical Planning and Inference*.

Lian, Z., and Del Castillo, E., (2005). "Adaptive Deadband Control of a Drifting Process With Unknown Parameters", submitted to *Statistics & Probability Letters*.

Ljung, L. (1999). *System Identification: Theory for the User*, 2nd. ed. Upper Saddle River,

NJ: Prentice Hall.

Luceño, A. (2003). "Dead-band Adjustment Schemes for On-line Feedback Quality Control", in Handbook of Statistics, Vol. 22, R. Khattree and C.R. Rao, eds., Elsevier Science B.V.

Luceño, A., and González, F.J., (1999). "Effects of Dynamics on the Properties of Feedback Adjustment Schemes with Dead Band". *Technometrics*, 41, pp. 142-152.

MacGregor, J. F., (1988). "On-Line Statistical Process Control", *Chemical Engineering Progress*, October, pp. 21-31.

MacGregor, J.F., 1990. "A Different View of the Funnel Experiment," Journal of Quality Technology, 22, pp. 255-259.

MacGregor, J.F., and Fogal, D.T. (1995). "Closed-loop identification: the role of the noise model and prefilters", *Journal of Process Control*, 5(3), pp. 163-171.

Milliken, G.A. and Johnson, D.E. (1984). *Analysis of Messy Data*. New York: Van Nostrand Reinhold.

Morari, M., and Zafiriou, A. (1989). *Robust Control*. Prentice Hall, Englewood Cliffs, NY.

Moyne, J., Del Castillo, E., and Hurwitz, A., eds., (2000). *Run to Run Process Control for Semiconductor Manufacturing*, CRC Press.

Ninness B., Henriksen, S., and Brinsmead, T. (2002). "System Identification via a Computational Bayesian Approach", *Proceedings of the 4st IEEE Conference on Decision and Control*, Las Vegas, NE, pp. 1820-1825.

Pan, R., and Del Castillo, E. (2001). "Identification and Fine Tuning of Closed-loop Processes under Discrete EWMA and PI Adjustments", *Quality and Reliability Engineering International*, 17, pp. 419-427.

Pan, R., and Del Castillo, E. (2003). "Integration of Sequential Process Adjustment and process Monitoring techniques," *Quality & Reliability Engineering International*, 19,4 pp. 371-386.

Pan, R. and Del Castillo, E. (2004). "Scheduling Methods for the Setup Adjustment problem," *International Journal of Productions Research*, 41,7, pp. 1467-1481, (2003). Correction, 42(1), pp. 211-212.

Sachs, E., Hu, A., and Ingolfsson, A., (1995). "Run by Run Process Control: Combining

SPC and Feedback Control", *IEEE Transactions on Semiconductor Manufacturing*, 8,1, pp. 26-43.

Söderström, T., Gustavsson, I. and Ljung, L. (1975). "Identifiability conditions for linear systems operating in closed loop", *International Journal of Control*, 21(2), pp. 243-255.

Söderström, T., Ljung, L. and Gustavsson, I. (1976). "Identifiability conditions for linear multivariate systems operating under feedback", *IEEE Transactions on Automatic Control*, 21(6), pp. 837-840.

Thil, S. and Gilson, M. (2004). "Closed loop identification: a bayesian approach", Working paper, Centre de Recherche en Automatique de Nancy (CRAN).

Trietsch, D., (1998). "The Harmonic Rule for Process Setup Adjustment with Quadratic Loss", *Journal of Quality Technology*, 30, 1, pp. 75–84.

Trietsch, D. (2000). "Process setup adjustment with quadratic loss". *IIE Transactions*, 32(4), pp. 299-307.

Tseng, S.-T., Chou, R.-J., and Lee, S.-P. (2002). "Statistical Design of double EWMA Controller". *Applied Stochastic Models in Business and Industry*, 18, pp. 313-322.

Tseng, S.-T., and Hsu, N.-J. (2005). "Sample Size Determination for Achieving Asymptotic Stability of a Double EWMA Control Scheme" *IEEE Transactions on Semiconductor Manufacturing*, 18, 1, pp. 104-111.

Tsung, F., and Shi, J.J. (1999). "Integrated Design of Run to Run PID Controllers and SPC Monitoring For Process Disturbance Rejection," *IIE Transactions*, 31(6), pp. 517-527.

Tsung, F., and Tsui, K.-L. (2003). "A mean-shift pattern study on the integration of SPC and APC for process monitoring", *IIE Transactions*, 35, pp. 231-242.

Tsung, F., Wu, H., and Nair, V., (1998), "On the Efficiency and Robustness of Discrete Proportional-Integral Control Schemes", *Technometrics*, 40, 3, 214-222.

Tucker, W.T., Faltin, F.W., and Vander Wiel, S.A. (1993). "ASPC: an ellaboration", Technometrics, 35, 4, pp.363-375.

Vander Wiel S.A., Tucker, W.T., Faltin, F.W., and Doganaksoy, N. (1992). "Algorithmic Statistical Process Control: Concepts and an Application", Technometrics, 34, 3, pp. 286-297.

Yang, J., and Makis, V. (2000). "Dynamic Response of Residuals to External Deviations in a Controlled Production Process", *Technometrics*, 42(3), pp. 290-299.