Spatio-Temporal Wafer-Level Correlation Modeling with Progressive Sampling: A Pathway to HVM Yield Estimation

Ali Ahmadi*, Ke Huang†, Suriyaprakash Natarajan‡, John M. Carulli Jr.§, and Yiorgos Makris*

*Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080
†Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA 92115
‡Intel Corp., 2200 Mission College Boulevard, Santa Clara, CA 95054
§Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, TX 75243

***Abstract*—Wafer-level spatial correlation modeling of probe-test measurements has been explored in the past as an avenue to test cost and test time reduction. In this work, we first improve the accuracy of a popular Gaussian process-based wafer-level spatial correlation method through two key enhancements: (i) confidence estimation-based progressive sampling, and, (ii) inclusion of spatio-temporal features for inter-wafer trend learning. We then explore a new application of the enhanced correlation modeling method in estimating High Volume Manufacturing (HVM) yield from a small set of early wafers and we demonstrate its effectiveness on a large set of actual industrial test data.**

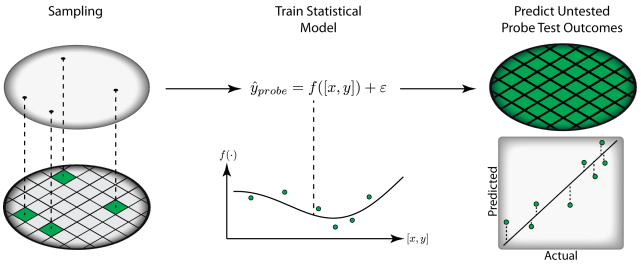


Fig. 1: Wafer measurement spatial interpolation [5]

I. Introduction

Recent research on modeling spatial measurement correlation has shown great promise in capturing wafer-level spatial variation and, thereby, reducing test cost of electrical measurements [1]–[7]. The underlying idea, as shown in Figure 1, 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. For example, in [2], the expectation-maximization (EM) algorithm is used to estimate spatial wafer measurements, assuming that data comes from a multivariate normal distribution and the Box-Cox transformation is used in case data is not normally distributed. The "Virtual Probe" (VP) approach [3] models spatial variation via a Discrete Cosine Transform (DCT) that projects spatial statistics into the frequency domain. The author of [1] laid the groundwork for applying Gaussian Process (GP) models to spatial interpolation of semiconductor data based on Generalized Least Square fitting and a structured correlation function. This fundamental model has been further enhanced using radial feature inclusion, multiple kernel evaluation and introduction of a regularization parameter [5], [6], as well as a clustering approach to handle spatial discontinuous effects [7]. The resulting comprehensive GP model has significantly improved both prediction accuracy and computational time, as compared to the VP model, and is therefore our starting point.

In the work presented herein, we first seek to improve the accuracy of the state-of-the-art wafer-level spatial correlation modeling methods. To this end, we introduce two enhancements, namely *progressive sampling* and *spatio-temporal feature inclusion*. Progressive sampling aims at improving the statistical information available in the sample based on which a spatial correlation model is constructed. This is achieved by starting with measurements from a small set of die locations, which is iteratively augmented with more samples. In each iteration, the accuracy of the spatial correlation model is improved by selecting new samples from regions where the confidence of the previous model is low. Spatio-temporal feature inclusion, on the other hand, extends the concept of correlation modeling across wafers, asserting that the spatial variation observed on a wafer reveals useful information for other wafers in the same lot. Accordingly, we construct a single spatio-temporal correlation model which captures variation of a parameter across all wafers in a lot as a function of die coordinates and wafer time-index. Furthermore, these two enhancements can also be combined, as they use different means to improve the accuracy of the correlation models.

Besides the straightforward objective of improving the accuracy of the models used for test cost / test time reduction, we also explore another application of the improved spatio-temporal correlation modeling with progressive sampling method. Specifically, we seek to employ such models in predicting the HVM yield of a device from measurements obtained on a sample of die from a small set of wafers available in early production. This resembles the typical problem faced in post-silicon validation, where the performance distribution of HVM devices needs to be extrapolated from a few sample die obtained from a few engineering wafers produced over the period of a few months. To address this problem, we propose to use these samples in order to build a spatio-temporal model

978-1-4799-4722-5/14/$31.00

that predicts the distribution of each parameter of interest in HVM production. In essence, the proposed method leverages better the information available on the few early wafers, even though only a few die are sampled on each such wafer. Accordingly, better statistical models can be derived either in isolation or in conjunction with previously developed synthetic population generation/enhancement methods [8].

