Process Monitoring through Wafer-level Spatial Variation Decomposition

Ke Huang*, Nathan Kupp†, John M. Carulli Jr‡, and Yiorgos Makris*
*Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080
†Department of Electrical Engineering, Yale University, New Haven, CT 06511
‡Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, TX 75243

***Abstract*—Monitoring the semiconductor manufacturing process and understanding the various sources of variation and their repercussions is a crucial capability. Indeed, identifying the root-cause of device failures, enhancing yield of future production through improvement of the manufacturing environment, and providing feedback to the designer toward development of design techniques that minimize failure rate rely on such a capability. To this end, we introduce a spatial decomposition method for breaking down the variation of a wafer to its spatial constituents, based on a small number of measurements sampled across the wafer. We demonstrate that by leveraging domain-specific knowledge and by using as constituents dynamically learned, interpretable basis functions, the ability of the proposed method to accurately identify the sources of variation is drastically improved, as compared to existing approaches. We then illustrate the utility of the proposed spatial variation decomposition method in (i) identifying the main contributor to yield variation, (ii) predicting the actual yield of a wafer, and (iii) clustering wafers for production planning and abnormal wafer identification purposes. Results are reported on industrial data from high-volume manufacturing, confirming the ability of the proposed method to provide great insight regarding the sources of variation in the semiconductor manufacturing process.**

I. Introduction

As complexity of modern Integrated Circuits (ICs) increases and minimum feature sizes continue to shrink, uncontrollable process variations constitute a mounting challenge in semiconductor manufacturing. Variability is introduced by various sources during manufacturing and each step, such as lithography, ion implantation, thermal treatments, etc., can be considered as a source of variation. For example, rotation of wafers to increase process uniformity can result in radial spatial variation, thermal gradients can result in linear or polynomial spatial variation, and reticle size can result in discontinuous effects in wafer-level measurements. With excessive process variations being a major contributor to yield loss during IC manufacturing [1], monitoring and understanding such variations is crucial for identifying the root-cause of device failures, enhancing yield for future device production, and providing valuable feedback to the designers.

A key step toward this end is the identification of the various sources that contribute to process variations and their repercussions. Prior literature commonly models the impact of process variations on wafer-level measurements as the sum of a systematic spatial component and a random component [2]. While random variation may be relatively easy to monitor by analysis of variance (ANOVA) methods [3], [4], systematic variation is much more intricate to model and deal with.

In this work, we propose a novel approach for identifying and analyzing systematic process variation through wafer-level spatial decomposition. More specifically, we employ a spatial decomposition method for breaking down the systematic variation of a wafer to a set of weighted basis functions, based on a small number of measurements from sampled die locations across the wafer. Figure 1 depicts this concept through an example of wafer-level spatial variation decomposition. In this example, the total variation on the wafer is decomposed into three distinct basis functions with the corresponding weight vector $A = [a_1, a_2, a_3]$. The main challenge in this effort is to identify an appropriate set of basis functions which can not only accurately reflect the spatial variation but which also have interpretable meaning which can assist the process engineer in understanding and moderating the source of variation. Accordingly, a key novelty of the method proposed herein over prior efforts is that, instead of employing a fixed set of statically-defined basis functions, it uses domain-specific knowledge to dynamically learn the most appropriate basis functions from the data. Thereby, its ability to pinpoint sources of variation and to provide actionable information to the process engineer is drastically improved.

As we demonstrate herein, the set of identified basis functions along with the vector of coefficients can be used to:

- identify the most prominent spatial variation component and the main contributor to yield variation by analyzing the correlation between the estimated weight vector and the yield,
- predict yield using correlation functions which map the estimated weight vector to the actual yield, and
- classify wafers through clustering analysis in order to detect abnormal wafers and plan future production.

The groundwork for applying wafer-level spatial variation decomposition was laid in [5], wherein the authors used a set of predefined basis functions. Herein, we extend the key ideas of [5] by introducing domain-specific knowledge in learning the basis functions and we demonstrate that, thereby, the capability of identifying sources of variation is greatly improved. Recent work on variation decomposition is also described in [6], wherein random process variation is removed and a single pattern of systematic spatial variation is exposed. In contrast, the method proposed herein delves into further decomposing the systematic process variation into domain-specific, interpretable basis functions.

978-1-4799-0859-2/13/$31.00 ©2013 IEEE

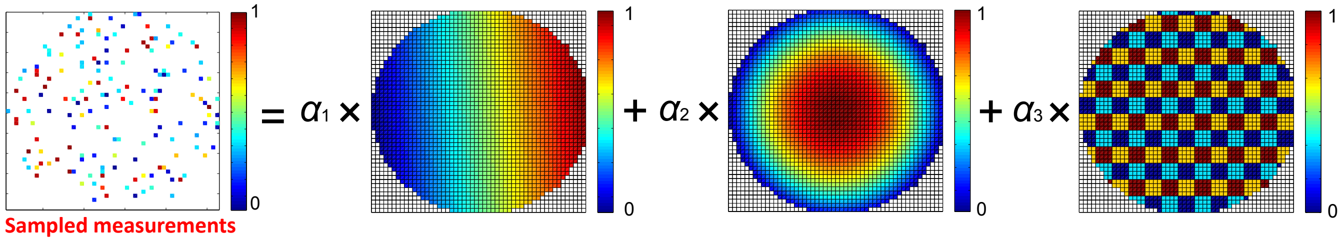

Fig. 1. An example of wafer-level spatial decomposition, where the total systematic variation on the wafer is decomposed into three distinct basis functions with the corresponding weight vector $A = [a_1, a_2, a_3]$

The remainder of this paper is organized as follows. In Section II, we discuss in detail prior work on wafer-level spatial correlation modeling and variation decomposition, using statistical analysis. In Section III, we introduce the proposed approach for understanding the various sources of wafer-level process variation and decomposing it into its spatial constituents. Experimental results which demonstrate the effectiveness of the proposed method using industrial data from high-volume manufacturing are provided in Section IV, and conclusions are drawn in Section V.

II. Prior work

A. Spatial correlation modeling of wafer-level measurements

Recent research on modeling spatial measurement correlation has shown great promise in capturing wafer-level spatial variation and, thereby, reducing test cost [4], [7]–[11]. The underlying idea is to collect measurements for a sparse subset of die on each wafer and subsequently train statistical spatial models to predict performance outcomes at unobserved die locations. In [4], the expectation-maximization (EM) algorithm is used to estimate spatial wafer measurements, assuming that data comes from a multivariate normal distribution. The Box-Cox transformation is used in case data is not normally distributed. The “Virtual Probe” (VP) approach [7]–[9] models spatial variation via a Discrete Cosine Transform (DCT) that projects spatial statistics into the frequency domain. Similarly, the author of [10] builds spatial models based on Generalized Least Square fitting and a structured correlation function. As recently shown in [11], [12], using Gaussian Process (GP) models can dramatically improve both prediction accuracy and computational time, as compared to the VP approach.

