A Methodology for Efficient Dynamic Spatial Sampling and Reconstruction of Wafer Profiles

Seán McLoone *Senior Member, IEEE*, Adrian Johnston, and Gian Antonio Susto *Member, IEEE*

Abstract—In semiconductor manufacturing, metrology is generally a high cost, non-value added operation that impacts significantly on cycle time. As such, reducing wafer metrology continues to be a major target in semiconductor manufacturing efficiency initiatives. A novel data-driven spatial dynamic sampling methodology is presented that minimises the number of sites that need to be measured across a wafer surface while maintaining an acceptable level of wafer profile reconstruction accuracy. The methodology is based on analysing historical metrology data using Forward Selection Component Analysis (FSCA) to determine, from a set of candidate wafer sites, the minimum set of sites that need to be monitored in order to reconstruct the full wafer profile using statistical regression techniques. Dynamic sampling is then implemented by clustering unmeasured sites in accordance with their similarity to the FSCA selected sites, and temporally selecting a different sample from each cluster. In this way, the risk of not detecting previously unseen process behaviour is mitigated. We demonstrate the efficacy of the proposed methodology using both simulation studies and metrology data from a semiconductor manufacturing process.

Note To Practitioners–Here we consider a practical metrology problem encountered in semiconductor manufacturing, namely, design of a wafer measurement plan where measurements have to be taken from several sites across a wafer surface in order to monitor the accuracy and spatial consistency of a given processing step. A methodology is presented for designing a dynamic measurement plan for this scenario that minimises the number of sites that need to be measured while maintaining an acceptable level of process visibility in terms of wafer profile reconstruction accuracy and the ability to detect previously unseen process behaviour. Our approach can significantly reduce the number of sites that need to be measured on each wafer enabling greater throughput on metrology tools. While developed for wafer metrology in semiconductor manufacturing, the methodology can potentially be applied to the design of measurement plans for any surface whose variation exhibits strong spatial correlation.

Index Terms—Dynamic Sampling, Forward Selection Component Analysis, Metrology, Principal Component Analysis, Semiconductor Manufacturing, Wafer Site Selection.

S. McLoone (corresponding author) is with Queen's University Belfast, Belfast, Northern Ireland, United Kingdom. E-mail: s.mcloone@ieee.org.

A. Johnston is with Seagate Technology, United Kingdom. E-mail: adrian.b.johnston@seagate.com.

G.A. Susto is with the Department of Information Engineering, University of Padova, Italy. E-mail: gianantonio.susto@dei.unipd.it.

I. INTRODUCTION

METROLOGY is a critical activity in industry [21], [23] and, in particular in semiconductor manufacturing [13], [27], where it is increasingly becoming the focus of attention as feature sizes continue to shrink and wafer diameters increase from the current industrial standards of 200 mm and 300 mm to 450 mm [11]. In Chemical Vapor Deposition (CVD), for example, the feature of interest is the depth of the layer/film of material deposited on the wafer [22], while in plasma etching it is the dimensions (depth and width) of the trenches etched into the wafer surface [18]. The spatial variation of these features over the full wafer, defined as the wafer profile, and their wafer-to-wafer (temporal) variation need to be tightly controlled to meet the demanding specifications of current and next generation semiconductor devices. This is achieved using Advanced Process Control (APC), typically run-to-run control, with metrology employed in one of two contexts; as a feedback control signal to adapt the operation of the current process for the next wafer, and/or as a feedforward control signal to enable the next process to compensate for the deviations introduced by the current process.

Ideally each wafer should be measured at a large number of locations to provide detailed performance information for APC, and also for Predictive Maintenance (PdM) and product quality assessment activities [10], [25], [26], [29]. In practice, such extensive metrology is not feasible due to the impact on cycle-time and the high cost of the precision metrology technologies needed [24]. Consequently, standard practice is to undertake limited sampling both temporally [9] and spatially, [1], [4], [22] and to rely on the information this provides for process monitoring and control.

A key consideration is how to determine the number, and location of measurement sites. Initial measurement plans are usually determined by process engineers based on *a priori* knowledge of wafer spatial variability patterns for a given process, or by employing space filling sampling designs such as Latin Hypercube Sampling [12]. Additional sites are then added during ramp-up and production to monitor new process issues as they arise. Typically, the spatial correlation structure of wafer variability is not explicitly taken into account in these approaches with the result that substantial redundancy can arise within measurement plans. Furthermore, the level of redundancy can increase over time as processes mature and process issues are engineered out.

In recent years there has been increasing interest in developing data driven wafer measurement plan optimization methodologies that can take account of spatial correlation to further reduce the number of sites that need to be measured. Vincent *et al*. [25] developed a methodology based on Principal Component Analysis (PCA) modelling and minimum-variance estimation. They considered both within wafer spatial patterns and temporal correlation patterns in their formulation, but concluded that only spatial patterns were present in the practical litho-etch process case study they used to validate their approach.

Borgoni *et al*. [2] proposed a simulated annealing and spatial prediction strategy to optimise the selection of a reduced set of sites from a larger candidate grid for a silicon oxide deposition process.

In [28], Zheng addresses the issue of optimal site

selection for monitoring and wafer map interpolation of electrical metrology data in the context of developing effective Fault Detection and Classification (FDC) schemes. Four approaches are investigated. Two are supervised selection techniques where sites are sequentially selected based on their utility as inputs for specified wafer classification tasks, and two are unsupervised techniques where the objective is to identify a subset of sites that best represent the remaining sites. Here, supervised refers to approaches that employ both input and target output data (a.k.a. labeled data) to guide the learning process, while unsupervised refers to approaches which learn based on input data only (a.k.a. unlabeled data). The first of the unsupervised approaches is a two stage methodology where candidate sites are initially clustered using k-means clustering and then a PCA based within-cluster site selection technique employed to select a subset of sites from within each cluster. The second approach involves estimating a bootstrapped forward selection Partial Least Square (PLS) model with measured sites as input and unmeasured sites as outputs to be predicted. Zheng concludes that the two stage clustering/PCA methodology yields the best FDC performance.

In [7], a Gaussian process model based sequential measurement strategy is developed where, for each wafer, an initial set of sites is measured and used to estimate a Gaussian Process (GP) model, which then guides the selection of additional measurements sites to update the GP model until its prediction error on test sites is within an acceptable level. The distinguishing characteristic of this approach is that it does not require historical data for model building, however the number and location of measurement sites change from wafer to wafer,

and the need for an initial set of measurements for in-line model estimation means that it is not suited to low measurement density scenarios.

In [15] we introduced a methodology for optimum wafer site selection for wafer sampling plan design based on Forward Selection Component Analysis (FSCA) [16], [17], an unsupervised extension of forward selection regression that determines the contribution that individual sites make to the variability observed in a process across a set of candidate wafer sites. In addition to eliminating redundancy, the methodology provides for accurate wafer profile reconstruction through the use of statistical regression models to predict unmeasured sites.

One of the concerns when using a reduced sampling plan is that there is risk of not detecting previously unseen abnormal process behavior. Here we extend our previous work in [15] to address this concern by developing a novel dynamic spatial sampling methodology that improves wafer coverage temporally with minimal information loss in terms of the ability to reconstruct the full wafer profile with the optimally selected FSCA sites. The paper also provides, for the first time, a complete description of the underpinning FSCA based site selection and wafer profile reconstruction methodology.

The remainder of the paper is organised as follows. The FSCA based site selection methodology is introduced in Section II. The dynamic sampling framework is then developed in Section III. Using metrology data from a semiconductor manufacturing case study, and two simulated datasets (described in Section IV), we demonstrate the efficacy of the overall methodology in Section V. Finally, conclusions are presented in Section VI.

