Model Context Selection for Run-to-Run Control

O. Arda Vanli[*], Nital S. Patel[†], Mani Janakiram[‡]

Intel Corporation, Chandler, AZ 86226

and Enrique Del Castillo[§]

The Pennsylvania State University, University Park, PA 16802

Abstract

In the design of run to run controllers one is usually faced with the problem of selecting a model structure that best explains the variability in the data. The variable selection problem often becomes more complex when there are large numbers of candidate variables and the usual regression modeling assumptions are not satisfied. This paper proposes a model selection approach that uses ideas from the statistical linear models and stepwise regression literature to identify the context variables that contribute most to the autocorrelation and to the offsets in the data. A simulation example and an application on lithography alignment control are presented to illustrate the approach.

Keywords: Run to run control, analysis of variance, ARIMA time series models, context selection.

1 Introduction

In semiconductor manufacturing, it is often common to find the same tool processing different types of products and operations. The particular combination of different factors related to

[*] omer.a.vanli@intel.com

[†] Corresponding author: nital.s.patel@intel.com, 5000 W. Chandler Blvd CH3-84, Chandler, AZ 85226

[‡] mani.janakiram@intel.com

[§] exd13@psu.edu

the batch, such as the product, operation, chamber and machine, is defined as the "context" of a batch, and the factors that define the context of the batch are called the context variables. Very often, run to run control has to account for this variation by using context-dependent models. The aim of this paper is to present a programmatic approach for identifying the context variables that best express process data for modeling purposes.

In production, different products and operations are processed in a random order and the batches which often have dissimilar contexts causes the parameters of the process model to be non-homogeneous across different contexts. Run to run control (see e.g. [8]), a class of control schemes developed for semiconductor manufacturing, takes into account this variability in the process parameters by defining a different process model for each context for calculating the recommended process settings from run to run.

Consider a single input single output process model

$$z_t = gu_t + y_t \tag{1}$$

which is a pure-gain transfer function model, where z is the output and u is the input of the process. The constant process gain g is assumed to be known. y_t is the stochastic disturbance which models the remaining variability not modeled by the transfer function. The input is observable, thus, the disturbance can be written as $y_t = z_t - gu_t$. Autoregressive-integrated-moving average (ARIMA) time series models [1] are commonly used to model stochastic disturbances.

Suppose that there are k different context variables that define the process model. Context variables are categorical, therefore, the disturbance can be represented using a k-way analysis of variance (ANOVA) model. The full model which contains the main effects and the two factor interactions is:

$$y_{ij...l,t} = \mu + \tau_i + \beta_j + \ldots + \gamma_l + (\tau\beta)_{ij} + ... + (\tau\gamma)_{il} + e_{ij...l,t} \tag{2}$$

where, τ_i is the effect of the ith level, or component, of the first variable, β_j is the effect of the jth level of the second variable, and so on, and γ_l is the effect of l-th level of the kth variable. The terms $(\tau\beta)_{ij} + ... + (\tau\gamma)_{il}$ represent the effects of the $\frac{k(k-1)}{2}$ two factor interactions. The model error $e_{ij...l,t}$ may be autocorrelated and may also be non-stationary, both of which are in violation of the ANOVA assumption that the errors are independently distributed.

The subscript $ij...l$ indicates the context of the process model and the subscript t denotes the time index of the observations collected in this context. Denoting the number of observations in this context by $n_{ij...l}$ we have that $t = 1, 2, ..., n_{ij...l}$. Further, if we denote the number of levels of each variable by $r_i (i = 1, 2, ..., k)$ we have that $i = 1, 2..., r_1$, $j = 1, 2..., r_2$, and so on up to $l = 1, 2..., r_k$.

The full model contains $d = k + \frac{k(k-1)}{2}$ variables. While for illustration purposes of this report we consider the model (2) that contain up to two factor interactions, the method also works with models that include higher order interactions. Our objective is to find a subset of the d variables which best explains the variability in the disturbance data. In particular, we want to obtain a subset of variables that contribute most to the autocorrelation and the non-stationarity of the data and another subset of variables that contribute most to the mean shifts, or the offsets, in the data. Figure 1 illustrates a typical case where data from different contexts can be represented by drifting or non-stationary time series and offset terms.

To model the autocorrelations we employ integrated autoregressive (AR) processes which we represent as regression models. Provided that the full regression model is general enough, a variable selection approach would allow one to determine, according to the data, the appropriate orders of the AR models that should be used for different variables. To model the offsets we employ ANOVA models.

The remainder of the paper is organized as follows. We introduce the proposed model selection approach in Section 2 using a simple example. The variable selection criterion and the stopping rule used in the approach are reviewed in Section 3. In Section 4, the implementation of the approach for the general case is discussed. Sections 5 and 6 contain the main results of the paper; in Section 5 the approach is applied on a simulated process data, and in Section 6 on lithography alignment control.

2 Proposed Approach

The model selection algorithm starts from an initial ANOVA model that is provided by the user. We usually use $y_t = \mu + e_t$ as the initial ANOVA model. To the residuals of this model it successively fits a set of regression models, each model being fitted to the residuals of the previous model, until a stopping rule is violated. The variables in each model are selected according to a variable selection criterion. We explain the variable selection criterion and the stopping rule used in the algorithm in Section 3.

The basic idea in sequentially fitting regression models is illustrated in Figure 2 using an example where a response y is regressed on the variables x_1 and x_2. This figure shows that a first order model (i.e. no $x_1 x_2$ interaction) can be decomposed in to two regressions, first that regresses y on x_1 where the prediction is y_{p1} and the residual is e_1; and second, that regresses the residuals e_1 of the first regression on x_2 where the prediction is y_{p2} and the residual is e_2. It can be seen that $y_p = y_{p1} + y_{p2}$ and $e = e_1 + e_2$. It will be explained below that in the autoregressive models that we employ in the model selection algorithm we use only linear functions of the regressors and this condition is satisfied.

After the regression step, an ANOVA model is fitted to the residuals of the regression model, where the variables are selected according to the variable selection criterion. The variables are entered until the stopping rule is violated at which point an iteration is completed. At the end of each iteration a convergence criterion is checked; if it is satisfied the algorithm stops and reports the model as the final model; otherwise the algorithm continues with a new iteration using the residuals of the ANOVA model of the previous iteration.

As the convergence criterion, we compare the variables selected (for the regression and the ANOVA models) in the current iteration to those in the previous iteration. If they are the same we conclude that the algorithm has reached convergence.

To illustrate the proposed approach we consider a simple process where the tool type and the product type are the two context variables and the interactions are not important. A two-way

ANOVA model to represent the disturbance is:

$$y_{ij,t} = \mu + \tau_i + \beta_j + e_{ij,t} \tag{3}$$

where τ_i is the effect of the ith tool and β_j is the effect of the jth product. Suppose that there are 2 tools and 2 products and during production, the levels of the context variables are varied according to a 2^2 full factorial design and in each combination 3 observations are collected. The disturbance and model error values and the context variable levels for this example are given in Table 1.

The proposed regression model approach to model the autocorrelations in the error terms and the ANOVA model approach to model the offsets proceed as follows.

2.1 Regression model

Suppose that the time effects of the tools and the products are given, respectively, by the sets of time series u_{i,t_1} and v_{j,t_2}. Here, t_1 is the time index of the observations on the ith tool (i.e. $t_1 = 1, ..., n_i$) and t_2 is the time index of the observations on the jth product (i.e. $t_2 = 1, ..., n_j$). The model error is, thus, the summation of the two series:

$$e_{ij,t} = u_{i,t_1} + v_{j,t_2}. \tag{4}$$

It is noted that, since the offsets are modeled by the parameters μ, τ_i and β_j in (3), all of the time series u_{i,t_1} and v_{j,t_2} have zero means. Suppose, without loss of generality, that each of the time series can be represented by an AR(2) process, that is:

$$u_{i,t_1} = \phi_1^{(i)} u_{i,t_1-1} + \phi_2^{(i)} u_{i,t_1-2} + \epsilon_{i,t_1} \tag{5}$$

and

$$v_{j,t_2} = \varphi_1^{(j)} v_{j,t_2-1} + \varphi_2^{(j)} v_{j,t_2-2} + \varepsilon_{j,t_2} \tag{6}$$

where $\phi_1^{(i)}, \phi_2^{(i)}, \varphi_1^{(j)}$ and $\varphi_2^{(j)}$ are the autoregressive parameters and ϵ_{i,t_1} and ε_{j,t_2} are white noise processes.