The utility of spatial interpolation models of wafer-level measurements has been demonstrated in various contexts. In [12], the authors extrapolate scribe line e-test measurements using spatial correlation models based on GP. In [11], the authors use GP to build spatial correlation models which dramatically reduce test time for probe-test specification measurements of RF devices. Handling discontinuous process variation effects in building spatial correlation models is also discussed in [13], wherein the authors employ a k-means clustering algorithm to ensure high prediction accuracy for measurements exhibiting spatially discontinuous effects.

B. Wafer-level spatial variation decomposition

The aforementioned work on spatial correlation modeling of wafer-level measurements shows the capability of extracting principal spatial variation patterns based on a sparse subset of die samples. Once these patterns are identified, they can be further analyzed to monitor process variation. The traditional analysis of variance (ANOVA) method provides an efficient way of quantifying the contribution of within-die, die-to-die, wafer-to-wafer or lot-to-lot variation to the total variation of a wafer [3], [4]. However, it cannot distinguish between wafers exhibiting the same total variation when the spatial distribution of this variation differs. To this end, a wafer-level variation decomposition method has been proposed, which takes into account the contribution of spatial variation patterns to the total variation. As mentioned earlier, the impact of process variations on wafer-level measurements can be modeled as the sum of a systematic spatial component and a random component [2]:

$$m(x, y) = g(x, y) + \epsilon \tag{1}$$

where $m(x, y)$ is the measurement under consideration, expressed as a function of a wafer’s Cartesian coordinate (x, y), $g(x, y)$ is the systematic spatial variation component, and ϵ is the random component often modeled as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Notice that a constant term C can also be added to (1) to represent wafer-to-wafer and lot-to-lot offset.

In [14], spatial variation of wafer-level measurements is decomposed into within field, field-to-field, across wafer, and random variation. A statistical filter is used to select various types of variation in the frequency domain using discrete Fourier transform. In [15], spatial variation is modeled by a combination of interpolation and regression models. The Nearest Neighbor Residual approach proposed in [16] aims at reducing the variance of spatially distributed wafer-level measurements to improve the detection of outliers. In [17], wafer-level spatial decomposition is accomplished by subtracting the effect of random variation from the mean value within a reticle. A dynamically learned linear plane spatial function is proposed to capture gradient effect caused by bake plate thermal gradients over different lots. Yet many other functions need to be incorporated in the overall analysis. All the above-mentioned approaches analyze wafer-level variation by taking all available measurements on the wafer, which could lead to high computational cost. In [6], a sparse subset of die samples are used to decompose systematic and random variations by projecting data into the frequency domain using a discrete cosine transform. The method described therein aims at identifying a spatial pattern which carries a unique signature in the frequency domain via sparse regression. In [5], spatial variation across a wafer is modeled by a linear combination of distinct basis functions representing different sources of variation:

$$m(x, y) = \sum_{i=1}^{n_b} \alpha_i b_i(x, y) + \epsilon \tag{2}$$

where $b_i(x, y)$ denotes the i-th spatial basis function, α_i denotes the coefficient of the i-th basis function, and n_b denotes the number of considered basis functions. Coefficients α_i are estimated using a linear regression method [18] and the null/alternative hypothesis method is used to determine the existence of a spatial pattern on a wafer. The residual of the model is used to represent random variation. As discussed earlier in the introduction, the choice of basis functions is crucial to the success of this method. Authors in [5] chose a set of predefined basis functions, while the work presented herein used domain-specific knowledge to dynamically learn the most appropriate basis functions from the data prior to performing spatial variation decomposition.

III. Proposed approach

The proposed approach consists of three principal phases:

- **Pre-decomposition learning,** during which appropriate basis functions are learned from a hold-out set of wafers.
- **Decomposition,** during which a target measurement is sampled on a small percentage of die-locations across each wafer under consideration and appropriate weights are attributed to each basis function, through a process similar to the one described in [5].
- **Post-decomposition analysis,** during which the correlation between the various sources of variation (as reflected in the basis functions) and yield is statistically learned, in order to identify the main contributors to yield variation. The estimated coefficients of the basis functions can also be used to predict the actual yield of a wafer, as well as to perform wafer clustering analysis.

Details for each of the three phases are provided next.

A. Pre-decomposition learning

Variability is introduced by several different sources during semiconductor manufacturing. While each piece of equipment, each knob, and each step in the process can be considered as a distinct source of variation, in practice the effects of variability can be cumulatively reflected through a relatively small set of basis functions. Such basis functions constitute a mechanism for communicating to process engineers what is being observed at probe in way that has interpretable meaning and be acted upon. Let $b_i(x, y)$ denote the i-th considered basis function as specified in (2). Examples of interpretable $b_i(x, y)$ include:

1) Linear basis function: This type of basis function represents linear spatial variation of wafer-level measurements, caused, for example, by thermal gradients. It can be expressed as

$$b_i(x, y) = ax + by \tag{3}$$

where a and b are used-defined coefficients which can be learned using a hold-out set of wafers. The basis function with coefficient α_1 shown in Figure 1 is an example of a linear basis function.

2) Cosine basis function: This type of basis function represents radial spatial variation of wafer-level measurements, caused, for example, by wafer spinning. It can be expressed as [5]

$$b_i(x, y) = \cos(n \frac{2\pi}{d_u} r) \tag{4}$$

where d_u is the usable wafer diameter, r is the distance from the center of the wafer, and n is a user-defined parameter which can also be learned using a hold-out set of wafers. The basis function with coefficient α_2 shown in Figure 1 is an example of a cosine basis function.

3) Discontinuous basis function: This type of basis function represents discontinuous spatial variation of wafer-level measurements, which can be caused by a number of reasons. For example, the reticle shot that produces several die patterns at the same time in the lithography process may result in individual rectangular regions. Similarly, a multi-site testing strategy may lead to systematic variations for die that are tested at the same time. If k denotes the number of "levels" caused by a discontinuous effect, then a discontinuous basis function can be expressed as

$$b_i(x, y) = \frac{l}{k} m_k \tag{5}$$

where l denotes the discontinuous "level" that the die on wafer coordinate (x, y) belongs to, and m_k denotes the measurement value of the highest "level". The basis function with coefficient α_3 in Figure 1 is an example of discontinuous basis functions.