II. FSCA Methodology

Let $\mathbf{X} \in \mathbb{R}^{N \times V}$, be a matrix of historical metrology data for a given process, where N is the number of samples (wafers measured), V is the number of candidate sites (measurement locations) on each wafer, and $N > V$. Thus, x_{ji} denotes the feature measurement taken at site i on the j-th wafer (e.g. etch depth, film thickness) and column vector $\mathbf{x}_i \in \mathbb{R}^{N \times 1}$ is a vector of measurements taken at site i for all N wafers in the data set. We are interested in determining the minimum set of sites from candidate set $\mathcal{I} = \{1, 2, ..., V\}$ to measure in order to capture the information contained in all sites. In terms of the data matrix $\mathbf{X}$ we can define this problem as searching for the subset of columns of $\mathbf{X}$ that best represent the information contained in all columns of $\mathbf{X}$.

Classical PCA [8] can be used to check for redundancy in the measurements (with respect to observing process driven wafer surface variability) and to establish a lower bound ϕ_{PCA} on the number of measurement sites needed [15]. If $\phi_{\mathrm{PCA}} << V$ significant redundancy exists, in which case it may be possible to select a reduced set of sites that yield similar levels of process visibility.

However, PCA does not tell us what sites to measure since each PCA loading (latent variable) is a linear combination of all candidates sites, and in general the contribution of individual sites does not reveal which sites are most important. In particular, identification of key sites is difficult if they are part of a highly correlated group, as their contribution to a loading will be distributed evenly across the group.

A. FSCA Wafer Site Selection

As introduced in [15], FSCA provides a solution to this problem by sequentially searching for individual sites that make the greatest contribution in terms of explaining the process variability observed across all sites. FSCA, is essentially the unsupervised equivalent to Forward Selection Regression [17], a well established technique for variable selection [5], [6]. It employs an iterative procedure consisting of the following steps:

1) *Initialization*: Set index set $\mathcal{I}_{\mathrm{FSC}} = \{\}$, iteration count $k = 1$ and $\widetilde{\mathbf{X}}_1 = \mathbf{X} - \bar{\mathbf{X}}$, where $\bar{\mathbf{X}}$ is as defined in eqt. (14), i.e. $\widetilde{\mathbf{X}}_1 =$ mean centered $\mathbf{X}$.

2) *Search*: Identify the index of the variable with maximum contribution to $\widetilde{\mathbf{X}}_k$ and add it to $\mathcal{I}_{\mathrm{FSC}}$. This is computed as

$$i^* = \operatorname*{argmin}_{i=1,...,V} \left\| \widetilde{\mathbf{X}}_k - \widehat{\mathbf{X}}_k(\widetilde{\mathbf{x}}_i) \right\|_F^2 \qquad (1)$$

where $\widetilde{\mathbf{x}}_i$ is the i-th column of $\widetilde{\mathbf{X}}_k$, $||.||_F$ is the Frobenius norm and $\widehat{\mathbf{X}}_k(\widetilde{\mathbf{x}}_i)$ is the estimate of $\widetilde{\mathbf{X}}_k$ obtained by regressing on $\widetilde{\mathbf{x}}_i$, that is:

$$\widehat{\mathbf{X}}_k(\widetilde{\mathbf{x}}_i) = \frac{\widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\mathsf{T}}{\widetilde{\mathbf{x}}_i^\mathsf{T} \widetilde{\mathbf{x}}_i} \widetilde{\mathbf{X}}_k. \qquad (2)$$

Vector $\widetilde{\mathbf{x}}_{i^*}$ is defined as a Forward Selection Component (FSC) and i^* is the index of the selected wafer site.

3) *Deflation*: Remove the contribution of $\widetilde{\mathbf{x}}_{i^*}$ from $\widetilde{\mathbf{X}}_k$:

$$\widetilde{\mathbf{X}}_{k+1} = \widetilde{\mathbf{X}}_k - \widehat{\mathbf{X}}_k(\widetilde{\mathbf{x}}_{i^*}). \qquad (3)$$

4) *Repeat*: While the cumulative variance explained (CVE) is less than a specified threshold τ_{CVE}, e.g. 99%, set $k = k + 1$ and repeat from *Step 2*. Here, the cumulative variance

explained by the first k FSC components is defined as

$$v_{\mathrm{c}}(k) = \frac{\left\|\widetilde{\mathbf{X}}_1\right\|_F^2 - \left\|\widetilde{\mathbf{X}}_1 - \widetilde{\mathbf{X}}_{k+1}\right\|_F^2}{\left\|\widetilde{\mathbf{X}}_1\right\|_F^2} \times 100. \quad (4)$$

5) *End*: Output the optimum number of sites k^*, the CVE vector $\mathbf{v_c}$, defined as

$$\mathbf{v_c} = [v_{\mathrm{c}}(1), v_{\mathrm{c}}(2), ..., v_{\mathrm{c}}(k^*)],$$

and the prioritized list of metrology sites $\mathcal{I}_{\mathrm{FSC}}$.

The final number of metrology sites k^* selected by FSCA is bounded in the range

$$\phi_{\mathrm{PCA}} \leq k^* \leq \mathrm{rank}(\mathbf{X}) \leq V$$

where ϕ_{PCA} is the number of principal components needed to exceed the specified CVE threshold τ_{CVE}. This threshold should be chosen taking into account the signal to noise ratio (SNR) of the metrology data. When the SNR is not known, analysis of the CVE trend (Scree plot) for a PCA analysis of the data can be used to estimate an appropriate threshold [8]. The PCA lower bound ϕ_{PCA} follows from the fact that PCA yields the linear combination of the variables that maximises the observed variance, hence any linear combination of a subset of these variables, as provided by FSCA, can only approach, but never exceed this optimum [16].

It should be noted that FSCA is not guaranteed to identify the optimum solution for a given problem. Determining the subset of columns from $\mathbf{X}$ which best represents $\mathbf{X}$ is an NP hard problem and not tractable for large V [6]. In this context, FSCA represents a pragmatic compromise, which although not guaranteed to be optimal, consistently yields good results at an acceptable computational cost.

B. Wafer Profile Reconstruction

In general the reduced set of measurements identified using FSCA are insufficient to accurately reconstruct the 3D profile of a wafer surface using interpolation techniques such as Biharmonic Spline Interpolation (BSI) [19]. However, as demonstrated in [15], unmeasured sites can be estimated using linear regression models, herafter referred to as Wafer Metrology Reconstruction (WMR) models, enabling accurate reconstructions to be achieved. The WMR models, which employ the measured sites as regressors, are trained using the historical metrology data. Specifically, denoting $\vec{\mathbf{x}}_{\mathrm{FSC}} \in \mathbb{R}^{1 \times k^*}$ as the measurements from the FSCA selected sites for a given wafer, the unmeasured sites $\vec{\mathbf{x}}_{\mathrm{u}} \in \mathbb{R}^{1 \times (V-k^*)}$ can be estimated using a linear regression model

$$\vec{\mathbf{x}}_{\mathrm{u}} = [\vec{\mathbf{x}}_{\mathrm{FSC}} \ 1]\boldsymbol{\beta}, \quad (5)$$

where $\boldsymbol{\beta} \in \mathbb{R}^{k^* \times (V-k^*+1)}$ is the matrix of regression coefficients. Given historical metrology data $\mathbf{X}$, partitioned into $\mathbf{X}'_{\mathrm{FSC}} \in \mathbb{R}^{N \times k^*}$ (metrology sites) and $\mathbf{X}_{\mathrm{u}} \in \mathbb{R}^{N \times (V-k^*)}$ (unmeasured sites), that is:

$$\mathbf{X}'_{\mathrm{FSC}} = \{\mathbf{x}_i | i \in \mathcal{I}_{\mathrm{FSC}}\} \text{ and } \mathbf{X}_{\mathrm{u}} = \{\mathbf{x}_i | i \in \mathcal{I}/\mathcal{I}_{\mathrm{FSC}}\},$$

$\boldsymbol{\beta}$ is computed as the least squares estimate:

$$\boldsymbol{\beta} = \left(\mathbf{X}_{\mathrm{FSC}}^{\mathsf{T}} \mathbf{X}_{\mathrm{FSC}}\right)^{-1} \mathbf{X}_{\mathrm{FSC}}^{\mathsf{T}} \mathbf{X}_{\mathrm{u}}. \quad (6)$$

where $\mathbf{X}_{\mathrm{FSC}} = \left[\mathbf{X}'_{\mathrm{FSC}} \ \vec{\mathbf{1}}\right]$, and $\vec{\mathbf{1}}$ is a column vector of ones of appropriate length.