It can be shown that, if we fit a regression model to $e_{ij,t}$ where the regressors are the lagged values of $e_{ij,t}$ defined over the tools, the regression will be significant and the residuals of this model will still have autocorrelation left due to the products. To show this, we rewrite (4) by renumbering the errors according to the tools as $e_{ij,t} \equiv e_{i,t_1}$ and by substituting (5) for u_{i,t_1}:

$$e_{i,t_1} = \phi_1^{(i)} u_{i,t_1-1} + \phi_2^{(i)} u_{i,t_1-2} + \epsilon_{i,t_1} + v_{j,t_2}. \tag{7}$$

In this expression, we can substitute e_{i,t_1-1} and e_{i,t_1-2} for u_{i,t_1-1} and u_{i,t_1-2} by using (4):

$$e_{i,t_1} = \phi_1^{(i)} e_{i,t_1-1} + \phi_2^{(i)} e_{i,t_1-2} + \epsilon_{i,t_1} + v_{j,t_2} + c \tag{8}$$

where the term c contains the past values of v_{j,t_2} which comes from the use of equation (4). It is clear that, the residuals of the estimated regression model (8) are autocorrelated over the products because v_{j,t_2} and c are autocorrelated over the products. Therefore, the procedure can be repeated using the residuals of (8) and defining the lagged values of these residuals over the products as the regressors.

It should also be noted that in (8) the model error of the regression is $\epsilon_{i,t_1} + v_{j,t_2} + c$ and it is assumed to be independently distributed for different t_1 on tool i. In order to satisfy this assumption, a randomly selected product number j must be used with tool i, because otherwise, according to (6), the errors will be correlated. This assumption is satisfied in production, because during production the levels of the context variables are usually randomized.

These two regression models are explained next. In the first step, we fit the regression model where the regressors are the lagged values of e_{ij} defined over the tools. The observations can be written according to this model as:

$$\boldsymbol{e} = \boldsymbol{Z}_1 \boldsymbol{g}_1 + \boldsymbol{a} \tag{9}$$

where $\boldsymbol{Z}_1$ is the regressor matrix, $\boldsymbol{g}_1$ is the vector of parameters and $\boldsymbol{a}$ is the vector of errors and they are defined as

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \end{bmatrix}, \boldsymbol{Z}_1 = \begin{bmatrix} \gamma_0 & \gamma_{-1} & 0 & 0 \\ e_1 & \gamma_0 & 0 & 0 \\ e_2 & e_1 & 0 & 0 \\ 0 & 0 & \eta_0 & \eta_{-1} \\ 0 & 0 & e_4 & \eta_0 \\ 0 & 0 & e_5 & e_4 \\ e_3 & e_2 & 0 & 0 \\ e_7 & e_3 & 0 & 0 \\ e_8 & e_7 & 0 & 0 \\ 0 & 0 & e_6 & e_5 \\ 0 & 0 & e_{10} & e_6 \\ 0 & 0 & e_{11} & e_{10} \end{bmatrix}, \boldsymbol{g}_1 = \begin{bmatrix} \phi_1^{(1)} \\ \phi_2^{(1)} \\ \phi_1^{(2)} \\ \phi_2^{(2)} \end{bmatrix}, \text{ and } \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{12} \end{bmatrix}. \tag{10}$$

γ_0 and γ_{-1} denote the two pre-sample values of $\boldsymbol{e}$ on tool 1 and η_0 and η_{-1} denote those on tool 2. The pre-sample values are unknown, and a common estimation approach, known as the conditional maximum likelihood estimation in time series literature (see [1], pg. 226), is to substitute suitable values for the pre-sample values and compute the estimates conditional on these values. When the sample size is sufficiently large, this assumption on the starting values of the series is expected to have negligible impact on the final results. In this study we assume that the pre-sample values are equal to the starting value of each series, that is, we set $\gamma_0 = \gamma_{-1} = e_1$ and $\eta_0 = \eta_{-1} = e_4$. Provided that the time series is long enough, this assumption on the starting values of the series would have negligible impact on the final results.

The ordinary least squares (OLS) estimates of the parameters are given by $\hat{\boldsymbol{g}}_1 = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{e}$ and the residuals of the fitted model are

$$\hat{\boldsymbol{a}} = \boldsymbol{e} - \boldsymbol{Z}_1\hat{\boldsymbol{g}}_1. \tag{11}$$

In the second step, we fit the regression model to the residuals $\hat{\boldsymbol{a}}$ where the regressors are the lagged values of $\hat{\boldsymbol{a}}$ defined over the products. The observations according to this model are written as:

$$\hat{\boldsymbol{a}} = \boldsymbol{Z}_2\boldsymbol{g}_2 + \boldsymbol{s} \tag{12}$$

where $\boldsymbol{Z}_2$ is the regressor matrix, $\boldsymbol{g}_2$ is the vector of parameters and $\boldsymbol{s}$ is the vector of errors and they are defined as

$$\hat{\boldsymbol{a}} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \hat{a}_8 \\ \hat{a}_9 \\ \hat{a}_{10} \\ \hat{a}_{11} \\ \hat{a}_{12} \end{bmatrix}, \boldsymbol{Z}_2 = \begin{bmatrix} \gamma'_0 & \gamma'_{-1} & 0 & 0 \\ \hat{a}_1 & \gamma'_0 & 0 & 0 \\ \hat{a}_2 & \hat{a}_1 & 0 & 0 \\ \hat{a}_3 & \hat{a}_2 & 0 & 0 \\ \hat{a}_4 & \hat{a}_3 & 0 & 0 \\ \hat{a}_5 & \hat{a}_4 & 0 & 0 \\ 0 & 0 & \eta'_0 & \eta'_{-1} \\ 0 & 0 & \hat{a}_7 & \eta'_0 \\ 0 & 0 & \hat{a}_8 & \hat{a}_7 \\ 0 & 0 & \hat{a}_9 & \hat{a}_8 \\ 0 & 0 & \hat{a}_{10} & \hat{a}_9 \\ 0 & 0 & \hat{a}_{11} & \hat{a}_{10} \end{bmatrix}, \boldsymbol{g}_2 = \begin{bmatrix} \varphi_1^{(1)} \\ \varphi_2^{(1)} \\ \varphi_1^{(2)} \\ \varphi_2^{(2)} \end{bmatrix}, \text{ and } \boldsymbol{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{12} \end{bmatrix}. \tag{13}$$

γ'_0 and γ'_{-1} denote the pre-sample values of $\hat{\boldsymbol{a}}$ on product 1 and η'_0 and η'_{-1} denote those on product 2. Similarly to (10) we set $\gamma'_0 = \gamma'_{-1} = \hat{a}_1$ and $\eta'_0 = \eta'_{-1} = \hat{a}_7$.

The OLS estimates are $\hat{\boldsymbol{g}}_2 = (\boldsymbol{Z}'_2\boldsymbol{Z}_2)^{-1}\boldsymbol{Z}'_2\hat{\boldsymbol{a}}$ and the residuals of the fitted model are

$$\hat{\boldsymbol{s}} = \hat{\boldsymbol{a}} - \boldsymbol{Z}_2\hat{\boldsymbol{g}}_2. \tag{14}$$

Since the autocorrelations due to the tools and the products are now modeled, the components of the residual vector $\hat{\boldsymbol{s}}$ are uncorrelated. From (11) and (14) it can be seen that

$$\boldsymbol{e} = \hat{\boldsymbol{e}} + \hat{\boldsymbol{s}} \tag{15}$$

where $\hat{\boldsymbol{e}} = \boldsymbol{Z}_1\hat{\boldsymbol{g}}_1 + \boldsymbol{Z}_2\hat{\boldsymbol{g}}_2$ is the vector of predictions of the two regression models.