A key contribution of this work is the use of domain-specific knowledge to learn basis functions. For example, to accurately learn the function in (5), we need to determine which "level" a die in coordinate (x, y) belongs to. Our experience with production test data shows that, for measurements which exhibit discontinuous effects, the spatial components are often relatively stationary across wafers (though the actual variation is certainly not). In other words, for a given measurement, most wafers exhibit very similar spatial discontinuous patterns. Based on this observation, we propose to learn the function in (5) by a k-means clustering algorithm using a single wafer (or a small set), on which all measurements for all die locations are explicitly collected. Formally, let the set

$$M^i = \{m_1, m_2, \ldots, m_n\} \tag{6}$$

include the values of the i-th measurement on all die of a wafer, with m_j denoting the measurement on the j-th die and n denoting the total number of die which are to be clustered. The k-means clustering algorithm aims to partition M^i into k sets ($k \leq n$): $\{S_1, S_2, \ldots, S_k\}$ so as to minimize the expected *distortion D*, which is defined as the sum of squared distances between each observation and its dominating cluster mean:

$$D = \sum_{j} \|\bar{m}_{k(j)} - m_j\|^2 \tag{7}$$

where $\bar{m}_{k(j)}$ denotes the nearest cluster mean value for observation m_j. In this work, we use the most common iterative refinement technique to refine the choices of cluster means in order to reduce the distortion D. The technique involves the following steps shown in Algorithm 1 [19].

```
1. Set $k$ cluster means $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_k\}$ to random values.
2. Assign each measurement in $M^i$ to the cluster with the nearest cluster mean. The assigned $p$-th cluster is denoted by $S_p$:
       $S_p = \{m_j : \|m_j - \bar{m}_p\|^2 \leq \|m_j - \bar{m}_q\|^2, \forall 1 \leq q \leq k\}$   (8)
3. Compute the new cluster means.
       $\bar{m}_p = \frac{1}{n_p} \sum_{m_j \in S_p} m_j$   (9)
   where $n_p$ is the number of observations in the $p$-th cluster.
4. Repeat steps 2 & 3 until the assignments do not change.
```

Algorithm 1: k-means algorithm for partitioning a wafer to clusters caused by discontinuous effects

The k-means clustering algorithm is a simple, unsupervised learning approach which allows us to separate the die on a wafer into k different clusters caused by various discontinuous effects, without assuming the shape of clusters. Notice that the clusters cannot be obtained by simply examining test site information or other reverse-engineering method, mainly because cluster shapes are often formed by multiple sources of variation, including discontinuous, radial or linear variations.

The question that naturally arises next concerns the choice of k. This choice is crucial in the k-means clustering algorithm. Underestimating k would result in clusters that still contain discontinuous patterns, while overestimating k would result in basis functions not reflecting the real underlying spatial pattern. The authors of [20] conducted a very comprehensive comparative study of 30 methods for determining the number of clusters in data. Among the variety of examined methods, the approach suggested in [21] generally outperformed the others. This approach consists of choosing an optimal value for k by maximizing the between-cluster dispersion and minimizing the within-cluster dispersion. Formally, the optimal value for k is defined as [21]

$$k = \underset{g}{\operatorname{argmax}}\, CH(g) \tag{10}$$

where $CH(g)$ is the Calinski and Harabasz index when g clusters are considered and is defined as

$$CH(g) = \frac{B(g)(g-1)}{W(g)(n-g)} \tag{11}$$

where n is the total number of die on the wafer, and $B(g)$ and $W(g)$ are the between- and within-cluster sums of squared errors computed as

$$B(g) = \sum_{p=1}^{g} n_p (\bar{m}_p - \bar{m})(\bar{m}_p - \bar{m})^T \tag{12}$$

$$W(g) = \sum_{p=1}^{g} (\sum_{m_j \in S_p} (m_j - \bar{m}_p)(m_j - \bar{m}_p)^T) \tag{13}$$

where n_p denotes the number of samples in the p-th cluster, $\bar{m}_p$ denotes the cluster mean of the p-th cluster, and $\bar{m}$ denotes the mean of all measurement samples in M^i.

Equation (10) allows us to automatically choose an optimal value for k for a particular measurement without making any assumptions about its discontinuity trends.

B. Decomposition

Once all the basis functions are specified, we can readily use them to identify different sources of variation in manufacturing. In particular, we analyze each wafer under consideration by taking wafer-level measurements from a sample of die on the wafer and using them to compute a weight for each basis function. For this purpose, we use robust regression [22] to estimate the weight of each basis function, which allows us to minimize the influence of outliers in the estimation. Formally, let $\mathbf{b}_j$ denote the basis function vector on the j-th Cartesian coordinate (x_j, y_j) of a particular wafer: $\mathbf{b}_j = [b_0, b_1(x_j, y_j), \ldots, b_{n_b}(x_j, y_j)]$, where $b_i(x_j, y_j)$ denotes the value of i-th basis function, n_b is the number of considered basis functions and b_0 is a constant term, and let $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_{n_b}]$ denote the corresponding weight coefficient associated with each basis function. Let m_j denote the considered measurement value on the j-th coordinate. Then m_j can be expressed as

$$m_j = \boldsymbol{\alpha} \mathbf{b}_j^\top + r_j \tag{14}$$

where r_j is the residual of the estimation on the j-th coordinate. In robust regression, we estimate $\boldsymbol{\alpha}$ by minimizing the objective function:

$$\sum_{i=1}^{n_m} \rho(r_i) = \sum_{i=1}^{n_m} \rho(m_i - \boldsymbol{\alpha} \mathbf{b}_j^\top) \tag{15}$$

where $\rho(\cdot)$ is a function which gives the contribution of each residual to the objective function (for example, for least-square estimation, $\rho(r_i) = r_i^2$), and n_m is the number of die locations used to estimate $\boldsymbol{\alpha}$. We minimize the function in (15) w.r.t. $\boldsymbol{\alpha}$ by taking derivatives,

$$\sum_{i=1}^{n_m} \psi(m_i - \boldsymbol{\alpha} \mathbf{b}_j^\top) \mathbf{b}_j^\top = 0 \tag{16}$$

where $\psi = \rho'$. If we define the weight function $w(r) = \psi(r)/r$, then the estimating equation (16) may be written as

$$\sum_{i=1}^{n_m} w(r_i)(m_i - \boldsymbol{\alpha} \mathbf{b}_j^\top) \mathbf{b}_j^\top = 0 \tag{17}$$

In this work, we use Huber's weight function which has the form:

$$w(r) = \begin{cases} 1 & \text{if } |r| \le k \\ k_r/|r| & \text{if } |r| > k \end{cases} \tag{18}$$

where k_r is a user-defined tuning constant specifying the boundary of "bad" observations. Also, to solve (17), we use the iteratively reweighted least squares method shown in Algorithm 2.

```
1. Set initial estimates $\boldsymbol{\alpha}^{(0)}$ using least-square estimates.
2. At each iteration $t$, calculate residual $r_i^{(t)}$ and
   associated weights $w_i^{(t-1)} = w\left(r_i^{(t-1)}\right)$.
3. Solve for new weighted-least-squares estimates
   $\boldsymbol{\alpha}^{(t)} = \left[\mathbf{B}'\mathbf{W}^{(t-1)}\mathbf{B}\right]^{-1}\mathbf{B}'\mathbf{W}^{(t-1)}\mathbf{m}$    (19)
   where $\mathbf{B}$ is the model matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{n_m}]^\top$,
   $\mathbf{W}^{(t-1)} = \text{diag}\{w_i^{(t-1)}\}$, and $\mathbf{m} = [m_1, \ldots, m_{n_m}]^\top$.
4. Repeat steps 2 and 3 until the estimated coefficients
   converge.
```

Algorithm 2: Iteratively reweighted least squares method

Equation (19) allows us to estimate the coefficient of each basis function, based on measurements taken on a subset of die locations (n_m samples) of a particular wafer.