III. DYNAMIC SAMPLING

The basic assumption of the data driven wafer site selection methodologies, as presented here, is that the historical data used to select the optimum sites and train the WMR models captures the full range of process behaviour. As such, one concern with implementing a fixed sampling plan with

a reduced set of sites is that previously unseen spatially localized process behaviour may go undetected. This risk can be mitigated by employing a dynamic sampling plan that measures a different subset of locations on each wafer, such that all candidate measurement sites are visited periodically. Depending on the sampling algorithm design, the periodicity may be different for different sites, hence, defining the 'site sampling interval' as the number of wafers processed without a given site being measured (i.e. a value that is one less than the site sampling rate), we introduce the metric *Maximum Site Sampling Interval* (MSSI), to define the largest sampling interval among all sites. MSSI is defined in this way so that MSSI = 0 when all sites are measured on each wafer, and MSSI = ∞ if a site is never measured. Hence MSSI takes a value in the range $[0 \; \infty]$.

The underlying requirement for detectability is that any new process behaviour that appears must persist beyond the MSSI of the sampling method. Otherwise it is not guaranteed to be detected. Obviously this implies that up to MSSI wafers may be processed before the abnormal behaviour is detected. Hence, it is desirable to keep the MSSI as low as possible.

Since the FSCA selected sites are the optimum subset of sites to measure for accurate reconstruction of the wafer profile, any deviation from this will lead to diminished performance. The challenge, therefore, is to have a dynamic sampling plan with low MSSI that minimizing the impact on wafer profile reconstruction accuracy. This can be achieved as follows.

Given V sites, with V_m measured and V_u estimated virtually we cluster all unmeasured sites around the V_m measured sites based on similarity.

The simplest approach to assessing similarity is to evaluate the correlation between sites. Denoting $\mathcal{I}_\mathrm{m} = \mathcal{I}_\mathrm{FSC}$ as the index set for measured sites, $\mathcal{I}_\mathrm{u} = \mathcal{I}/\mathcal{I}_\mathrm{FSC}$ as the index set for unmeasured sites, and $\mathcal{I}_\mathrm{m}[k]$ as the k-th element of $\mathcal{I}_\mathrm{m}$, the clusters are defined as

$$\mathcal{C}_k = \mathcal{C}_k^\mathrm{u} \cup \mathcal{I}_\mathrm{m}[k] \tag{7}$$

with

$$\mathcal{C}_k^\mathrm{u} = \{i | \mathrm{corr}(\mathbf{x}_i, \mathbf{x}_p) \geq \max_j(\mathrm{corr}(\mathbf{x}_i, \mathbf{x}_j)),$$
$$p = \mathcal{I}_\mathrm{m}[k], j \in \mathcal{I}_\mathrm{m}, j \neq p, i \in \mathcal{I}_\mathrm{u}\} \tag{8}$$

for $k = 1, 2, .., V_\mathrm{m}$. Dynamic sampling is then performed by sequentially selecting one site from each of the V_m clusters, $\mathcal{C} = \{C_1, C_2, ..., C_{V_\mathrm{m}}\}$, at each process iteration. Hence, at the tth process iteration the sampling plan is:

$$\mathcal{I}_\mathrm{DS}(t) = \{C_1[i_1(t)], C_2[i_2(t)], ..., C_{V_\mathrm{m}}[i_{V_\mathrm{m}}(t)]\} \tag{9}$$

where

$$i_k(t) = t \bmod (\mathrm{card}(C_k)) + 1. \tag{10}$$

In this way each site is guaranteed to be visited at least once every $\max_k(\mathrm{card}(C_k))$ wafers.

An alternative cluster assignment strategy, which is computationally much more expensive to determine, is to assign each unmeasured site to the FSCA cluster where it has the least impact on prediction performance (as defined below) when used instead of the FSCA site for that cluster.

With reference to eqt. (5) and eqt. (6), the FSCA estimate of unmeasured sites for all N wafers in dataset $\mathbf{X}$ can be expressed as

$$\hat{\mathbf{X}}_\mathrm{u} = \mathbf{\Phi}(\mathbf{X}_\mathrm{FSC})\mathbf{X}_\mathrm{u} \tag{11}$$

where

$$\mathbf{\Phi}(\mathbf{X}_\mathrm{FSC}) = \mathbf{X}_\mathrm{FSC}\left(\mathbf{X}_\mathrm{FSC}^\mathsf{T}\mathbf{X}_\mathrm{FSC}\right)^{-1}\mathbf{X}_\mathrm{FSC}^\mathsf{T} \tag{12}$$

and the prediction for all V sites is given by

$$\hat{\mathbf{X}} = \mathbf{\Phi}(\mathbf{X}_{\text{FSC}})\mathbf{X}. \quad (13)$$

Defining $\bar{\mathbf{X}}$ as a matrix with column entries equal to the mean of the corresponding columns of $\mathbf{X}$, that is:

$$\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_i | \bar{\mathbf{x}}_i = \bar{x}_i \mathbf{1}_N, \bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ji}, i = 1, ..., V\} \quad (14)$$

the normalized mean squared error (NMSE) over all sites and all wafers can be expressed as

$$\mathcal{N}(\mathbf{X}_{\text{FSC}}) = \frac{\|\mathbf{X} - \mathbf{\Phi}(\mathbf{X}_{\text{FSC}})\mathbf{X}\|_F^2}{\|\mathbf{X} - \bar{\mathbf{X}}\|_F^2} \times 100. \quad (15)$$

Finally, denoting $\mathbf{X}_{\text{FSC}}^{(i)}(\mathbf{x})$ as matrix $\mathbf{X}_{\text{FSC}}$ with its i-th column replaced by $\mathbf{x}$, the optimal clusters can be defined as

$$\mathcal{C}_k = \mathcal{C}_k^{\text{u}^*} \cup \mathcal{I}_{\text{m}}[k] \quad (16)$$

where

$$\mathcal{C}_k^{\text{u}^*} = \{i | \mathcal{N}(\mathbf{X}_{\text{FSC}}^{(k)}(\mathbf{x}_i)) \le \max_j(\mathcal{N}(\mathbf{X}_{\text{FSC}}^{(j)}(\mathbf{x}_i))),$$

$$j \in \mathcal{I}_{\text{m}}, j \ne k, i \in \mathcal{I}_{\text{u}}\} \quad (17)$$

for $k = 1, 2, .., V_{\text{m}}$.

A consequence of dynamic sampling is that the regressors for the WMR models change at each process iteration, and hence new models need to be computed as follows. Denoting $\mathbf{x}_{\text{DS}}(t)$ as the dynamically measured sites at process iteration t, as defined by $\mathcal{I}_{\text{DS}}(t)$, and partitioning the historical data matrix $\mathbf{X}$ as $[\mathbf{X}'_{\text{DS}} \; \mathbf{X}'_{\text{u}}]$ then

$$\mathbf{x}_{\text{u}}(t) = [\mathbf{x}_{\text{DS}}(t) \; 1]\boldsymbol{\beta} \quad (18)$$

where $\boldsymbol{\beta}$ is computed as in eqt. (6), but with $\mathbf{X}_{\text{FSC}}$ replaced by $\mathbf{X}_{\text{DS}} = [\mathbf{X}'_{\text{DS}} \; \vec{\mathbf{1}}]$.

The overall FSCA clustering-based sequential dynamic sampling (SDS) methodology, which consists of an off-line training phase followed by an on-line sampling and virtual metrology phase, is summarised in Algorithm 1.