The remainder of this paper is organized as follows: in section II, we discuss the Gaussian Process model for capturing wafer-level spatial correlation. In section III, we introduce the two proposed enhancements, namely progressive sampling and inclusion of spatio-temporal features. Then, in section IV, we explore the utility of the enhanced spatial correlation modeling method in accurately predicting HVM yield from a small number of early silicon wafers. Experimental results demonstrating the effectiveness of the proposed methods on industrial data are presented in section V and conclusions are drawn in section VI.

II. Gaussian Process

In this section, we briefly present the Gaussian process model; a detailed explanation can be found in [9]. A Gaussian process is a collection of random variables, any finite number of which exhibits a joint Gaussian distribution. A Gaussian process is fully specified by its mean function and its kernel-based covariance function. Consider a training set $\mathbf{D}$ of n_t data points, $\mathbf{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i))|i = 1, \ldots, n_t\}$, where $\mathbf{x}$ denotes an input vector (in this work, the input vector is the Cartesian coordinate of a die denoted as $\mathbf{x} = [x, y]$) and $f(\mathbf{x})$ is the output (herein, a measurement value). Accordingly, a Gaussian process can be viewed as a group of random variables $f(\mathbf{x}_i)$ with joint Gaussian distribution:

$$f(\mathbf{x}_1), ..., f(\mathbf{x}_n) \sim \mathcal{N}(0, K) \tag{1}$$

where element $\mathbf{K}_{ij}$ of the covariance matrix $\mathbf{K}$ is the covariance between values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$. This covariance function can be formed as an inner product, permitting us to leverage the kernel trick [10] and express it as a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. Thus, the covariance between the outputs can be written as a function of inputs using a kernel function. Many kernel functions exist, and any function $k(\cdot, \cdot)$ that satisfies Mercer's condition [11] is a valid kernel function. However, only a handful of kernels are commonly used. Among these common kernels, the most prevalent one is the squared exponential, also known as the radial basis function kernel. In this work, we employ a squared exponential kernel of the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2l^2}|\mathbf{x}_i - \mathbf{x}_j|^2\right) \tag{2}$$

where l is some characteristic length-scale of the squared exponential kernel. This function expresses that neighbor instances will have highly correlated outputs. Employing this kernel is equivalent to training a linear regression model with an infinite-dimensional feature space. Substituting our squared-exponential covariance function into the definition of the Gaussian process, we arrive at a Gaussian process formulation as:

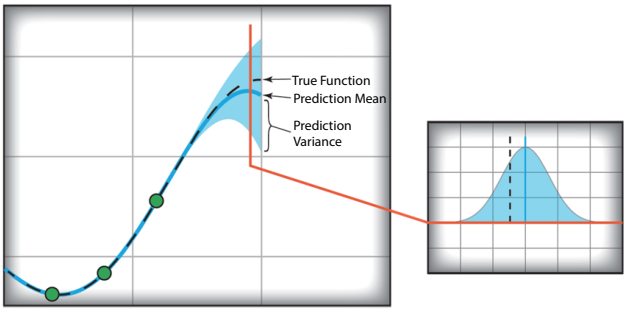


Fig. 2: Distribution prediction using Gaussian process

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}_i, \mathbf{x}_j)) \tag{3}$$

For a new data point with input $\mathbf{x}_*$, the predictive distribution of the output $f_*(\mathbf{x}_*)$ can be computed by using conditional distributions of the joint Gaussian distribution:

$$f_*|X, \mathbf{t}, \mathbf{x}_* \sim \mathcal{N}(\mathbf{k}_*^\top K^{-1}\mathbf{t}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top K^{-1}\mathbf{k}_* \tag{4}$$

where $\mathbf{X}$ is the input training matrix, $\mathbf{t}$ is a training output vector, $\mathbf{x}_*$ is a test point, $\mathbf{k}_* = K(X, \mathbf{x}_*)$ which is the kernel evaluation between the test point and all training instances, $\mathbf{K}$ is the matrix of the kernel function evaluated at all pairs of training points and $k(\mathbf{x}_*, \mathbf{x}_*)$ is the variance of the kernel function at test point $\mathbf{x}_*$.

The original GP method is primarily concerned with point predictions, so it simply uses the distribution mean $\bar{f}_* = \mathbf{k}_*^\top K^{-1}\mathbf{t}$ to generate a point prediction from the predictive distribution. This corresponds to decision-theoretic risk minimization [12] using a squared-loss function. However, in addition to simply providing a prediction with a fixed value, this approach also reports a *confidence level* for the prediction by computing the distribution of predicted values at unobserved locations. Figure 2 illustrates the distribution prediction using the Gaussian process. This feature will be leveraged in the next section in order to improve the model through intelligent selection of the samples from which the model is learned.