C. Post-decomposition analysis

1) Identifying main contributor to yield variation: Once the weight vector, $\hat{\boldsymbol{\alpha}}$, is estimated, we can use it to identify the most "important" spatial variation component and the main contributor to yield variation by building the correlation functions that map $\hat{\boldsymbol{\alpha}}$ to actual yield. Let y_{kl} denote the yield of the k-th measurement on the l-th wafer, and let $\hat{\boldsymbol{\alpha}}_{kl}$ denote the coefficient vector of the basis functions, estimated on the l-th wafer for the k-th measurement. Then y_{kl} is expressed as

$$\begin{aligned} y_{kl} &= f(\hat{\boldsymbol{\alpha}}_{kl}) + \epsilon \\ &= f(\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_{n_b}) + \epsilon \end{aligned} \tag{20}$$

where ϵ denotes additive stochastic error, whose expected value is defined to be zero, and f denotes a function mapping $\hat{\boldsymbol{\alpha}}_{kl}$ to y_{kl}. By effectively learning f using a set of training samples, the impact of different sources of variation on the actual yield can be learned. In this work, we use the Multivariate Adaptive Regression Splines (MARS) regression method [23] to estimate f, where y_{kl} is expressed as a weighted sum of individual functions

$$\begin{aligned} y_{kl} &= \sum f_i(\hat{\alpha}_i) \\ &\quad + \sum f_{i,j}(\hat{\alpha}_i, \hat{\alpha}_j) \\ &\quad + \sum f_{i,j,h}(\hat{\alpha}_i, \hat{\alpha}_j, \hat{\alpha}_h) \end{aligned} \tag{21}$$

The first term in (21) denotes the sum of all individual functions involving only $\hat{\alpha}_i$, the second term denotes the sum of all individual functions involving only $\hat{\alpha}_i$ and $\hat{\alpha}_j$, and so on. Note that i, j and h vary from 1 to n_b. As suggested in [23], the "importance" of each input variable $\hat{\alpha}_i$ can be assessed by computing the standard deviation of $f_i(\hat{\alpha}_i)$, computed across all considered wafers:

$$\sigma\big(f_i(\hat{\alpha}_i)\big) = \sqrt{\frac{1}{N_w - 1}\sum_{l=1}^{N_w}\big(f_i^l(\hat{\alpha}_i) - \bar{f}_i(\hat{\alpha}_i)\big)} \tag{22}$$

where $f_i^l(\hat{\alpha}_i)$ denotes the value of $f_i(\hat{\alpha}_i)$ estimated on the l-th wafer, N_w denotes the number of considered wafers in the data set, and $\bar{f}_i(\hat{\alpha}_i)$ denotes the sample mean of $f_i(\hat{\alpha}_i)$ computed across N_w wafers. In this work, we omit the second and higher order interaction analysis for brevity. Note that the "importance" of these interaction terms can be computed similarly as in (22). The greater the $\sigma\big(f_i(\hat{\alpha}_i)\big)$, the more important the $\hat{\alpha}_i$ in the model. We then define the principal yield contributor α_p as

$$\alpha_p = \operatorname*{argmax}_j \, \sigma\big(f_j(\hat{\alpha}_j)\big) \tag{23}$$

The above equation allows us to identify the principal contributor to yield variation, which can be further explored by process engineers to improve process and enhance yield.

2) Yield prediction: Once the correlation function f that maps $\hat{\boldsymbol{\alpha}}$ to y is learned, we can use it to accurately predict yield for new wafers, using Equation (20). Note that predicting yield can also be accomplished by employing spatial correlation models to predict measurements at untested die locations, as discussed in Section II-A. In this work, we show that by using the estimated $\hat{\boldsymbol{\alpha}}$ and the correlation function f, the yield can be accurately predicted without explicitly estimating measurements of untested die.

3) Wafer clustering analysis: In order to understand the impact of sources of variation on the wafers over time and different production lots/sites, the estimated $\hat{\boldsymbol{\alpha}}$ can also be used as a signature to classify wafers into different bins. A $k-$clustering approach similar to Algorithm 1 in Section III-A3 can be used on a training set of wafers to identify the common types of produced wafers and define a number of clusters/bins. Then, for each wafer coming out of production, the vector $\hat{\boldsymbol{\alpha}}$ is computed and classified into the appropriate cluster. By monitoring the distribution of wafers into bins, we can identify abnormal wafers and process excursions, as well as obtain useful information to assist planning of future production runs.

IV. Experimental results

We now demonstrate results of applying the proposed method on semiconductor data from high-volume manufacturing. The device under consideration is an RF transceiver with multiple radios built in a 65nm technology. Our dataset contains a total of 690 wafers, each of which has approximately 2,000 devices, with 78 probe test measurements collected on each device. For each wafer, 10% randomly chosen die locations are used to estimate $\hat{\boldsymbol{\alpha}}$. In this case study, we employ 4 basis functions, namely linear, cosine, discontinuous #1, and discontinuous #2, thus $n_b = 4$. Figure 2 shows the normalized