2.2 ANOVA model

After modeling the autocorrelations in the error term by the fitted regression model, we model the offsets in the data by an ANOVA model. We can write the model (3) by using (15) as

$$y_{ij,t} = \mu + \tau_i + \beta_j + \hat{e}_{ij,t} + \hat{s}_{ij,t}. \tag{16}$$

Defining $\tilde{y}_{ij,t} = y_{ij,t} - \hat{e}_{ij,t}$ we can see that the model

$$\tilde{y}_{ij,t} = \mu + \tau_i + \beta_j + \hat{s}_{ij,t} \tag{17}$$

satisfies the ANOVA model assumptions because $\hat{s}_{ij,t}$ are independently distributed. The observations $\tilde{y}_{ij,t}$ can be written according to this model as:

$$\tilde{\boldsymbol{y}} = \boldsymbol{X} \boldsymbol{b} + \hat{\boldsymbol{s}}$$

or

$$\begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \hat{s}_1 \\ \vdots \\ \hat{s}_{12} \end{bmatrix} \tag{18}$$

where $\mathbf{1}$ is a 3×1 vector of ones and $\mathbf{0}$ is a 3×1 vector of zeros. The columns of the regressor matrix $\boldsymbol{X}$ are linearly dependent, thus $\boldsymbol{X}'\boldsymbol{X}$ is less than full rank and hence is not invertible. This also implies that a unique solution for $\boldsymbol{b}$ does not exist, however, a solution that satisfies the normal equations can be obtained by using a generalized inverse. We use the Moore-Penrose generalized inverse [7] which gives the unique minimum norm solution. Let $\boldsymbol{X}^-$ denote the Moore-Penrose generalized inverse of $\boldsymbol{X}$. Thus, the unique minimum norm solution of (18) can directly be obtained as $\hat{\boldsymbol{b}} = \boldsymbol{X}^-\tilde{\boldsymbol{y}}$ (If $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, the Moore-Penrose inverse satisfies $\boldsymbol{X}^- = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$).

It can be shown that (see e.g. [9], pg. 170) the predictions and consequently the sum of squared errors of the fitted model are invariant to the choice of the generalized inverse. The variable selection, as will be explained next, is made on the basis of the sum of squared errors, and thus, the selected models are not affected by this choice as well.

3 Variable Selection Criterion and the Stopping Rule

Variable selection in regression, that is obtaining a subset of a larger set of regressor variables, has attracted considerable attention in the statistics literature. The most commonly used methods are the best subsets analysis, which enumerates all subsets (in the case of the model (2) there are 2^d subsets) to select the subset according to an optimality criterion, and the class of methods that includes the forward selection, backward elimination and stepwise regression, which utilize a search method to enumerate some of the subsets. Miller ([6]) gives a review of some of the widely used variable selection algorithms.

In this study we adopt a forward-selection approach and as the variable selection criterion we use the C_p statistic suggested by Mallows [5]. The C_p statistic of a model with p parameters is defined as:

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2} + 2p - n \tag{19}$$

where SSE is the sum of squared errors of the fitted model and $\hat{\sigma}^2$ is the estimated error variance. In variable selection problems $\hat{\sigma}^2$ is usually computed as the mean square error of the full model. The variable to be entered to the model is selected as the one that minimizes C_p. According to (19) a variable that minimizes C_p provides the smallest SSE with fewest possible parameters.

It is noted that, in an ANOVA model, when a variable is entered to the model, the parameters associated with all the different levels of this variable are entered. For example, if τ_i is being entered, then all the parameters $\{\tau_1, ..., \tau_{r_1}\}$ are entered. Thus, in this case $p = r_1$.

If there are no variables already in the model, the selected variable is entered to the model if the fitted model is significant. If the model already contains p parameters, then a new variable is selected as the one that minimizes C_{p+q}, where q is the number of parameters in the new variable. The selected variable is entered to the model if it satisfies

$$C_p \geq C_{p+q} \tag{20}$$

which is also the stopping rule of our algorithm, because we stop entering more variables once it is violated. It can be shown that this condition is equivalent to

$$SSE(p) - SSE(p+q) \geq 2q\hat{\sigma}^2. \tag{21}$$

4 Algorithm

In this section we explain how the model selection algorithm is applied for the general case of k context variables. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of disturbance observations and $\boldsymbol{X}$ denote the regressor matrix for all dummy variables including the overall mean μ, the main effects and the two factor interactions, that is $\boldsymbol{X}$ is $n \times d'$ where $d' = 1 + d$.

Let $\boldsymbol{X}_0$ denote the columns of $\boldsymbol{X}$ that correspond to the variables in the user provided initial ANOVA model. Therefore, the residuals of the initial model are $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}_0\hat{\boldsymbol{b}}_0$ where $\hat{\boldsymbol{b}}_0 = \boldsymbol{X}_0^- \boldsymbol{y}$. The remaining of the algorithm consists of two parts.

4.1 Regression Model Selection

The regression model selection algorithm can be given as follows.

0. Set variable counter $count = 0$; set the flag that indicates that the main effects have been checked to $flag_{me} = 0$; Let $\boldsymbol{e}_0 = \boldsymbol{e}$ denote the original residual vector

1. If $flag_{me} = 0$ then go to (1.a) otherwise go to (1.b)

   1.a Main effects: Create the regressor matrices $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$ from the lagged values of $\boldsymbol{e}$ for the main effects $i = 1, 2, ..., k$, where $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$ correspond, respectively to the AR(1) and AR(2) models on the variable i.

   1.b Two-factor interactions: Create the regressor matrices $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$ from the lagged values of $\boldsymbol{e}$ for the two factor interaction effects $i = 1, 2, ..., \frac{k(k-1)}{2}$, where $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$ correspond, respectively to the AR(1) and AR(2) models on the interaction i

2. Calculate the C_p value of the AR(1) and AR(2) models of each effect i that is not already in the model. Let C_p^* denote the minimum C_p value, let i^* denote the corresponding effect number, and let $\boldsymbol{Z}^*$ denote the corresponding regressor matrix

3. Estimate the selected regression model. If $count = 0$ then test the significance of the estimated model, otherwise check the stopping rule $C_{p,model} \geq C_p^*$.

4. If the estimated model is significant or if the stopping rule is not violated then enter the variable

   - Set $count := count + 1$
   - Test for unit autoregressive roots
   - Set $C_{p,model} = C_p^*$

- Compute the residuals of this model $\hat{\boldsymbol{a}} = \boldsymbol{e} - \boldsymbol{Z}^* \hat{\boldsymbol{g}}$ where $\hat{\boldsymbol{g}} = (\boldsymbol{Z}^{*'} \boldsymbol{Z}^*)^{-1} \boldsymbol{Z}^{*'} \boldsymbol{e}$
- Set $\boldsymbol{e} := \hat{\boldsymbol{a}}$
- If $flag_{me} = 0$ and if all main effects are checked then set $flag_{me} = 1$ and go to (1). If $flag_{me} = 1$ and if all two factor interactions are checked then go to (6)

5. If $flag_{me} = 0$ then set $flag_{me} = 1$ and go to (1)
6. Compute the residuals $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{e}}$ of the regression model where $\hat{\boldsymbol{e}} = \boldsymbol{e}_0 - \boldsymbol{e}$ is the vector of predictions of the residuals. Go to ANOVA Model Selection.

Remark: (Sample size). In estimation of ARMA models, it is recommended to have a sample size of at least 100 observations (see e.g. [2]). In this study we are considering a large number of variable combinations and this may result in a considerably large sample size requirement. Furthermore, we are estimating low order linear AR models and, thus, a relatively small sample size would provide satisfactory estimates. According to this, we set the minimum sample size for AR models to 20 observations. When fitting an AR model to a variable, the number of observations in each level of this variable is checked. The levels that do not satisfy this condition are not considered in the estimation.