Algorithm 1: FSCA Clustering-based Sequential Dynamic Sampling (SDS)

Offline:

    Data: $\mathbf{X}, \tau_{\text{CVE}}, \mathcal{I}$

      1. $\mathcal{I}_{\text{FSC}} \leftarrow \text{FSCA}(\mathbf{X}, \tau_{\text{CVE}})$

      2. $\mathcal{C} \leftarrow \text{cluster}(\mathbf{X}, \mathcal{I}, \mathcal{I}_{\text{FSC}})$

    Result: $\mathcal{C}, \mathcal{I}_{\text{FSC}}$

Online (at process iteration t):

    Data: $\mathbf{X}, \mathcal{C}, t$

      1. $\mathcal{I}_{\text{DS}}(t) \leftarrow \text{SDS}(\mathcal{C}, t)$

      2. $\mathbf{x}_{\text{DS}}(t) \leftarrow \text{Measure}(\mathcal{I}_{\text{DS}}(t))$

      3. $\boldsymbol{\beta}(t) \leftarrow \text{WMRmodel}(\mathbf{X}, \mathcal{I}_{\text{DS}}(t))$

      4. $\hat{\mathbf{x}}_{\text{u}}(t) \leftarrow \mathbf{x}_{\text{DS}}(t)\boldsymbol{\beta}(t)$

    Result: $\mathcal{I}_{\text{DS}}, \mathbf{x}(t) = [\mathbf{x}_{\text{DS}}(t) \; \hat{\mathbf{x}}_{\text{u}}(t)]$

IV. CASE STUDIES

This section introduces three cases studies, which will be used to investigate and demonstrate the effectiveness of the proposed SDS methodology.

Case study 1 [Industrial]: This consists of wafer metrology data for a process used in read-write head formation within disk drive semiconductor manufacturing. This dataset was collected over several weeks from a single production tool for the process in question using a static 50-site measurement plan ($V = 50$), yielding a dataset consisting of metrology for $N = 316$ wafers of different product types. For confidentiality reasons the data has been normalized. Fig. 1 shows four sample wafer profiles from this dataset. As can be seen the process exhibits substantial variation with regard to the shape of the wafer profiles produced.

Case study 2 [RBF]: In this case study wafer profiles are simulated as sums of randomly generated Gaussian Radial Basis Functions (RBF) defined on

Fig. 1: [Industrial] Sample wafer surfaces belonging to the industrial case study.

Fig. 2: [RBF] Sample wafer surfaces generated by the RBF model: $S_f = 0.3$ (left), $S_f = 0.6$ (right).

the unit radius disc centred on the origin, that is:

$$z(x,y) = \sum_{i=1}^{N_g} h_i exp(\frac{(x - c_{x_i})^2 + (y - c_{y_i})^2}{S_f^2}) + \epsilon \quad (19)$$

where $h_i \sim N(0,1)$, $c_{x_i}, c_{y_i} \sim U(-1,1)$, $z(x,y)$ is the profile height at the coordinates (x,y), and $\epsilon \sim N(0,0.02)$ simulates measurement noise. The smoothness of the resulting wafer profiles is con-

trolled by the number of RBFs (N_g) and the spread factor S_f. In particular, S_f can be adjusted to vary the spatial correlation of the variation occurring on the synthesized wafer surfaces. Fig. 2 shows typical wafer profiles generated by the RBF model for $N_g = 100$, $S_f = 0.3$ (left) and $N_g = 100$, $S_f = 0.6$ (right).

Fig. 3: [Zernike] Sample wafer surfaces generated by the Zernike model.

Case study 3 [Zernike]: Here wafer profiles are generated based on randomly weighted combinations of Zernike polynomials. Zernike polynomials are an infinite set of orthonormal functions defined on a unit disk that are widely used in the field of optics to describe complex non-rotationally symmetric surfaces over a circular domain [3]. A finite set of Zernike polynomials is defined by the *degree N* and a function Z_n^m parameterized by two indices n and m, with $n = 0, 1, \ldots, N$ and $m = -n, -n+1, \ldots, n-1, n$. For a given N the generic Zernike polynomial is defined in polar coordinates as

$$Z_n^m(\rho, \theta) = R_n^m(\rho) \begin{cases} \cos(m\theta) & \text{if } m \geq 0 \\ \sin(m\theta) & \text{if } m < 0 \end{cases}$$

where $\theta \in [0, 2\pi]$, $\rho \in [0, 1]$, and the radial function $R_n^m(\rho) = 0$ if $(n - m)$ is odd, and

$$R_n^m(\rho) = \sum_{l=0}^{(n-m)/2} \frac{(-1)^l (n-l)!}{l! \left[\frac{1}{2}(n+n) - l\right]! \left[\frac{1}{2}(n-n) - l\right]!} \rho^{n-2l}$$

if $(n - m)$ is even. For the case study $N = 7$ giving a basis set of 36 Zernike polynomials, denoted as Z_i, $i = 1, 2, .., 36$. Wafer surfaces are then generated as

$$z(\rho, \theta) = \sum_{i=1}^{36} \alpha_i Z_i(\rho, \theta) \tag{20}$$

where $\alpha_i \sim N(0, 8\exp(-0.3i))$. Fig. 3 shows typical wafer profiles generated by this model.

V. EXPERIMENTAL RESULTS

A. Performance Evaluation

To provide a statistically robust evaluation of the performance of the various wafer sampling and profile reconstruction approaches Monte Carlo cross-validation (MCCV) is employed [14], [20]. Here the dataset of N samples is randomly split into:

- a training dataset of qN samples (where $0 < q < 1$) used to estimate the FSCA sites and construct the WMR models for the unmeasured sites; and
- a validation dataset of $(1 - q)N$ samples used to assess wafer profile reconstruction performance, as measured by the NMSE (eqt. 15).

This is repeated K times, each time with a different random split of the data, and the average and standard deviation of the reconstruction NMSE over the K repetitions used as performance metrics.

A second consideration with dynamic sampling is the frequency with which individual sites are visited (i.e. measured), as encapsulated by the *Maximum Site Sampling Interval* (MSSI) metric. For static sampling approaches MSSI $= \infty$, while for dynamic sampling techniques it is bounded below by $\lceil V/V_m \rceil - 1$, a bound which is achieved when sites are distributed evenly across all clusters, i.e. clusters are balanced. Thus, a *Wafer Observability Index* (WOI) can be defined as:

$$\text{WOI}(V_m) = \frac{\lceil V/V_m \rceil - 1}{\text{MSSI}(V_m)} \times 100 \qquad (21)$$

for $V_m < V$, to quantify the effectiveness of a given sampling algorithm in terms of providing wafer coverage with a given number measurements V_m. WOI provides a normalized measure (%) of the MSSI performance of a given method with respect to the theoretical minimum MSSI achievable.

B. FSCA based site selection

The results of a PCA and FSCA analysis of the industrial metrology dataset are summarized in Table I. The analysis reveals a very high level of redundancy in the measurement plan. With τ_{CVE} set at 99% there are only 5 significant modes of variation identified by PCA ($\phi_{\text{PCA}} = 5$) and this variation can be captured with as few as 7 FSCA selected sites ($k^* = 7$). Thus, using FSCA based site selection a 7-fold reduction in metrology (V/k^*) can be achieved for this process.