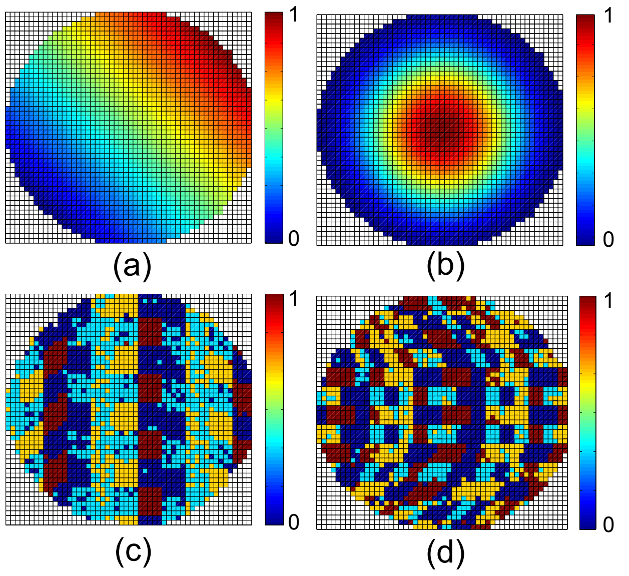

Fig. 2. Normalized wafer map of the 4 considered basis functions

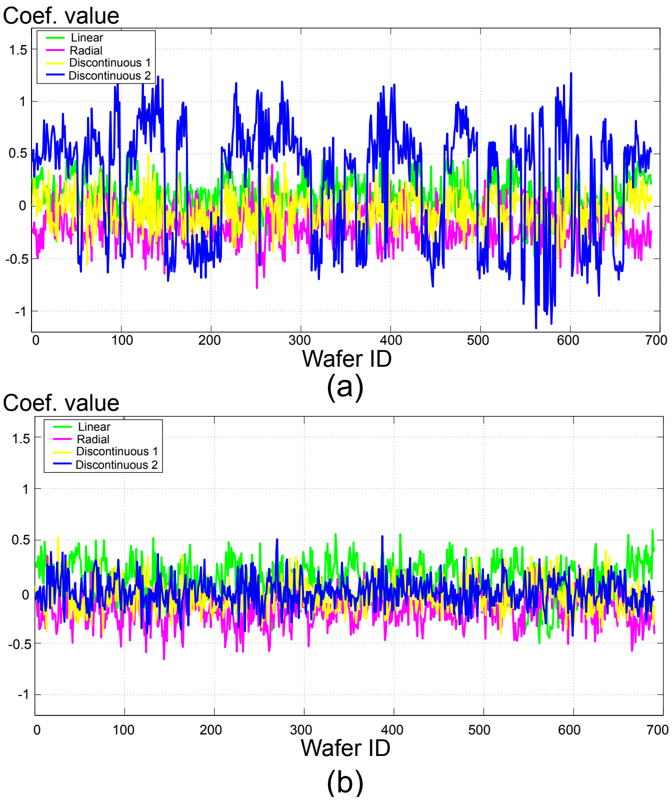

Fig. 3. Estimated coefficient $\hat{\boldsymbol{\alpha}}$ of measurement 61 computed for all the 690 wafers, using (a) proposed approach (b) approach in [5]

wafer map of the 4 considered basis functions. The parameters of basis functions $b_1(x,y)$ and $b_2(x,y)$ are dynamically learned on the first wafer using equations (3) and (4), while the number and shape of clusters in the discontinuous basis functions $b_3(x,y)$ and $b_4(x,y)$ are dynamically learned using the procedure described in section III-A3.

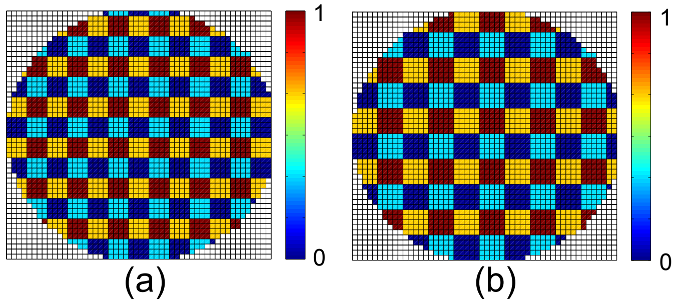

Fig. 4. Normalized wafer map of two statically chosen discontinuous basis functions

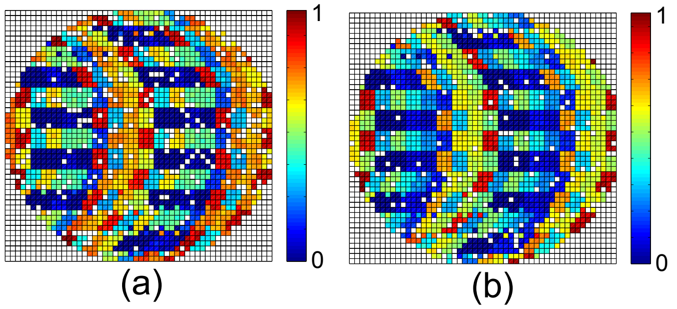

Fig. 5. Wafer maps of two randomly chosen wafers for a measurement having $b_4(x,y)$ as the most prominent basis function and yield variation contributor

A. Identification of sources of variation

As discussed in Section III-C, the principal contributor to yield variation can be identified by analyzing the constituents of the regression function mapping $\hat{\boldsymbol{\alpha}}$ to y. Accordingly, since these constituents are interpretable by the process engineer, actions to reduce process variability and enhance yield can be taken. Consider, for example, Figure 3(a), which plots the estimated coefficient $\hat{\boldsymbol{\alpha}}$ for measurement 61, wherein $b_4(x,y)$ is considered the principal contributor to yield variation, as computed through (23). The estimated coefficient vector $\hat{\boldsymbol{\alpha}}$ is computed for all the 690 wafers and the weights of the four basis functions are shown in different colors in this figure. As may be visually observed, $\hat{\alpha}_4$ exhibits the highest variance and the highest absolute value for most wafers, i.e., the yield variation for this particular measurement is mainly contributed by $b_4(x,y)$, which is in-line with the results of the prominent variation source identification analysis using (23).

1) Comparison to wafer decomposition approach in [5]: In order to demonstrate the effectiveness improvement achieved through the use of dynamically-learned basis functions, Figure 3(b) plots the estimated coefficient $\hat{\boldsymbol{\alpha}}$ of the same measurement, using only statically defined basis functions to represent discontinuous spatial variation, as introduced in [5]. In other words, instead of using the learned discontinuous basis functions shown in Figure 2(c)(d), this time we use the two basis functions shown in Figure 4, which seek to capture the same types of discontinuity in a more generic fashion. The estimated coefficients for all the 690 wafers are shown in Figure 3(b).

As can be observed even through visual inspection, the

TABLE I
COMPARISON OF ABSOLUTE MEAN VALUE OF BASIS FUNCTION COEFFICIENTS FOR MEASUREMENT 61

	$\overline{\alpha_1}$	$\overline{\alpha_2}$	$\overline{\alpha_3}$	$\overline{\alpha_4}$
Proposed approach	0.17	0.2	0.14	0.57
Approach in [5]	0.18	0.2	0.12	0.1

decomposition method using the statically defined functions is unable to accurately pinpoint the main source of variance, with linear and radial basis functions exhibiting the highest variance and the highest absolute values for most wafers, while the contribution of the discontinuous functions appears to be significantly smaller. Table I justifies this observation, by computing the absolute mean value of each coefficient: $\overline{\alpha_i}, i = 1, \ldots, 4$ over the 690 wafers. It can be observed that the proposed approach provides the highest value for $\overline{\alpha_4}$, while the approach in [5] has the lowest value for $\overline{\alpha_4}$, implying that $b_4(x, y)$ has the lowest contribution to total variation.