4.1.1 Tests for significance

In the test for significance of the regression models, we test, for an AR(1) model, the null hypothesis H_0 : all $\phi_1^{(j)} = 0$ against the alternative hypothesis H_1 : at least one $\phi_1^{(j)} \neq 0$ and for an AR(2) model H_0 : all $\phi_1^{(j)} = 0$ and $\phi_2^{(j)} = 0$ against H_1 : at least one $\phi_1^{(j)} \neq 0$ or $\phi_2^{(j)} \neq 0$ where $j = 1, 2, ..., m$ and m is the number of columns in $\boldsymbol{Z}^*$. We reject the null hypothesis if $F_0 \geq F_{\alpha,m,n-m}$ where $F_0 = \frac{SSR/m}{SSE/(n-m)}$, $SSR = \hat{\boldsymbol{g}}' \boldsymbol{Z}^{*'} \boldsymbol{e}$, $SSE = \boldsymbol{e}'\boldsymbol{e} - \hat{\boldsymbol{g}}' \boldsymbol{Z}^{*'} \boldsymbol{e}$ and $F_{\alpha,m,n-m}$ is the upper 100α percentile point of an F distribution with m and $n - m$ degrees of freedom.

4.1.2 Test for unit autoregressive roots

For all selected regression models we test for the significance of unit autoregressive roots. An AR process with unit root is non-stationary and it drifts if the constant term is non-zero (see

[1]). In order to detect the variables that drift, the non-stationary context variables are given priority in the ANOVA model by starting from these variables in the tests for significance.

For an AR(1) model where the parameter estimates are $\hat{\boldsymbol{g}} = (\hat{\phi}_1^{(1)}, ..., \hat{\phi}_1^{(m)})'$ we test the null hypothesis $H_0^{(j)} : \phi_1^{(j)} = 1$ for each $j = 1, 2, ..., m$. We use the test statistic $F_0 = \frac{(\hat{\phi}_1^{(j)} - 1)^2}{var(\hat{\phi}_1^{(j)})}$ and compare it to $F_{\alpha,1,n-m}$. Here, $var(\hat{\phi}_1^{(j)})$ is the jth diagonal of the parameter covariance matrix

$$var(\hat{\boldsymbol{g}}) = s^2(\boldsymbol{Z}^{*'}\boldsymbol{Z}^*)^{-1} \tag{22}$$

where s^2 is the model mean square error

$$s^2 = \frac{1}{n-m}(\boldsymbol{e}'\boldsymbol{e} - \boldsymbol{e}'\boldsymbol{Z}^*\hat{\boldsymbol{g}}). \tag{23}$$

If we fail to reject the null hypothesis for at least one j then we conclude that there is evidence of at least one unit root in the sets of time series that correspond to this effect.

For an AR(2) model where the parameter estimates are $\hat{\boldsymbol{g}} = (\hat{\phi}_1^{(1)}, ..., \hat{\phi}_1^{(m)}, \hat{\phi}_2^{(1)}, ..., \hat{\phi}_2^{(m)})'$ we test the null hypothesis $H_{0,1}^{(j)} : \phi_2^{(j)} - \phi_1^{(j)} = 1$ and $H_{0,2}^{(j)} : \phi_2^{(j)} + \phi_1^{(j)} = 1$ for each $j = 1, 2, ..., m$. We use the test statistics

$$F_{0,i} = \frac{(\boldsymbol{C}_i\hat{\boldsymbol{g}} - 1)^2}{s^2\boldsymbol{C}_i(\boldsymbol{Z}^{*'}\boldsymbol{Z}^*)^{-1}\boldsymbol{C}_i'}$$

for the hypotheses $i = 1$ and 2. Here, s^2 is obtained using (23) by replacing the denominator with $n - 2m$. $\boldsymbol{C}_1$ is a $1 \times 2m$ vector with jth entry equal to -1 and $(j+m)$th entry equal to 1 and all other entries equal to 0. $\boldsymbol{C}_2$ is a $1 \times 2m$ vector with jth and $(j+m)$th entries equal to 1 and all other entries equal to 0. We compare the test statistics to $F_{1,n-2m,\alpha}$. If we fail to reject the null hypothesis $H_{0,1}^{(j)}$ or $H_{0,2}^{(j)}$ for at least one j then we conclude that there is evidence of at least one unit root.

4.1.3 Exponential weighting of the past data

One major assumption made in formulating the AR models as in (9) and (12) is that the sampling interval is uniform for all observations. As it is clear from (10) and (13) the sampling interval is changing at the time instants where the components of the context variables are switching; for example for tool 1, the observation vector is $(e_1, e_2, e_3, e_7, e_8, e_9)$, and the sampling interval changes from 1 to 4 time steps at the fourth observation.

Non-uniform sampling intervals would cause inaccuracy in modeling the autocorrelations using a regression model as this model assumes that the past data are separated by an equal time interval. We propose to use an exponential weighting approach to discount the effect of the past data by how much it is separated from the current observation when using the past data as a regressor. Consider, for example e_8. When regressing this point on e_7 and e_3 we weigh e_7 by λ and e_3 by λ^{8-3}, where λ is the exponential weight. We recommend a relatively high value for λ, such as 0.8.

4.2 ANOVA Model Selection

The ANOVA model selection algorithm can be given as follows.

0. Set $count = 0$. Test the significance of the overall mean $\hat{\mu}$; if it is significant then enter it into the model, that is $\boldsymbol{X}_r = \mathbf{1}$ where $\mathbf{1}$ is $n \times 1$ vector of ones otherwise $\boldsymbol{X}_r = \emptyset$

1. Test the significance of the variables that have unit autoregressive roots in the regression model. If they are significant then enter the variables and

    - set $count := count + 1$
    - Update the regressor matrix: Let $\boldsymbol{X}_s$ denote the columns of $\boldsymbol{X}$ that correspond to the variables with unit roots and set $\boldsymbol{X}_r := [\boldsymbol{X}_r \;\; \boldsymbol{X}_s]$

2. Calculate the C_p value for each of the main effects and two factor interactions $i = 1, 2, ..., k + \frac{k(k-1)}{2}$ that is not already in the model. Let C_p^* denote the minimum C_p value and let i^* denote the corresponding variable number

3. Estimate the selected ANOVA model. If $count = 0$ then test the significance of the estimated model, otherwise check the stopping rule $C_{p,model} \geq C_p^*$.

4. If the estimated model is significant or if the stopping rule is not violated then enter the variable and

    - Set $count := count + 1$

- Set $C_{p,model} = C_p^*$
- Update the regressor matrix: let $\boldsymbol{X}_s$ denote the columns of $\boldsymbol{X}$ that correspond to variables i^* and set $\boldsymbol{X}_r := [\boldsymbol{X}_r \ \ \boldsymbol{X}_s]$
- Go to (2)

otherwise check the convergence criterion. If it is not satisfied calculate the vector of residuals $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}_r\hat{\boldsymbol{b}}_r$ where $\hat{\boldsymbol{b}}_r = \boldsymbol{X}_r^- \boldsymbol{y}$ and go to Regression Model Selection. If it is satisfied, stop and report the variables selected in the regression model (along with the order of the AR filter) and the variables selected in the ANOVA model as the final model.

Remark: (Checking for nested factors). In both the ANOVA and the regression models, in order to define an interaction effect, we check the variables for nestedness. In linear models, if the levels of one variable are nested under the levels of another factor, then the interaction between these two variables is not defined (see e.g. [9], pg. 155-159).