Fig. 4 shows the average NMSE performance for $K = 100$ MC simulations of FSCA as a function

TABLE I: [Industrial] Variance explained by the top 10 Principal Components (PCs) and Forward Selection Components (FSCs)

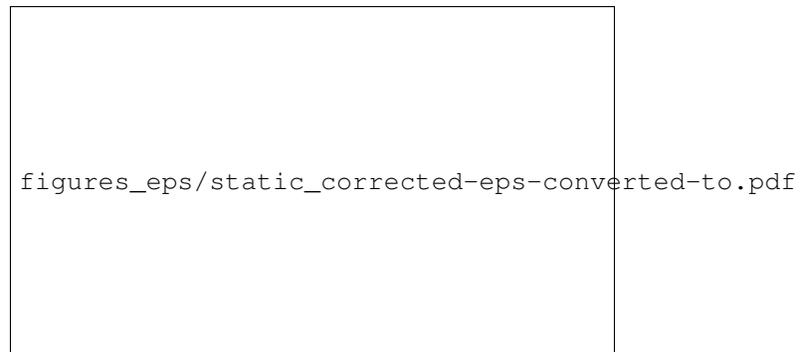| (a) PCA		(b) FSCA		
PC	Var. Exp. [%]	FSC	Site ID	Var. Exp. [%]
1	41.05	1	45	38.81
2	70.20	2	27	67.86
3	88.35	3	1	86.28
4	98.47	4	24	96.68
5	**99.07**	5	9	97.87
6	99.43	6	49	98.53
7	99.64	**7**	**14**	**99.02**
8	99.72	8	21	99.42
9	99.79	9	28	99.60
10	99.85	10	11	99.69

Fig. 4: [Industrial] Average NMSE over $K = 100$ MC simulations of WMR wafer profile reconstructions using Random and FSCA based metrology site selection and least squares WMR model estimation as a function of the number of measured sites.

of the number of measured sites (V_m). For comparison purposes the performances of WMR models based on randomly selected metrology sites are also included. As expected, FSCA based site selection strongly outperforms randomly selected sites. Fig. 5 shows a typical wafer profile reconstruction obtained using FSCA when $k^* = 7$. The actual profile
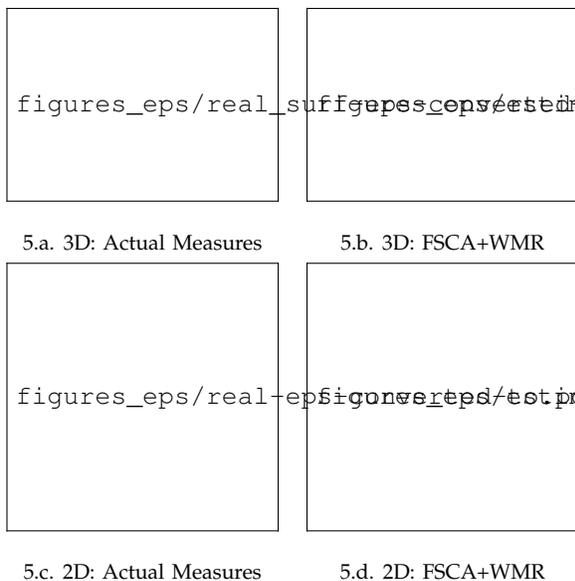
figures_eps/real_suffgepescenvestedmabeddsurf-eps-converted-to.pdf

5.a. 3D: Actual Measures   5.b. 3D: FSCA+WMR

figures_eps/real-epsi-conveneps/estimated_eps-convertedto.pdf

5.c. 2D: Actual Measures   5.d. 2D: FSCA+WMR

Fig. 5: [Industrial] Typical wafer profile reconstruction obtained using FSCA based site selection ($k = 7$): 3D plots (top) and corresponding contour plots (bottom) for the actual (left) and estimated wafer profile (right). Dots (•) and squares (□) indicate measured and WMR sites, respectively. The axes have been normalised for confidentially reasons.

is displayed on the left and the estimate profile on the right. The circular black markers in the contour plots show the locations of the measured sites while the square markers are the locations of the unmeasured sites estimated using WMR.

It may seem counterintuitive that linear WMR models yield such good performance in terms of profile reconstruction, since the profiles themselves are highly nonlinear, however, this is a misconception. The reason that the linear WMR models are effective is that they are not trying to reproduce the full nonlinear profile or explicitly consider the spatial coordinates; rather, they are tracking the relative changes from one site to another from the mean profile and these relative changes can

be approximated to be linear (at least locally). As such they can be adequately represented by linear models. The accuracy of the models, and hence ultimately the number of sites that need to be measured, is dictated by the level of spatial correlation across the wafer surface.

To illustrate the link between spatial correlation of wafer profile variation and metrology site selection, Fig. 6 shows the number of FSCA selected sites needed to achieve 99% wafer profile reconstruction accuracy as a function of the spread factor S_f when using the RBF model (Case study 2). Results are presented for two different values of N_g and the corresponding PCA lower bound is also plotted. The dataset used for each S_f, N_g combination consisted of $N = 500$ randomly generated wafers sampled using a uniform sampling grid of 0.1 units. This yielded 317 candidate metrology sites per wafer, over which reconstruction accuracy was computed. As expected, as S_f increases fewer metrology sites are needed. A similar pattern is observed with the Zernike model (Case study 3) as the model order N is reduced. For example, when $N = 6$, $k^* = 13$ and when $N = 3$, $k^* = 9$.

C. Dynamic Sampling

To evaluate the performance of the correlation and NMSE based FSCA clustering implementations of the proposed SDS methodology Monte Carlo simulations were performed for each of the three case studies. For the RBF and Zernike model cases studies 500 wafer datasets were generated, with the RBF and Zernike model parameters selected as $S_f = 0.6, N_g = 100$ and $N = 7$. For each Monte Carlo run $V = 50$ candidate metrology sites were randomly selected subject to the constraints that; (1) the Euclidean distance between any two sites was

Fig. 6: [RBF] A plot of the number of FSCA selected sites needed to achieve 99 % reconstruction accuracy and the corresponding PCA lower bound as a function of the spread factor S_f for the RBF model wafer profiles with $N_g = 10$ (solid) and $N_g = 100$ (dashed)

Fig. 7: [Industrial] Average Wafer Profile Reconstruction NMSE ($K = 100$ MC simulations) with a number of static and dynamic sampling approaches as a function of the number of measured sites. (The V_m range is plotted over two intervals for clearer presentation of the differences between algorithms at lower NMSE levels)

greater than or equal to 0.16, and; (2) the distance between a site and the edge of the unit disk was at least 0.07. The SDS algorithms were compared with the 'static' Random and 'static' FSCA approaches and with:

- 'Random Dynamic Sampling' (RDS) where wafer sites are ordered randomly then visited sequentially, V_m sites at a time, and;
- 'Conservative Dynamic Sampling' (CDS) where $V_m - 1$ sites are fixed according to FSCA with only the V_m-th site selected dynamically in a similar fashion to RDS.

It should be noted that the 'static' sampling approaches provide approximate upper and lower bounds on the performance achievable with dynamic sampling algorithms which seek to visit all sites over time. In particular, for statistically stationary processes FSCA defines an approximate lower bound on the achievable NMSE, since by design dynamic sampling trades of reconstruction performance to obtain better wafer coverage tem-

porally. It is not a precise lower bound since, as already noted, FSCA is not guaranteed to identify the optimal combination of V_m sites.

Figs 7, 8 and 9 show the average wafer profile reconstruction NMSE (over $K = 100$ MC simulations) as a function of the number of measured sites for each sampling strategy, for the Industrial, RBF and Zernike case studies, respectively. To illustrate the MC variability, the average and standard deviation of the reconstruction NMSEs is also reported in Table II for the Industrial case study, while Fig. 10 and 11 show the corresponding NMSE boxplots for each approach for $V_m = 4$ and $V_m = 7$, respectively.

It can be seen that static 'Random' sampling provides the worst NMSE performance of all methods while static FSCA achieves the best performance for all values of V_m, and therefore, serve as bounds on the NMSE performance achievable with Dynamic Sampling.