Finally, in order to verify that the dynamically-learned basis function $b_4(x, y)$ is indeed the most prominent one, Figure 5 shows the wafer maps of two randomly chosen wafers for this measurement. A simple visual inspection of Figure 5 and contrasting to Figure 2(d), corroborates the finding of our method and underlines the benefits of using dynamically-learned basis functions.

2) Comparison to wafer decomposition approach in [6]: The spatial variation decomposition approach proposed in [6] uses a discrete cosine transform that projects wafer spatial data into the frequency domain, in order to decompose process variation into systematic spatially correlated variation and uncorrelated random variation. For example, Figures 6(a) and (b) show the wafer decomposition of measurement 1 in our dataset, computed from the same 10% sampled die locations on the wafer, using the approach in [6]. The original wafer map is shown in Figure 6(a) and the estimated systematic variation pattern is shown in Figure 6(b). As may be observed, the estimation captures very well the underlying systematic variation, which is a radial spatial pattern. To further justify this observation, in Figure 7 we plot the estimated coefficients of basis functions for measurement 1, as computed by the method proposed herein. Evidently, our method identifies the radial basis function as the principal spatial pattern for this measurement, which is in agreement with the wafer-maps of Figures 6(a) and (b).

While the approach proposed in [6] performs very well in identifying radial spatial variation, it is not as efficient when dealing with discontinuous spatial patterns. This is demonstrated in Figures 6(c) and (d) for the case of measurement 61 in our dataset. The original wafer map for this measurement is shown in Figure 6(c) and the estimated systematic variation pattern identified by the approach in [6] is shown in Figure 6(d). As may be observed, the underlying systematic spatial variation cannot be correctly captured in this case. However, using the approach herein, the discontinuous spatial pattern is accurately captured, as was shown in Figure 3(a). Another advantage of using the proposed approach over the method described in [6] is the fact that the basis functions employed provide more actionable information. In other words, instead of showing a single systematic pattern (which could potentially be further broken down to frequency-domain components), the proposed approach decomposes the systematic variation into domain-specific basis functions which can be easily interpreted by process engineers towards improving yield.

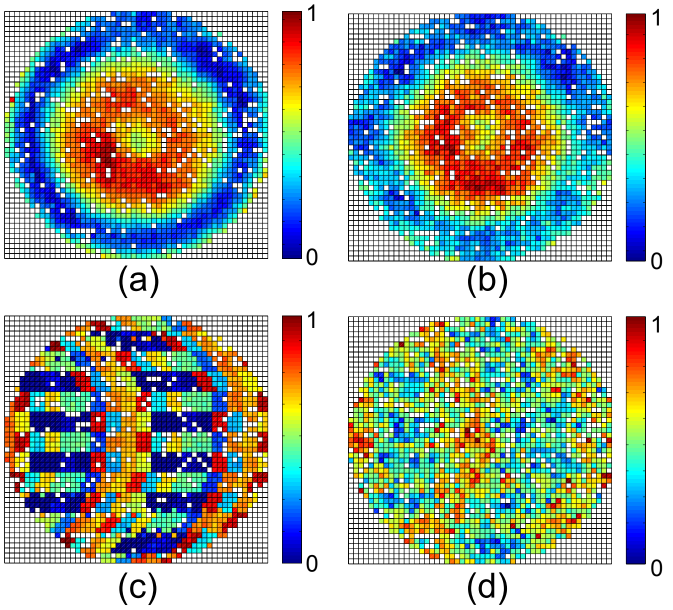

Fig. 6. Wafer spatial variation decomposition using [6], (a) original wafer map of measurement 1, (b) estimated spatially correlated systematic variation of measurement 1, (c) original wafer map of measurement 61, (b) estimated spatially correlated systematic variation of measurement 61

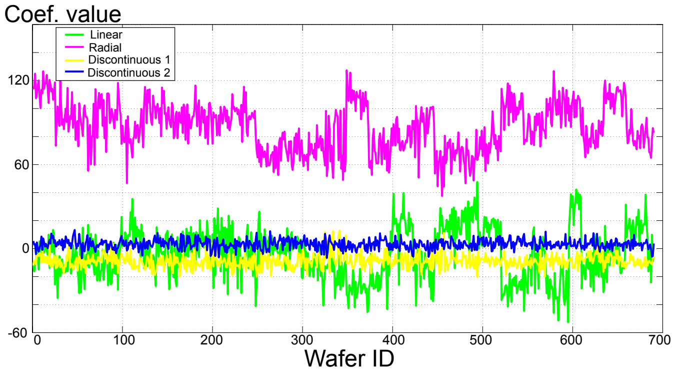

Fig. 7. Estimated coefficient $\hat{\boldsymbol{\alpha}}$ of measurement 1 computed for all the 690 wafers using the proposed approach

B. Yield prediction

To demonstrate the ability of the proposed method to accurately predict yield for each measurement from the coefficient vector $\hat{\boldsymbol{\alpha}}$, we split our wafers into two sets and generated the following data:

- The set S_t contains data collected from 345 wafers. The l-th wafer contains $\hat{\boldsymbol{\alpha}}_{kl}$ estimated by (19) for all 78 measurements: $k = 1, \ldots, 78$, using 10% of the available

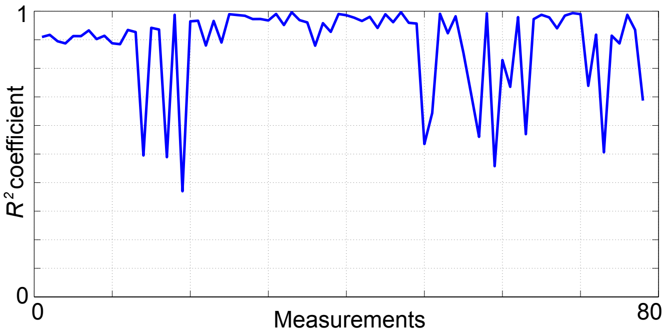

Fig. 8. R^2 coefficient between estimated and actual yield for each measurement, averaged over all 345 wafers in the validation set S_v

die locations randomly chosen on the wafer. The actual yield of the k-th measurement on the l-th wafer y_{kl} is also computed using all available die locations on the wafer. Thus, $S_t = \{(\hat{\boldsymbol{\alpha}}_{kl}, y_{kl})\}, k = 1, \ldots, 78, l = 1, \ldots, 345$.

- The set S_v contains data collected from other 345 wafers. As before, the basis function coefficient vector $\hat{\boldsymbol{\alpha}}_{kl}$ for the k-th measurement on the l-th wafer is estimated using 10% of the available die locations randomly chosen on the wafer. Thus, $S_v = \{\hat{\boldsymbol{\alpha}}_{kl}\}, k = 1, \ldots, 78, l = 346, \ldots, 690$.