4.2.1 Tests for significance

In the test for significance of the ANOVA models, we test, for example for the effects $\tau_i, i = 1, 2, ..., r_1$, the null hypothesis H_0 : all $\tau_i = 0$ against the alternative hypothesis H_1 : at least one $\tau_i \neq 0$. We reject the null hypothesis if $F_0 \geq F_{\alpha,m,n-m}$ where $F_0 = \frac{SSR/m}{SSE/(n-m)}$, $SSR = \hat{\boldsymbol{b}}_r' \boldsymbol{X}_r' \tilde{\boldsymbol{y}}$, $SSE = \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \hat{\boldsymbol{b}}_r' \boldsymbol{X}_r' \tilde{\boldsymbol{y}}$, $\hat{\boldsymbol{b}}_r = \boldsymbol{X}_r^- \tilde{\boldsymbol{y}}$. $\boldsymbol{X}_r$ denote the columns of $\boldsymbol{X}$ that correspond to τ_i (and μ, if it is significant) and $F_{\alpha,m,n-m}$ is the upper 100α percentile point of an F distribution with m and $n - m$ degrees of freedom where $m = rank(\boldsymbol{X}_r)$.

5 Simulation Example

In this example, we consider tool, product and operation as the context variables, and assume that the effects of the tools and products are significant, but the effects of the operations are not significant.

We consider two different scenarios. In the first one, the disturbance due to the tools and the products follow independent ARIMA(1,1,0) processes. If we denote the time effect due to the tool i by u'_{i,t_1}, this can be expressed as (see e.g. [1])

$$u'_{i,t_1} = \frac{1}{(1-\phi\mathcal{B})(1-\mathcal{B})}\epsilon_{i,t_1} + \frac{\delta^{(i)}}{1-\mathcal{B}}$$

where $\mathcal{B}$ is the back shift operator (i.e. $\mathcal{B}u_{i,t_1} = u_{i,t_1-1}$) and $\delta^{(i)}$ is the drift rate of the ith tool. A constant AR parameter ϕ is assumed for all tools. We can rewrite this as

$$u'_{i,t_1} = (\phi+1)u'_{i,t_1-1} - \phi u'_{i,t_1-2} + \delta^{(i)}(1-\phi) + \epsilon_{i,t_1}.$$

Similarly, for the products a constant AR parameter φ is assumed. If we denote the time effect due to product j by v'_{j,t_2} and the drift rate by $\xi^{(j)}$ we have that

$$v'_{j,t_2} = (\varphi+1)v'_{j,t_2-1} - \varphi v'_{j,t_2-2} + \xi^{(j)}(1-\varphi) + \varepsilon_{j,t_2}.$$

It can be shown that the process output can be represented as shown in (3) and the model error as shown in (4), where the constant terms are $\mu = 0$, $\tau_i = \delta^{(i)}(1-\phi)$, $\beta_j = \xi^{(j)}(1-\varphi)$ and the individual error terms are $u_{i,t_1} = u'_{i,t_1} - \delta^{(i)}$ and $v_{j,t_2} = v'_{j,t_2} - \xi^{(j)}$.

The disturbance process is simulated with 5 tools, 5 products and 5 operations by assuming equal proportions for the levels of each context variable (i.e. the probability of occurrence of each level of each variable is 0.2). We used $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ as the tool drift rates $\delta^{(i)}, i = 1, ..., 5$, and $\{-0.1, -0.3, -0.5, -0.7, -0.9\}$ as the product drift rates $\xi^{(j)}, j = 1, ..., 5$, and $\phi = 0.4$ and $\varphi = -0.4$ as the AR parameters. Furthermore, the white noise processes are assumed to be normally distributed with mean 0 and variance 1.

In the second scenario, the disturbance follows a different ARIMA(1,1,0) process in each tool-product combination. If we denote the time effect due to tool i and product j by $e_{ij,t}$ and represent it as

$$e_{ij,t} = (\phi+1)e_{ij,t-1} - \phi e_{ij,t-2} + \delta^{(ij)}(1-\phi) + \epsilon_{ij,t}$$

then the disturbance model (3) can be written as $y_{ij,t} = (\tau\beta)_{ij} + e_{ij,t}$ where $(\tau\beta)_{ij} = \delta^{(ij)}(1-\phi)$.

The process is simulated by assuming the drift rates $\delta^{(ij)}$ given in Table 2, and a constant AR parameter $\phi = 0.4$. The same white noise properties and the context variable proportions as in the first scenario are assumed. For both scenarios, 1000 lots were simulated.

Tables 3 and 4 show the iterations of the algorithm for the first and second scenarios, respectively. The symbols T and P are used to represent the effects of the tools and the products, respectively. For both scenarios we used $y_t = \mu + e_t$ as the initial ANOVA model. As it can be seen, in both of the scenarios, the algorithm converged to the true model of the process. In the first scenario, we obtained the ANOVA model $y_{ij,t} = \tau_i + \beta_j + e_{ij,t}$ where $e_{ij,t}$ follow AR processes due to the tools and the products, and in the second scenario we obtained the ANOVA model $y_{ij,t} = \mu + (\tau\beta)_{ij} + e_{ij,t}$ where $e_{ij,t}$ follow AR processes due to products, tools and tool-product interactions. Since the interaction can model the effects due to the main effects, we can simplify the regression model of the second scenario as an AR process due to the tool-product interactions.

Figure 3 shows the comparison of the actual data to the residuals after fitting each model, and the comparison of the actual data to the predictions computed by the selected models.

6 Industrial Application

In this section we apply the proposed model selection algorithm to a well studied problem in semiconductor manufacturing, the alignment control for lithography process. [4] illustrates an application and explain the outputs and the inputs of the process and [10] presents a run to run control algorithm for alignment control.

In this study we consider 3 response variables in a lithography process: $yshift, xscale$ and *yscale*, which are the displacement in the y direction, the elongation in the x direction and the elongation in the y direction of the pattern in the current layer measured with respect to the pattern in the previous layer. The context variables of this process are *scanner*, *prior scanner*, *operation*, *product*, *recipe* and *reticle*. See [10] for definitions of these variables.

We use a data set that contains 10000 disturbance observations for these responses. Disturbance data is obtained from the input-output data of the process by assuming that the responses are not correlated and that the each process has unit gain. Table 5 shows the number of levels of each context variable and figure 4 shows the pairwise plots of the levels of the context variables under which the data is collected. It is also noted that, in this data the levels of the recipe and reticle are nested under the levels of operation.

From figure 4, we can see that the levels of *product* and *reticle*, *product* and *reticle*, and *recipe* and *reticle* are highly correlated (The sample correlation coefficients between these pairs are 0.991, 0.679 and 0.699, respectively); that is, the levels of the regressors *product*, *recipe* and *reticle* "move together". Highly correlated variables in general cause the variable selection algorithm to perform poorly because it is not possible to distinguish the effects of each variable. To remedy this problem, we removed *recipe* and *product* from the list of variables and considered *scanner*, *prior scanner*, *operation* and *reticle* (and their interactions) as potential variables. Table 5 shows the symbols used for these variables.

We run the model selection algorithm on the disturbance data of each of the response variables. To validate the models selected, we split the data into two, and use the first 5000 observations for model selection and estimation and the second 5000 rows to compare the predictions made by the estimated model (here the row order is by time).

As shown in Table 6 using the first 5000 rows of the $yshift$ response data the algorithm converged in 3 iterations to the ANOVA model $y_t = S \times R + P \times O + e_t$ where e_t follow AR(2) processes due to the scanners. The initial ANOVA model in this case was $y_t = \mu + e_t$. The ϕ_1 and ϕ_2 estimates, their standard errors and the results of the unit autoregressive root tests are shown in Table 7. The summary of the regression and the ANOVA models fitted to the $yshift$ response data are given in Table 8; as it can be seen both models are highly significant. The same ANOVA and regression models were obtained using the second 5000 rows of the data.

For the $xscale$ response we obtained the ANOVA model $y_t = S \times R + P \times O + e_t$ where e_t follow AR(1) processes due to the scanners and for the $yscale$ response we obtained the ANOVA model $y_t = S \times R + P \times R + e_t$ where e_t follow AR(2) processes due to the scanners. The same model was obtained for each response using the first and second 5000 rows of the data.

Figure 5 shows the residuals $\tilde{y}_t$ of the selected regression model for $yshift$. The residuals are plotted for the different levels of *scanner*, *prior scanner* and the different levels of *reticle* on the first level of *scanner*. These plots show that an ANOVA model that contains only *scanner* or only *reticles* would not adequately represent the mean shifts in the data, but, as also selected by the algorithm, it must at least contain a $scanner - reticle$ interaction.