TABLE II: [Industrial] Average (standard deviation) wafer profile reconstruction NMSE for a number of static and dynamic sampling approaches ($K = 100$ MC simulations)

V_m	Random Static	FSCA Static	RDS Dynamic	CDS Dynamic	SDS-Corr Dynamic	SDS-NMSE Dynamic
2	48.87 (7.30)	36.54 (2.04)	48.68 (2.77)	43.79 (2.21)	46.64 (2.94)	46.64 (3.00)
3	26.80 (6.82)	18.35 (1.30)	27.43 (2.54)	21.35 (2.00)	24.05 (1.90)	23.58 (2.05)
4	11.61 (5.57)	3.51 (0.32)	11.76 (2.01)	8.40 (1.02)	8.19 (1.17)	8.29 (1.26)
5	6.66 (4.98)	2.45 (0.31)	6.26 (1.48)	2.81 (0.28)	4.00 (0.61)	3.93 (0.59)
6	3.60 (2.27)	1.64 (0.30)	3.73 (0.62)	1.93 (0.27)	2.35 (0.40)	2.48 (0.49)
7	2.53 (1.31)	0.96 (0.25)	2.64 (0.63)	1.29 (0.25)	1.43 (0.23)	1.44 (0.23)
8	1.65 (0.80)	0.70 (0.11)	1.79 (0.39)	0.80 (0.19)	1.01 (0.12)	1.06 (0.17)
9	1.25 (0.58)	0.51 (0.06)	1.35 (0.26)	0.57 (0.09)	0.74 (0.08)	0.67 (0.14)

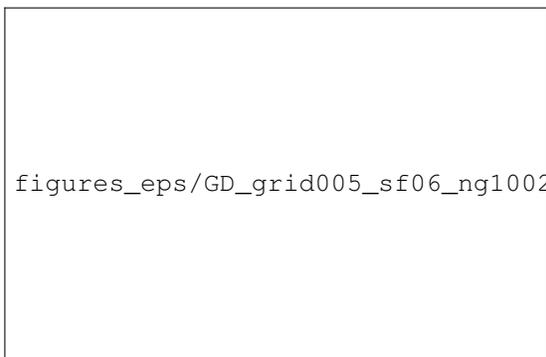
Fig. 8: [RBF] Average Wafer Profile Reconstruction NMSE ($K = 100$ MC simulations) with a number of static and dynamic sampling approaches as a function of the number of measured sites
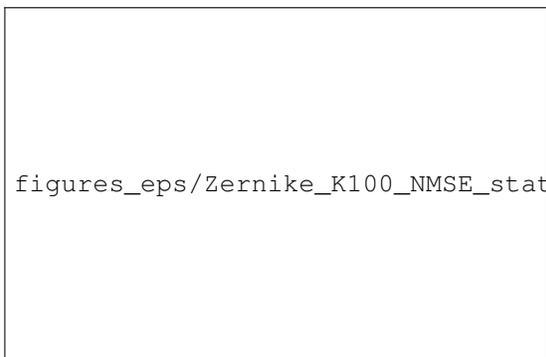

Fig. 9: [Zernike] Average Wafer Profile Reconstruction NMSE ($K = 100$ MC simulations) with a number of static and dynamic sampling approaches as a function of the number of measured sites
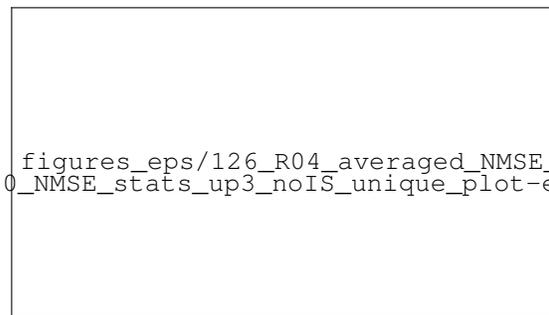

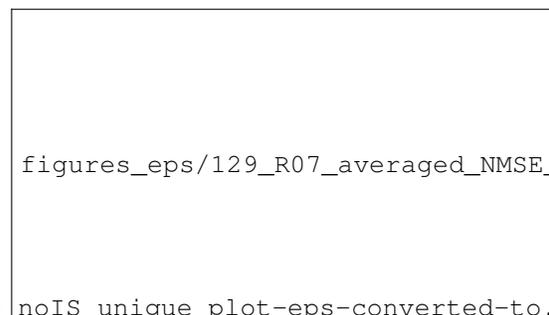Fig. 10: [Industrial] Boxplots of the NMSE with $V_m = 4$

Fig. 11: [Industrial] Boxplots of the NMSE with $V_m = 7$

There is little difference in average NMSE performance between random static sampling and RDS, but RDS exhibits substantially lower variability in performance. This is simply a consequence of the

fact that static random sampling is much more sensitive to the impact of selecting good or bad site combinations that RDS which averages this effect out over multiple wafers.

In theory, SDS-NMSE should be superior to SDS-Corr, since it is specifically designed to form clusters based on minimizing the impact on reconstruction accuracy (i.e. a supervised learning algorithm), whereas SDS-Corr simply selects cluster members based on correlation between sites with no consideration of the overall objective of minimizing the NMSE (i.e. it is an unsupervised method). However, comparing the performance of the two SDS implementations it can be seen that there is little to choose between them. Both implementations substantially outperform RDS and are only marginally inferior to CDS, the most conservative form of dynamically sampling. As expected, static FSCA yields the best NSME performance for a given number of measurement sites. However, SDS can in general achieve a similar level of accuracy by including one additional measurement per wafer. For example, with the optimum number of measurement sites ($V_m = 7$) the average NMSE reconstruction error with static FSCA is 0.96%. In contrast SDS-Corr yields a NMSE of 1.43%, but this drops to 1.01% when V_m is increased to eight.

The average WOI performances of the four dynamic sampling algorithms (RDS, CDS, SDS-Corr and SDS-NMSE) for the three case studies are plotted as a function of V_m in Figs 12-14. The mean and standard deviation of WOI for each algorithm for the Industrial case study are also reported in Table III. Note that WOI is independent of the data in the case of RDS and CDS and hence the WOI standard deviation is zero for these algorithms. By design RDS achieves optimal wafer coverage (WOI=100%)
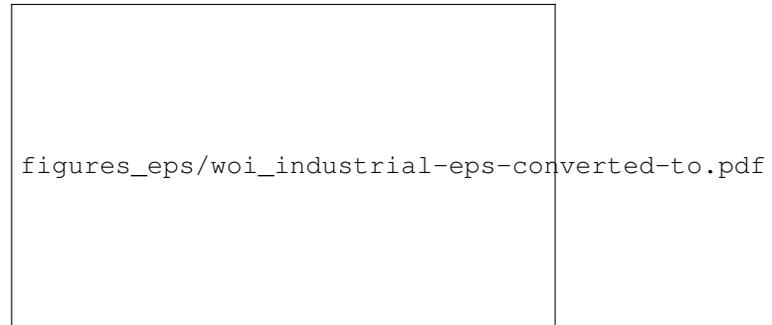

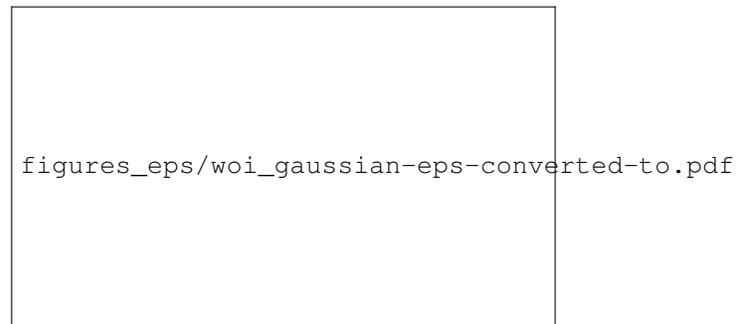Fig. 12: [Industrial] Average WOI ($K = 100$ MC simulations) with various dynamic sampling approaches

Fig. 13: [RBF] Average WOI ($K = 100$ MC simulations) with various dynamic sampling approaches

and hence is the upper bound on the performance that can be achieved by other methods. While CDS yielded the best NMSE performance it is the poorest algorithm in terms of wafer coverage with a WOI of less than 20% for $V_m \geq 5$. In contrast, the proposed SDS-Corr algorithm performs much better achieving WOI values of 70% or better for $V_m \leq 5$.