We use S_t to train the regression functions f mapping $\hat{\boldsymbol{\alpha}}$ to y, as shown in (20). S_v is then used to validate the accuracy of yield prediction through the learned functions. Figure 8 shows the R^2 coefficient between the estimated and actual yield for each measurement, averaged over all the 345 wafers in the validation set S_v. As may be observed from Figure 8, most measurements have R^2 coefficient close to 1, which indicates an excellent ability of our method in predicting yield from $\hat{\boldsymbol{\alpha}}$. It should also be noted that the yield prediction is less accurate with R^2 at around 0.5 for a handful of measurements, for which random variation tends to dominate systematic spatial variation.

To gain further insight about yield prediction, Figures 9(a) and (b) plot the actual and estimated yield[1] for two randomly chosen measurements, with 0.9 and 0.98 as their corresponding R^2 coefficient, respectively. The reported yield is computed for all 345 wafers in S_v. As may be observed, the predicted yield accurately tracks the actual yield for all wafers in S_v.

C. Wafer clustering analysis

As discussed in Section III-C, the estimated $\hat{\boldsymbol{\alpha}}$ can also be used as a signature to classify wafers, in order to plan production and/or detect process excursions resulting in abnormal wafers. To illustrate the effectiveness of wafer clustering using $\hat{\boldsymbol{\alpha}}$, we consider measurement 73 in our dataset, for which $b_4(x, y)$ was identified by our method as the primary source of variation. Using $\hat{\boldsymbol{\alpha}}$ as the signature we run $k-$means clustering (see Algorithm 1 in Section III-A), using (10) to select the optimal k, which in this case is $k = 3$. Figure 10 depicts two randomly chosen wafers from each of the three clusters, (a)/(b), (c)/(d), and (e)/(f), respectively. As may be observed, wafers in the same cluster have a very similar spatial variation pattern, while wafers in different clusters are clearly distinguishable. This example demonstrates that the coefficient vector $\hat{\boldsymbol{\alpha}}$ is, indeed, a powerful spatial signature for performing wafer clustering.

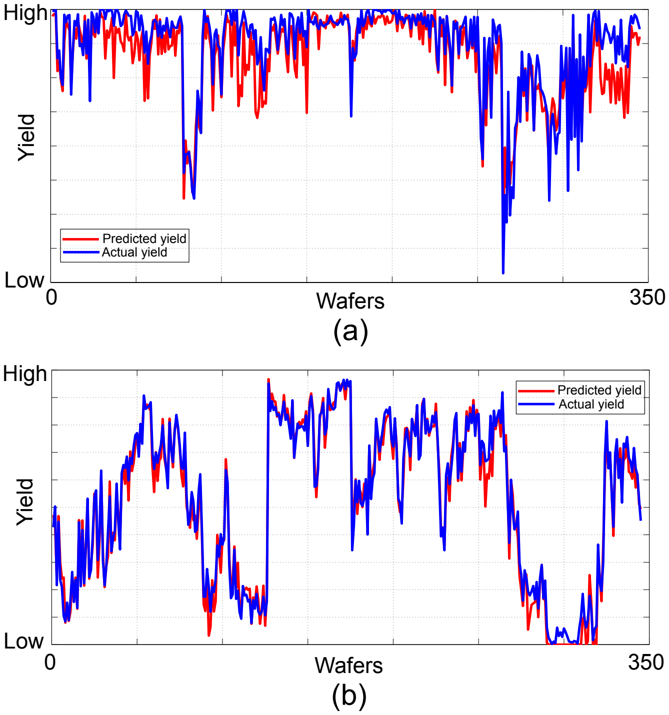

Fig. 9. Actual and estimated yield for two randomly chosen measurements, computed for all 345 wafers in set S_v.

To gain further insight on wafer clustering, we arrange the estimated vector $\hat{\boldsymbol{\alpha}}$ of measurement 73 for all wafers in a 4×690 matrix A. We then perform a Principal Component Analysis (PCA) on A, resulting in a transformed 4×690 matrix A_p, in which rows are linearly uncorrelated. In Figure 12, we project the wafers on the first 2 principal components of A_p, color-coding the cluster to which each wafer belongs. As may be observed through simple visual inspection, wafers in different clusters are clearly separated in this space.

We can also use $\hat{\boldsymbol{\alpha}}$ as a signature to detect abnormal wafers, as discussed in Section III-C3. Any outlier wafer can be easily detected by computing d_w, which is defined as the Euclidean distance between the considered wafer and the nearest cluster center in the space of $\hat{\boldsymbol{\alpha}}$. By setting a threshold value d_{th}, we can classify a wafer as an outlier if $d_w > d_{th}$. Note that d_{th} can be properly learned using a hold-out set in manufacturing. To illustrate this outlier wafer detection capability, we generate a synthetic outlier wafer[2], by randomly choosing the estimated $\hat{\boldsymbol{\alpha}}$ of one wafer, multiplying the coefficient $\hat{\alpha_2}$ of $b_2(x, y)$ by 10, and generating its wafer map using Equation (2), which is shown in Figure 11. In this way we generate realistic

[1] An NDA with Texas Instruments prohibits us from disclosing actual yield figures, hence the “High” and “Low” markers on the Y-axis of the plots.

[2] Our dataset does not contain any outlier wafers.

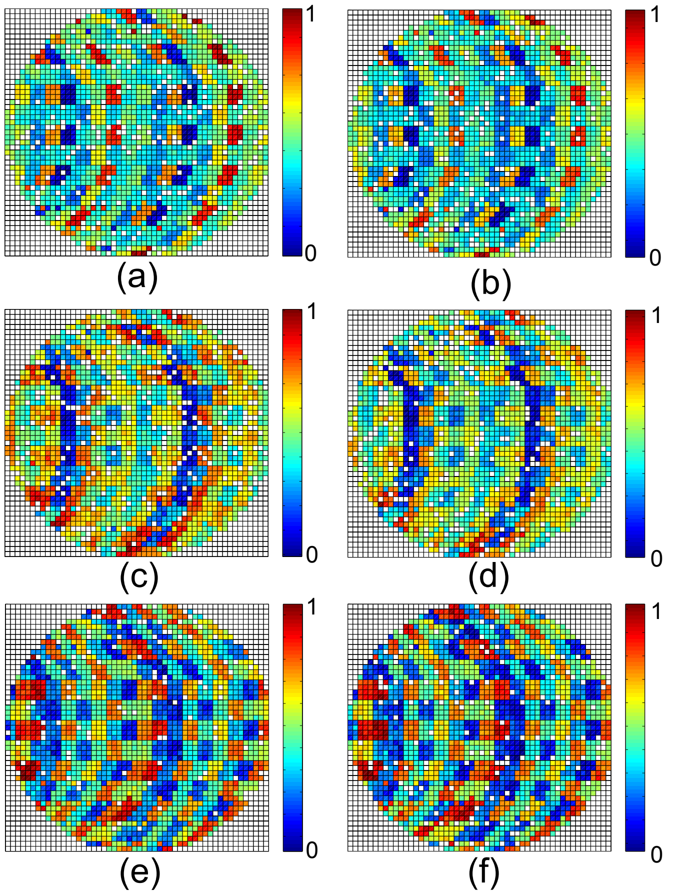

Fig. 10. Wafer clustering for measurement 73: Pairs (a)/(b), (c)/(d) and (e)/(f) show two randomly chosen wafers from each cluster, respectively

outlier wafer by considering excessive variance in radial basis function $b_2(x, y)$.