Figure 6 shows the disturbance values for $yshift$ on different $scanner - reticle$ combinations after removing the offsets due to *prior* $scanner - reticle$. It illustrates that the disturbance data wanders in similar patterns on the same *scanner* and the separation of the time series reflects the offsets due the *reticle*s. The autocorrelations due to the scanners and those of the operations and reticles were also compared by applying Fisher's white noise test [3]. This procedure tests the null hypothesis that the series is a normal white noise against the alternative that it has a periodic component, where the null is rejected if the test statistic κ is large. The κ values and the corresponding p-values are reported in Table 9. This indicates that the autocorrelations due to the *scanner*s are more significant than those due to *operation*s and *reticle*s.

Figure 7 gives the predictions and the actual data of the second 5000 rows of $yshift$ for different $scanner - reticle$ combinations. Here, the predictions are the one-step-ahead forecasts and are computed using the parameter estimates obtained from the first half of the data set. As it can be seen, there is a good fit between the predictions and the actual data.

7 Conclusions

This paper presented an algorithmic approach for context based model selection for run to run control. In addition the algorithm explicitly differentiates between the contexts that experience autocorrelation and those that can be represented by offsets. This has direct bearings on the controller design. One assumption used in this work is that the context variables are sufficiently randomized during production so that the white noise assumption is satisfied at each step in the algorithm. Extensions to this procedure would involve inclusion of additional operational events (such as preventive maintenance) to context variables to further improve accuracy.

References

[1] Box, G. E. P., Jenkins, G. W., Reinsel, G. C. (1994), *Time Series Analysis, Forecasting and Control*, Prentice Hall, Inc.

[2] Castillo, E. D. (2002), *Statistical Process Adjustment for Quality Control*, Wiley & Sons, New York.

[3] Fisher, R. A. (1929), "Tests of Significance in Harmonic Analysis", *Proceedings of the Royal Society of London. Series A*, 125, No. 796, 54-59.

[4] Janakiram, M. and Goernitz, S. (2005), "Real-Time Lithography Registration, Exposure and Focus Control - A Framework for Success", *IEEE Transactions on Semiconductor Manufacturing*, 18, No. 4, 534-538.

[5] Mallows, C. L. (1973), "Some Comments on C_p", *Technometrics*, 15, No. 4, 661-675.

[6] Miller, A. J. (1984), "Selection of Subsets of Regression Variables", *Journal of the Royal Statistical Society. Series A*, 147, No. 3, 389-425.

[7] Penrose, R. A. (1955), "A Generalized Inverse for Matrices", *Proceedings of the Cambridge Philosophical Society*, 51, 406-13.

[8] Sachs, E., Hu, A., and Ingolfsson, A. (1995), "Run by run process control: Combining SPC and feedback control", *IEEE Transactions on Semiconductor Manufacturing*, 8, No. 1, 26-43.

[9] Searle, S. R. (1971), *Linear Models*, Wiley & Sons, New York.

[10] Toprac, A., and Wang Y. (2004), "Solving the High-mix Control Problem", *Proceedings of the SEMATECH AEC/APC Symposium.*

8 Figures and Tables

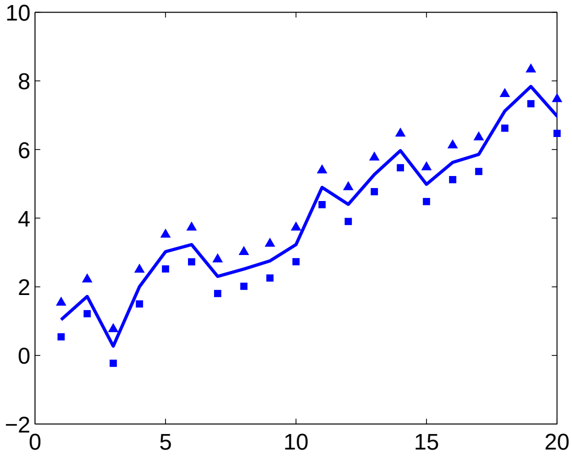

Figure 1: This figure illustrates how data from different contexts can be represented with drift and offset terms. Square marker: context 1, Triangle marker: context 2, Solid line: context 3

Disturbance	Model error	Tool	Product
y_1	e_1	1	1
y_2	e_2	1	1
y_3	e_3	1	1
y_4	e_4	2	1
y_5	e_5	2	1
y_6	e_6	2	1
y_7	e_7	1	2
y_8	e_8	1	2
y_9	e_9	1	2
y_{10}	e_{10}	2	2
y_{11}	e_{11}	2	2
y_{12}	e_{12}	2	2

Table 1: Disturbance data, model error and context variable levels for the 2 tool 2 product example

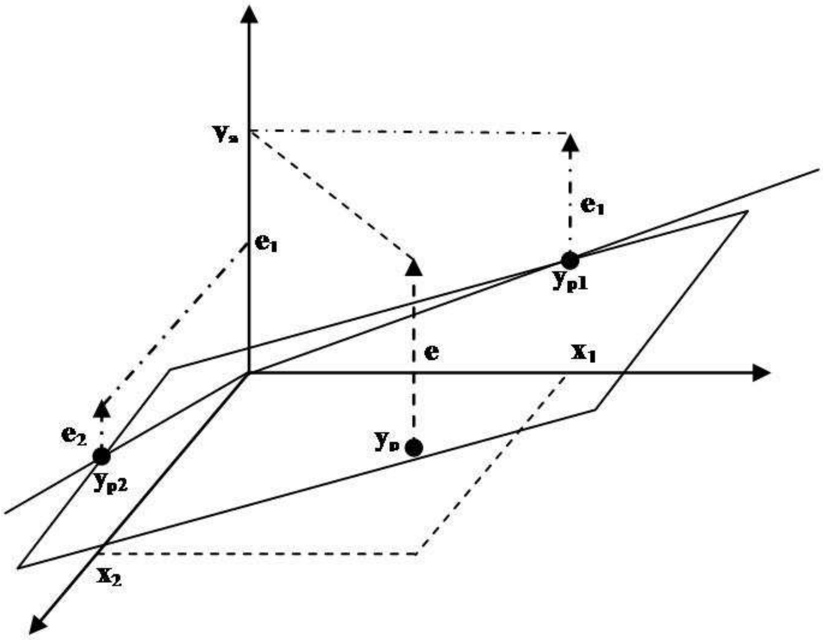

Figure 2: Sequential regression: This figure illustrates how a regression of y on x_1 and x_2 can be decomposed in to two regressions, first that regresses y on x_1 and second, that regresses the residuals e_1 of the first regression on x_2

	1	2	3	4	5
1	-0.005	-0.003	-0.001	0.001	0.003
2	-0.003	-0.001	0.001	0.003	0.005
3	-0.001	0.001	0.003	0.005	0.007
4	0.001	0.003	0.005	0.007	0.009
5	0.003	0.005	0.007	0.009	0.011

Table 2: Drift rates, $\delta^{(ij)}$. Rows are for tools($i = 1, ..., 5$), columns are for products ($j = 1, ..., 5$)

Iteration	Regression model	ANOVA model	Convergence ?
1	AR(2) on T, AR(2) on P	$T + P$	No
2	AR(2) on T, AR(2) on P	$T + P$	Yes

Table 3: Simulation example: iterations in the first scenario. The algorithm converged to the true solution