Comparing SDS-Corr and SDS-NMSE reveals that in general the former produces a higher mean WOI and is more stable (smaller WOI variance) than the latter. Fig. 15 shows typical cluster assignments obtained with Correlation and NMSE based clustering for $V_m = 2, 4, 6$ and 7 for the Industrial
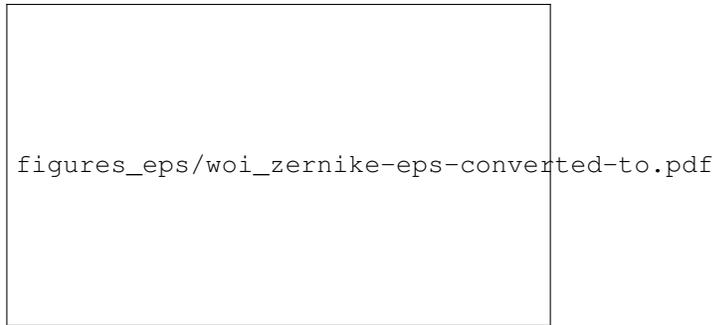
Fig. 14: [Zernike] Average WOI ($K = 100$ MC simulations) with various dynamic sampling approaches

TABLE III: [Industrial] Average (standard deviation) WOI for the proposed dynamic sampling approaches ($K = 100$ MC simulations)

V_m	RDS	CDS	SDS-Corr	SDS-NMSE
2	100 (-)	50.0 (-)	92.3 (4.7)	82.0 (5.7)
3	100 (-)	34.0 (-)	73.4 (3.6)	73.0 (7.8)
4	100 (-)	26.1 (-)	67.4 (8.3)	84.4 (8.1)
5	100 (-)	20.0 (-)	69.7 (9.7)	45.9 (10.42)
6	100 (-)	18.2 (-)	66.4 (10.3)	55.8 (18.5)
7	100 (-)	16.3 (-)	61.4 (13.1)	56.2 (20.46)
8	100 (-)	14.3 (-)	65.4 (10.2)	45.7 (14.5)
9	100 (-)	12.2 (-)	59.6 (9.2)	30.7 (16.5)

Case Study. As can be seen NMSE clusters are much less balanced than those obtained using correlation, hence the poorer WOI performance of SDS-NMSE.

Overall, considering both NMSE and WOI performance metrics, SDS-Corr provides the best all-round performance among the sampling techniques considered.

D. Detecting Previously Unseen Process Behaviour

To demonstrate the effectiveness of SDS over static FSCA sampling, as a final example, we investigate their performance for the industrial case study where, following training on normal wafers, a previously unseen localized anomaly is introduced into the test wafers. Recall that the Industrial
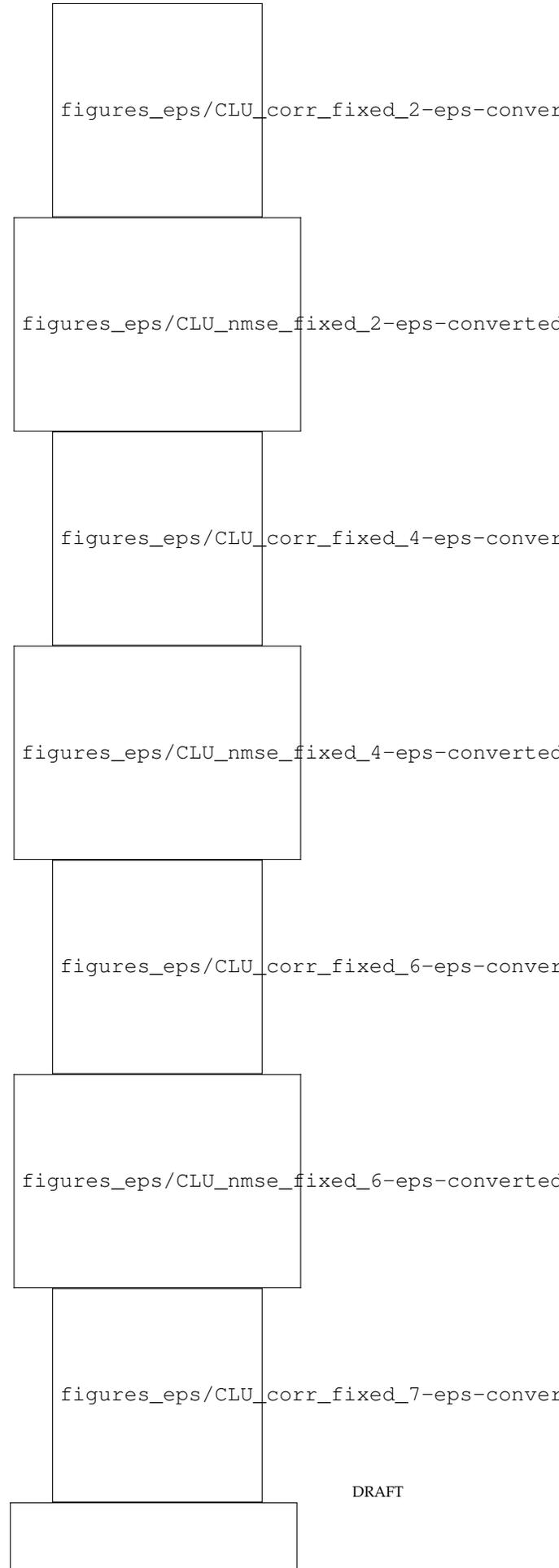
dataset is composed of $N = 316$ wafers, each measured at $V = 50$ locations. Using a seven-point measurement plan (i.e., $V_m = 7$), 200 randomly selected wafers are used as a historical dataset to train the static and dynamic sampling and wafer profile reconstruction methodologies described in the previous sections. The remaining 116 wafers are ordered randomly and modified to simulate the occurrence of a previously unseen, persistent, spatially localized, production anomaly. The anomaly is generated by adding a RBF perturbation at a fixed random location on each wafer surface. The location is selected at random among the V candidate measurement sites and is the same location on all test wafers. The height $h^{(n)}$ and spread $S_f^{(n)}$ of the perturbation for the n-th wafer are drawn from the following normal distributions:

$$h^{(n)} \approx N(0.8, 0.2),$$

$$S_f^{(n)} \approx N(0.4, 0.1).$$

Starting from the first wafer in the test dataset, and proceeding sequentially through the wafers, the following anomaly detection procedure is run:

1) Using linear models derived from the training dataset, each of the V_m sites measured for the current wafer is estimated from the measured values of the other $V_m - 1$ sites;

2) If the mismatch between the real and estimated measurement is too high, as defined by the 95% confidence intervals of the linear prediction model, an anomaly flag is raised.

3) If an anomaly is detected, signifying that previously unseen process behaviour has been detected, wafer processing stops, and the wafer count until anomaly detection recorded; otherwise, processing proceeds to the next test wafer.

TABLE IV: [Industrial] Anomaly detection rate (ρ_D), and Median (N_D) and Mean (μ_D) wafer count to anomaly detection, with various static and dynamic sampling approaches

Sampling Method	Type	ρ_D	N_D	μ_D
Random	Static	29.7	1	1.12
FSCA	Static	43.8	1	1.14
RDS	Dynamic	100	7	7.32
CDS	Dynamic	99.8	24	24.41
SDS-Corr	Dynamic	100	8	8.76
SDS-NMSE	Dynamic	100	10	11.09

To generate statistically robust results, the above procedure was repeated $K = 1000$ times, that is, the experiment is repeated for 1000 different instances of the test data set, where each instance has a different random sequence of wafers and a different random anomaly location. The results obtained are reported in Table IV. This shows the anomaly detection rate (ρ_D), defined as the percentage of the 1000 repetitions in which the anomaly was detected, and the median (N_D) and mean (μ_D) wafer counts to anomaly detection, for each sampling method.

As expected, the Random and FSCA static sampling methods perform poorly with detection rates of 29.7% and 43.8%, respectively. In contrast, the dynamic sampling strategies, achieves a 100% detection rate, with the exception of CDS which achieves a 99.8% detection rate. Recalling that RDS has the lowest possible MSSI ($= 8$), it is not surprising that is has the lowest N_D and μ_D values of the dynamic sampling methods. However, SDS-Corr is a close second with $\mu_D = 8.76$ versus 7.32 for RDS. It is interesting to note that SDS-Corr is superior to SDS-NMSE, which in turn is substantially superior to CDS. This is consistent with the WOI performances of each method, as

discussed in the previous section and plotted in Figure 12.