This outlier wafer is detected correctly by the procedure described above. The red dot shown in Figure 12 shows the projection of this wafer on the first two principal components using the same PCA transform. As may be observed, the outlier wafer is clearly separated from other wafers and does not belong to any cluster, demonstrating the utility of $\hat{\alpha}$ as a signature for detecting outlier wafers.

V. Conclusion

Wafer-level spatial variation decomposition offers excellent insight into the impact of process-induced uncertainty in semiconductor manufacturing. By breaking down the systematic wafer-level variation into a set of weighted spatial basis functions, the method described herein identifies and assesses the importance of different process variation sources. Its key novelty lies in the use of domain-specific, dynamically-learned, interpretable basis functions, which drastically improve its ability to accurately pinpoint variation sources over existing approaches. Using industrial high-volume manufacturing data, we demonstrated the utility of the proposed wafer-level spatial decomposition method in identifying prominent yield variation contributors, predicting yield, and clustering wafers based on their spatial variation pattern, in order to plan production and to detect process excursions resulting in abnormal wafers.

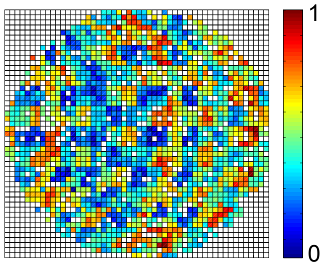

Fig. 11. Wafer map of synthetic outlier wafer

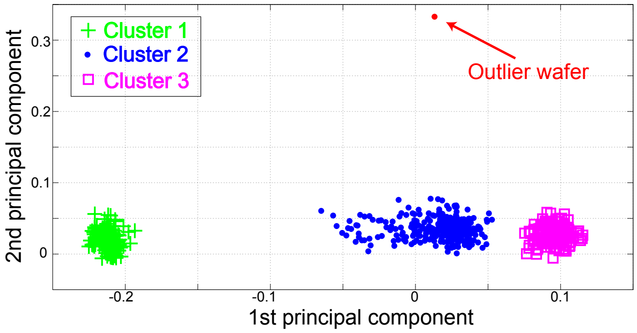

Fig. 12. Projection of wafers on the first two principal components

VI. Acknowledgement

This research has been partially supported by the National Science Foundation (NSF 1149463) and the Semiconductor Research Corporation (SRC 1836.092).

References

[1] Semiconductor Industry Association (SIA), "International technology roadmap for semiconductors (ITRS)," http://www.itrs.net/Links/2011ITRS/Home2011.htm, 2011 edition.

[2] J.K. Kibarian and A. Strojwas, "Using spatial information to analyze correlations between test structure data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 4, no. 3, pp. 219–225, 1991.

[3] L.K. Garling and G.P. Woods, "Enhancing the analysis of variance (ANOVA) technique with graphical analysis and its application to wafer processing equipment," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, vol. 17, no. 1, pp. 149–152, 1994.

[4] S. Reda and S. R. Nassif, "Accurate spatial estimation and decomposition techniques for variability characterization," *IEEE Transactions on Semiconductor Manufacturing*, vol. 23, no. 3, pp. 345–357, 2010.

[5] B.J. Whitefield, P.J. Rudolph, J.N. McNames, and B.Moon, "Pattern detection for integrated circuit substrates," U.S. Patent 7 277 813 B2, Oct. 2, 2007.

[6] W. Zhang, K. Balakrishnan, X. Li, D. Boning, and R. Rutenbar, "Toward efficient spatial variation decomposition via sparse regression," in *IEEE/ACM International Conference on Computer-Aided Design*, 2011, pp. 162–169.

[7] W. Zhang, X. Li, E. Acar, F. Liu, and R. Rutenbar, "Multi-wafer virtual probe: Minimum-cost variation characterization by exploring wafer-to-wafer correlation," in *IEEE/ACM International Conference on Computer-Aided Design*, 2010, pp. 47–54.

[8] W. Zhang, X. Li, F. Liu, E. Acar, R.A. Rutenbar, and R.D. Blanton, "Virtual probe: a statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 12, pp. 1814–1827, 2011.

[9] H.-M. Chang, K.-T. Cheng, W. Zhang, X. Li, and K.M. Butler, "Test cost reduction through performance prediction using virtual probe," in *IEEE International Test Conference*, 2011, pp. 1–9.

[10] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Design Automation Conference*, 2007, pp. 817–822.

[11] N. Kupp, K. Huang, J. Carulli, and Y. Makris, "Spatial correlation modeling for probe test cost reduction," in *IEEE/ACM International Conference on Computer-Aided Design*, 2012, pp. 23–29.

[12] N. Kupp, K. Huang, J. Carulli, and Y. Makris, "Spatial estimatin of wafer measurement parameters using gaussian process models," in *IEEE International Test Conference*, 2012, pp. 1–8.

[13] K. Huang, N. Kupp, J. Carulli, and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," in *Design Automation and Test in Europe*, 2013.

[14] C. Yu, H.Y. Liu, and C. J. Spanos, "Patterning tool characterization by causal variability decomposition," *IEEE Transactions on Semiconductor Manufacturing*, vol. 9, no. 4, pp. 527–535, 1996.

[15] B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, no. 1, pp. 24–41, 1997.

[16] W.R. Daasch, J. McNames, D. Bockelman, and K. Cota, "Variance reduction using wafer patterns in I_{ddQ} data," in *IEEE International Test Conference*, 2000, pp. 189–198.

[17] A. Gattiker, "Unraveling variability for process/product improvement," in *IEEE International Test Conference*, 2008, pp. 1–9.

[18] J. McNames, B. Moon, B. Whitefield, and D. Abercrombie, "Robust linear regression for modeling systematic spatial wafer variation," *Proc. SPIE, Data Analysis and Modeling for Process Control II*, vol. 5755, no. 87, 2005.

[19] D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[20] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrica*, vol. 50, no. 2, pp. 159–179, 1985.

[21] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[22] P. Huber, *Robust statistics*, Wiley series in probability and statistics. John Wiley & Sons, 1981.

[23] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.