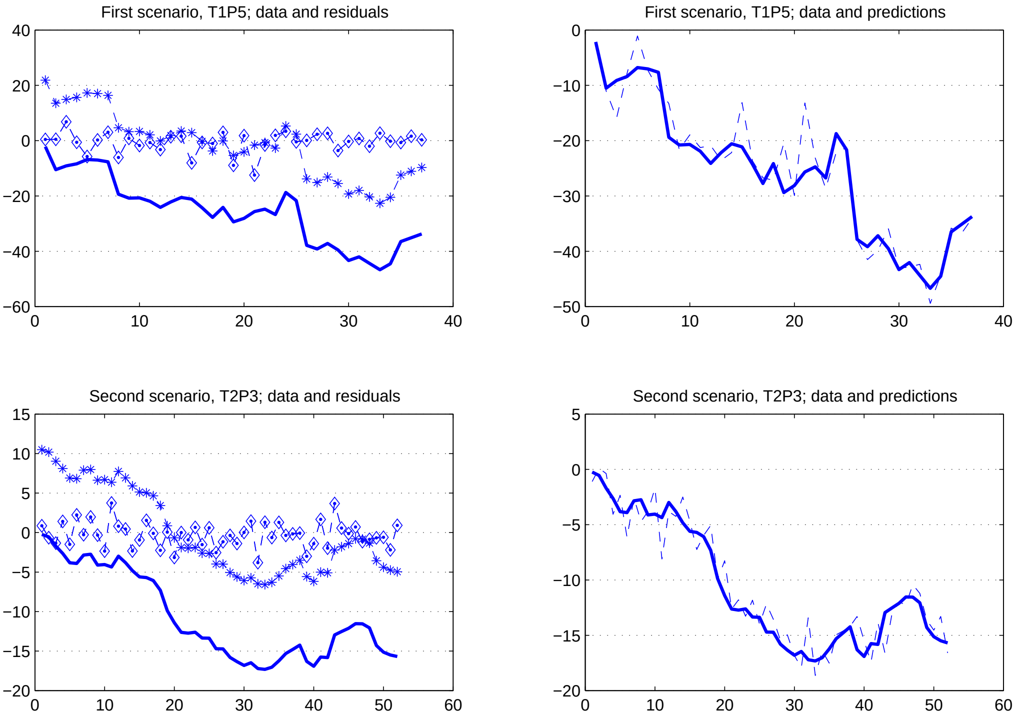

Figure 3: Simulation example. The actual simulated data (solid line), the residuals of the ANOVA model (dash line with asterisk), the residuals of the ANOVA and the regression models (dash line with diamond) and the predictions (dash line) are shown for the two scenarios. Top: First scenario, Tool 1 and Product 5, Bottom: Second scenario, Tool 2 and Product 3.

Iteration	Regression model	ANOVA model	Convergence ?
1	AR(2) on P, AR(1) on T, AR(1) $T \times P$	$\mu + T \times P$	No
2	AR(2) on P, AR(2) on T, AR(1) $T \times P$	$\mu + T \times P$	No
3	AR(2) on P, AR(2) on T, AR(1) $T \times P$	$\mu + T \times P$	Yes

Table 4: Simulation example: iterations in the second scenario. The algorithm converged to the true solution

Variable	Number of levels	Symbol
scanner	17	S
prior scanner	21	P
operation	11	O
product	25	n.a.
recipe	116	n.a.
reticle	107	R

Table 5: Context variables. The variables *product* and *recipe* are not considered in the selection algorithm.

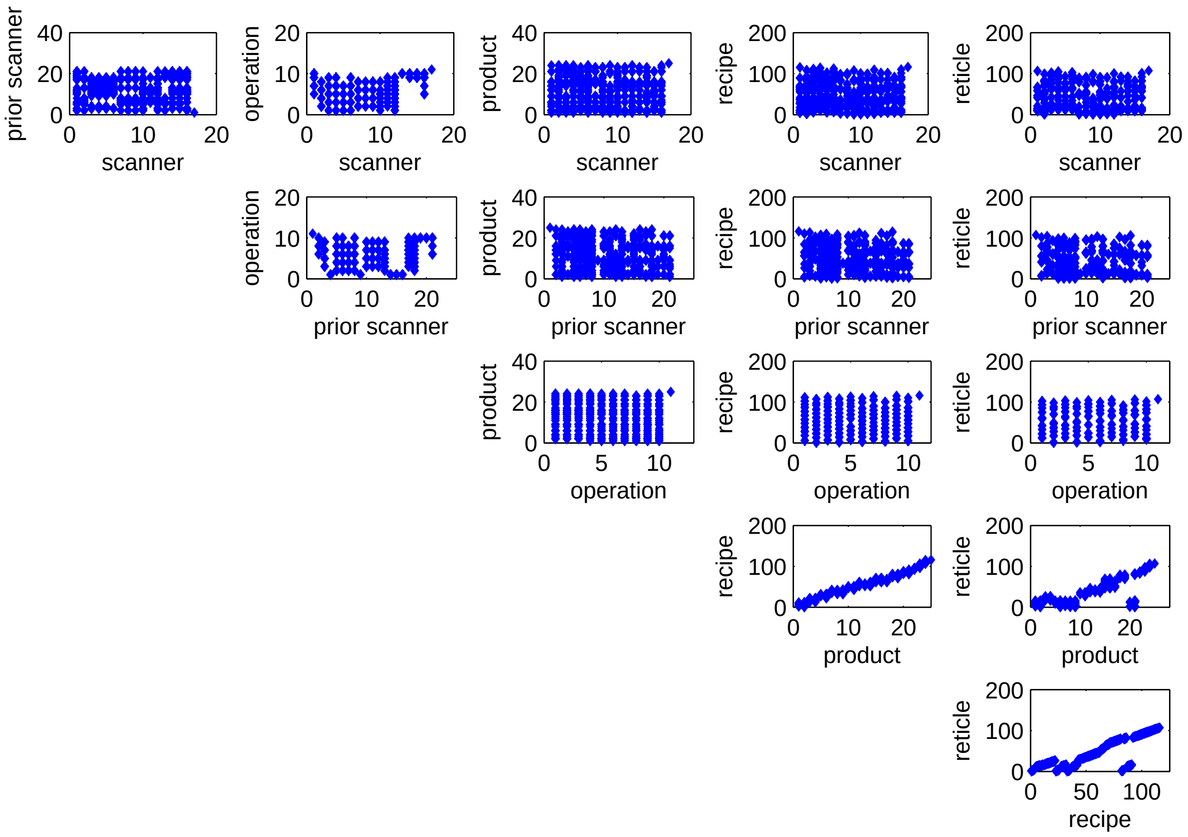

Figure 4: Pairwise plots of the context variables

Iteration No	Regression model	ANOVA model	Convergence ?
1	AR(2) on $S \times P$, on $S \times O$ on $P \times R$, on $P \times O$	$S \times O + P \times R + P \times O + S \times R$	No
2	AR(2) on S	$S \times R + P \times O$	No
3	AR(2) on S	$S \times R + P \times O$	Yes

Table 6: Iterations of the variable selection algorithm using the first 5000 rows of the *yshift* response data

ϕ_1			ϕ_2			
scanner	estimate	s.e.	*scanner*	estimate	s.e.	Unit Root ?
1	0.27	0.174	1	0.83	0.41	1
2	0.37	0.09	2	0.39	0.17	1
3	0.31	0.09	3	0.27	0.16	0
4	0.18	0.09	4	0.33	0.18	0
5	0.36	0.09	5	0.39	0.16	1
6	0.38	0.09	6	0.48	0.17	1
7	0.47	0.08	7	0.72	0.13	1
8	0.26	0.13	8	0.34	0.27	1
9	0.24	0.11	9	0.65	0.22	1
10	0.33	0.13	10	0.56	0.26	1
11	-0.74	0.77	11	0.00	1.37	1
12	-0.03	0.16	12	0.10	0.37	0
13	0.27	0.18	13	0.28	0.42	1
14	0.14	0.13	14	0.70	0.26	1
15	0.26	0.09	15	0.31	0.20	0
16	0.15	0.11	16	0.72	0.23	1

Table 7: Autoregressive parameter estimates using *yshift* as the response (1st 5000 rows)

Regression model

Source	d.f.	SS	MS	F	$Prob > F$
Model	32	0.013	4.0E-04	9.74	> 0.0001
Error	4966	0.205	4.1E-05		
Total	4998	0.218			