VI. DISCUSSION AND CONCLUSIONS

A novel methodology and algorithms for dynamic spatial sampling and reconstruction of wafer profiles have been presented. The key elements of the methodology are FSCA based site selection to eliminate measurement redundancy, WMR estimation of unmeasured sites to enable accurate profile reconstruction, and sequential dynamic sampling (SDS) of sites from FSCA clusters (formed by clustering unmeasured wafer sites around the FSCA selected sites) to mitigate the risk of missing previously unseen process behaviour. By design the proposed methodology can detect localized anomalies that persist over several runs (greater than the MSSI) and anomalies whose extent is such that they impact on neighbouring sites that are being measured. Significantly, detection of the former class of anomaly is not guaranteed by traditional static sampling plans.

Two variations of SDS were considered, one employing correlation based clustering and the other NMSE based clustering. Results from both practical and simulated case studies have shown that while both approaches provide similar wafer reconstruction accuracy, the former is preferred by virtue of its superior wafer coverage and substantially lower algorithm computational complexity. The effectiveness of SDS in providing good wafer coverage (as measured by the WOI metric) with minimal degradation in wafer reconstruction accuracy has also been demonstrated through comparisons with a number of static and dynamic sampling alternatives including static FSCA, the optimal NMSE wafer reconstruction benchmark, and RDS, the optimal wafer coverage benchmark.

The fundamental requirement when applying the FSCA based methodology is that the training dataset $\mathbf{X}$ used to optimise the FSCA sites and clusters, and estimate the WPR models, must be representative of the spatial correlation (exhibited across the wafer surface) and the production variability over time that will be encountered going forward. The industrial case study presented in the paper was for wafers corresponding to several different products/recipes processed through a single chamber. The resulting variability in wafer profiles (see Fig.1) did not present an issue for WPR and DS performance. If wafers come from more than one chamber there is no guarantee that this will be the case if the dataset does not contain data from each chamber. For optimum performance, training of bespoke models for each chamber is likely to be needed if there is significant chamber mismatch.

A number of avenues exist for future research. In SDS sites are sequentially selected in random order within each cluster independently of the selections made in other clusters. Therefore, the possibility exists of sites in close proximity, but residing in different clusters, being selected simultaneously, potentially leading to reduced wafer profile reconstruction performance. While our experimental studies have not shown this to be a major concern in terms of performance, and interesting question is how to coordinate the selection of sites across clusters to maximize reconstruction accuracy. A further, challenging task is to develop an approach to updating the WMR models online with the incomplete data that arises with dynamic sampling. Finally, it would be of interest to evaluate to what extent wafer profile reconstruction accuracy can

be enhanced by incorporating production/process variables into the proposed WMR approach.

REFERENCES

[1] J. Blue and A. Chen, "Spatial variance spectrum analysis and its application to unsupervised detection of systematic wafer spatial variations," *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 1, pp. 56–66, 2011.

[2] R. Borgoni, L. Radaelli, V. Tritto, and D. Zappa, "Optimal reduction of a spatial monitoring grid: Proposals and applications in process control," *Computational Statistics and Data Analysis*, vol. 58, pp. 407 – 419, 2013.

[3] C.-H. Chang, M. Akilian, and M. Schattenburg, "Describing isotropic and anisotropic out-of-plane deformations in thin cubic materials by use of zernike polynomials," *Applied Optics*, vol. 45, no. 3, pp. 432–437, 2006.

[4] F.-T. Cheng, H.-C. Huang, and C.-A. Kao, "Developing an automatic virtual metrology system," *IEEE Transactions on Automation Science and Engineering*, vol. 9, no. 1, pp. 181–188, 2012.

[5] N. Draper and H. Smith, *Applied Regression Analysis*. Wiley, 1966.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[7] R. Jin, C.-J. Chang, and J. Shi, "Sequential measurement strategy for wafer geometric profile estimation," *IIE Transactions*, vol. 44, no. 1, pp. 1–12, 2012.

[8] I. Jolliffe, *Principal Component Analysis*. John Wiley & Sons, 2005.

[9] D. Kurz, C. De Luca, and J. Pilz, "A sampling decision system for virtual metrology in semiconductor manufacturing," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 75–83, 2015.

[10] C.-Y. Lee and B.-S. Chen, "Mutually-exclusive-and-collectively-exhaustive feature selection scheme," *Applied Soft Computing*, 2017.

[11] C. Martin, "Towards a new scale," *Nature Nanotechnology*, vol. 11, p. doi:10.1038/nnano.2016.8, 2016.

[12] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. pp. 239–245, 1979.

[13] T.-H. Pan, B.-Q. Sheng, D. S.-H. Wong, and S.-S. Jang, "A virtual metrology system for predicting end-of-line electrical properties using a mancova model with tools cluster-ing," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 2, pp. 187–195, 2011.

[14] R. Picard and R. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, pp. 575–583, 1984.

[15] P. Prakash, A. Johnston, B. Honari, and S. McLoone, "Optimal wafer site selection using forward selection component analysis," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2012, pp. 91–96.

[16] L. Puggini and S. McLoone, "Forward selection component analysis: Algorithms and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2395–2408, 2017.

[17] E. Ragnoli, S. McLoone, S. Lynn, J. Ringwood, and N. Macgearailt, "Identifying key process characteristics and predicting etch rate from high-dimension datasets," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2009, pp. 106–111.

[18] J. Ringwood, S. Lynn, G. Bacelli, B. Ma, E. Ragnoli, and S. McLoone, "Estimation and control in semiconductor etch: Practice and possibilities," *IEEE Transactions on Semiconductor Manufacturing*, vol. 23, no. 1, pp. 87–98, 2010.

[19] D. Sandwell, "Biharmonic spline interpolation of geos-3 and seasat altimeter data," *Geophysical Research Letters*, vol. 14, pp. 139–142, 1987.

[20] J. Shao, "Linear model selection by cross-validation." *Journal of the American Statistical Association*, vol. 88, pp. 486–494, 1993.

[21] F. Souza and R. Araujo, "Online mixture of univariate linear regression models for adaptive soft sensors," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 937–945, 2014.

[22] G. A. Susto and A. Beghi, "Least angle regression for semiconductor manufacturing modeling," in *IEEE International Conference on Control Applications (CCA)*. IEEE, 2012, pp. 658–663.

[23] ——, "A virtual metrology system based on least angle regression and statistical clustering," *Applied Stochastic Models in Business and Industry*, vol. 29, pp. 362–376, 2013.

[24] G. A. Susto, S. Pampuri, A. Schirru, A. Beghi, and G. De Nicolao, "Multi-step virtual metrology for semiconductor manufacturing: A multilevel and regularization methods-based approach," *Computers & Operations Research*, vol. 53, pp. 328–337, 2015.

[25] T. Vincent, J. Stirton, and K. Poolla, "Metrology sampling strategies for process monitoring applications," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, pp. 489–498, 2011.

[26] W.-M. Wu, F.-T. Cheng, T.-H. Lin, D.-L. Zeng, and J.-F. Chen, "Selection schemes of dual virtual-metrology outputs for enhancing prediction accuracy," *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 2, pp. 311–318, 2011.

[27] C. Yugma, J. Blue, S. Dauzère-Pérès, and P. Vialletelle, "Integration of scheduling and advanced process control in semiconductor manufacturing: review and outlook," in *IEEE International Conference on Automation Science and Engineering (CASE)*.   IEEE, 2014, pp. 93–98.

[28] D. Zeng, "Statistical methods for enhanced metrology in semiconductor/photovoltaic manufacturing," Ph.D. dissertation, Electrical Engineering and Computer Sciences Department, University of California, Berkeley, Dec 2012. [Online]. Available: www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-237.html

[29] L. Zhang, K. Wang, and N. Chen, "Monitoring wafers' geometric quality using an additive gaussian process model," *IIE Transactions*, vol. 47, pp. 1–15, 2016.