ANOVA model

Source	d.f.	SS	MS	F	$Prob > F$
Model	312	3.001	9.6E-03	220.34	> 0.0001
Error	4686	0.205	4.4E-05		
Total	4998	3.206			

Effects Test

Source	no of prms	d.f.	SS	F	$Prob > F$
$S \times R$	1545	247	9.8E-01	91.16	> 0.0001
$P \times O$	171	55	4.0E-02	16.48	> 0.0001

Table 8: Summary of the estimated model using the first 5000 rows of the *yshift* response data

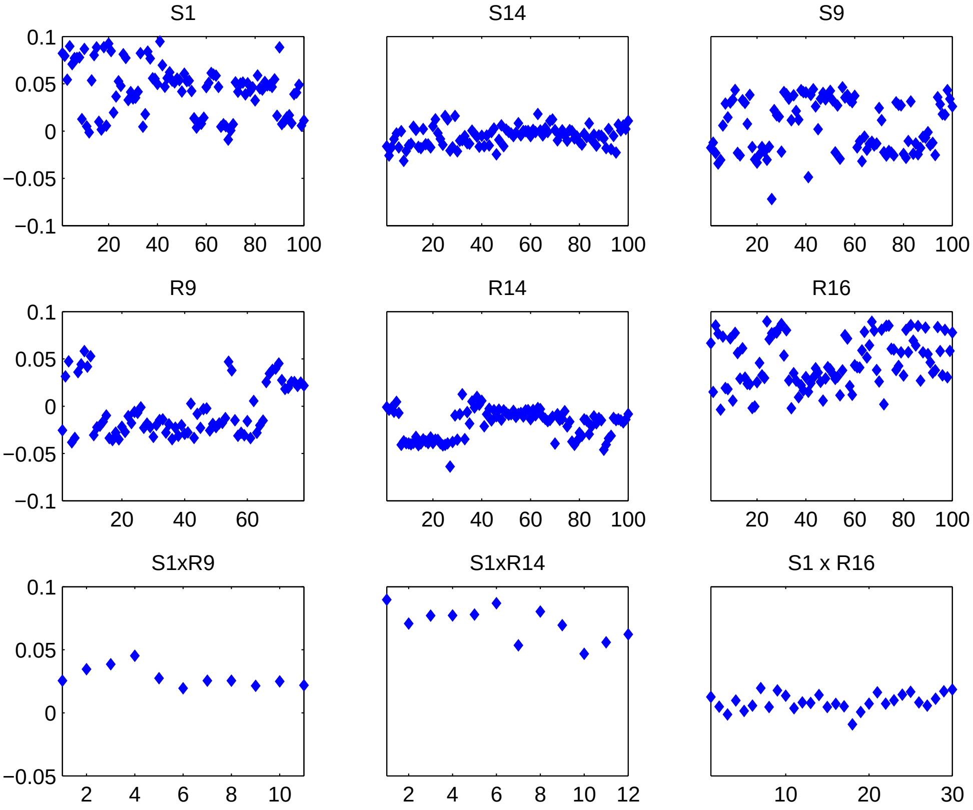

Figure 5: Residuals of the regression model on different scanners, reticles and scanner-reticle combinations (1st 5000 rows of $yshift$). As can be seen $scanner$ only (top) or $reticle$ only (middle) models are not able to account for the mean shifts in the data adequately. $scanner \times reticle$ model (bottom) accounts for the mean shifts adequately.

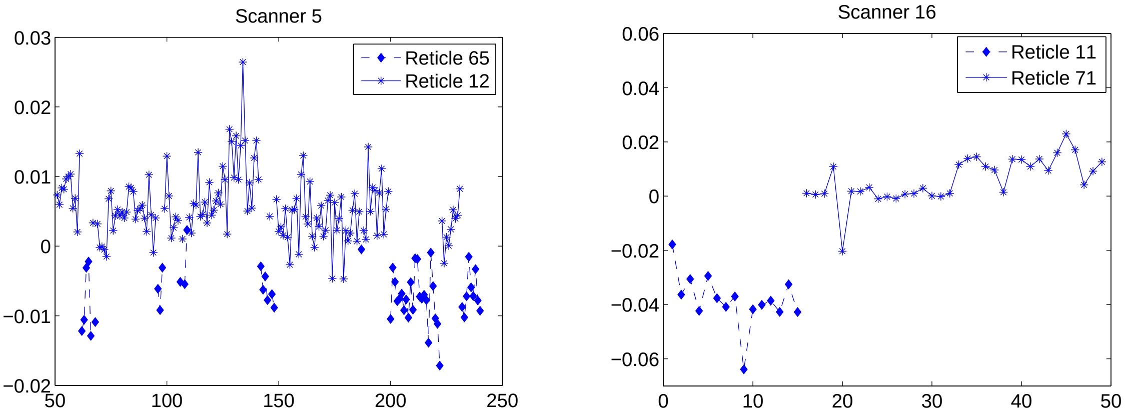

Figure 6: Disturbance values for different $scanner - reticle$ combinations from the 1st 5000 rows of $yshift$ after removing the offsets due to $P \times O$. Left panel: On $scanner$ 5 and with $reticle$ 65 (diamond marker) and $reticle$ 12 (asterisk marker). Right panel: On $scanner$ 16 and with $reticle$ 11 (diamond marker) and $reticle$ 71 (asterisk marker)

	κ	$Prob > \kappa$
$scanner$ 5	14.21	1.2E-04
$scanner$ 6	9.23	0.019
$scanner$ 7	15.50	2.7E-05
$operation$ 3	6.34	0.333
$operation$ 4	13.91	1.732E-04
$operation$ 5	7.04	0.220
$reticle$ 10	5.34	0.105
$reticle$ 11	3.07	0.753
$reticle$ 12	8.69	0.031

Table 9: Fisher's white noise test results on the residuals of the ANOVA model on $scanner$s, $operation$s and $reticle$s. The 1st 5000 rows of the $yshift$ response data was used.

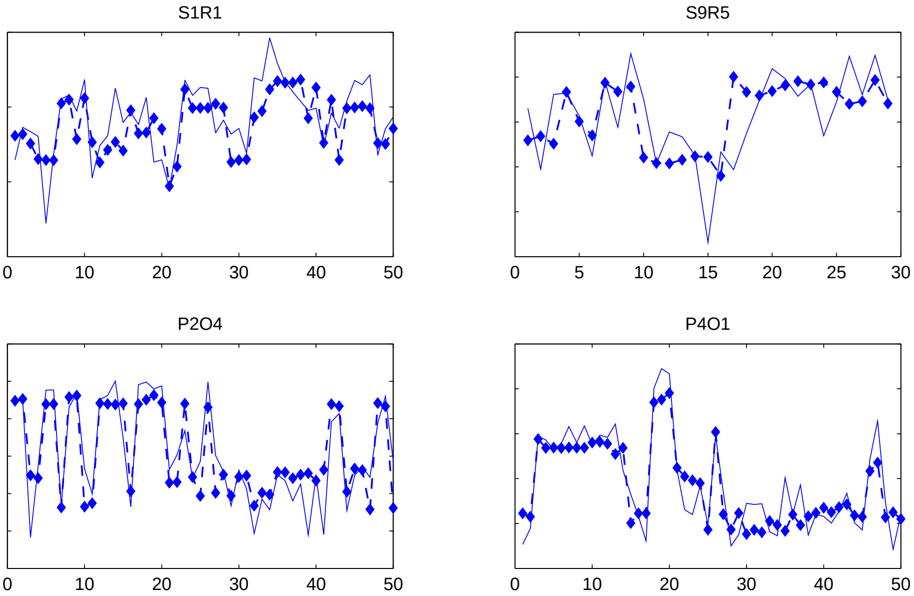

Figure 7: Predictions (dash line with asterisk marker) and the actual measurements (solid line) of the 2nd 5000 rows of the $yshift$ data on different $scanner-reticle$ (top) and $prior\ scanner-operation$ (bottom) combinations. The predictions are computed using the parameter estimates obtained from the 1st 5000 rows. Due to the classified nature of the data the y- axis labels are not given.