III. Proposed Enhancements

In this section, we introduce two enhancements to the aforementioned Gaussian Process model. The first enhancement seeks to improve the information available in the sample from which the model is learned, by starting with a small set and progressively selecting additional samples in areas where confidence of the learned model is low. The second enhancement seeks to expand the correlation model across wafers within the same lot by introducing a spatio-temporal feature, namely the order of the wafer within the lot. The combination of these two enhancements is also explored. Finally, practical considerations related to the deployment of

these enhancements in the context of test cost / test time reduction are discussed.

A. Progressive Sampling (GP-PS)

Selecting a sample of die locations that accurately reflect the spatial variation across a wafer is crucial in any wafer-level spatial correlation modeling method. To date, most methods rely on a set of randomly selected die locations. Attempts to guarantee sufficient coverage by employing a Latin hypercube sampling approach to evenly choose random sample points over the entire wafer have also been made [3]. Yet all samples are usually taken at once, without taking into consideration any *a priori* knowledge of spatial variation patterns on the wafer. In contrast, herein we propose an iterative progressive sampling approach, in order to select training samples which better represent the spatial variation pattern across a wafer. To achieve this, we leverage the ability of the GP model to provide a confidence level for all predicted samples in each iteration. Algorithm 1 outlines the proposed progressive sampling approach.

In particular, we begin the sampling procedure with n' samples randomly chosen on the wafer. Note that n' is set to be significantly smaller than the total number of samples, n, allowed in our budget for building the model. Using these samples we build a GP model and we predict the values along with the prediction confidence level at each unobserved die location. This confidence level is then used to guide our sampling at the next iteration, towards reducing the uncertainty and improving the accuracy of the GP model. Specifically, we identify a set of k locations for which the predictions have the highest uncertainty and lowest confidence level, we sample them to obtain the true values, and we then use them to augment our training sample. We note that the selection of these k new locations also uses their Euclidean distance as a metric, in order to distribute the new samples across many areas of low confidence. This progressive sampling process is repeated until a stopping criterion is reached. In Figure 3, we illustrate an example of progressive sampling in 2 iterations. As can be observed, carefully selecting a new sample in the training set can significantly improve the accuracy of the spatial correlation model.

The stopping criterion of Algorithm 1 depends on the application and prediction problem. There are two standard methods: i) when the highest uncertainty of prediction drops below a given threshold, or ii) when a given budget of samples, n, is reached. In this work, we use the latter stoping criterion and all experiments are based on a given sampling budget (i.e. 10%) of die on a wafer. The number of samples added in each iteration, k, is also problem-dependent. In this work, we chose to add the same number of samples (i.e. 2.5%) in each iteration, resulting in 4 iterations of the progressive sampling algorithm, until our sampling budget is reached.

B. Spatio-Temporal Feature Inclusion (GP-ST)

A key contribution of this work is the extension of Gaussian process modeling over spatial coordinates to a joint spatio-temporal space, capturing our intuition that wafer-level spatial variance is also time-dependent. Indeed, we expect that wafers processed together, such as wafers within the same lot, will exhibit similar intra-wafer spatial variation and will also exhibit time-dependent inter-wafer variation which would be beneficial to include in our model. Essentially, our conjecture is that *a single spatio-temporal model learned from samples*

1. Randomly select n' samples on the wafer as initial training set: $S = \{\mathbf{x}_1|t_1, \cdots, \mathbf{x}_{n'}|t_{n'}\}$
2. Build spatial GP model using set S and predict values and confidence at unobserved die locations (set U)
3.1 For each $\mathbf{x}_i$ in U, calculate $\mathbf{d}_i = min\{|\mathbf{x}_i - \mathbf{x}_j|^2, \forall \mathbf{x}_j \in S\}$
3.2 Select location $\mathbf{x}_i$ which has highest variance and maximum Euclidean distance from current training set
3.3 Add $\mathbf{x}_i$ to the set S and remove it from set U and obtain corresponding true value $\mathbf{t}_{x_i}$
3.4 Repeat 3.1-3.3 until k locations are added to the training set
4. Augment the training set $S = \{S, \mathbf{x}_{h_1}|t_{h_1}, \cdots, \mathbf{x}_{h_k}|t_{h_k}\}$
5. Repeat steps 2-4, until stopping criterion is reached

Algorithm 1: Progressive sampling of information-rich training locations in spatial correlation modeling

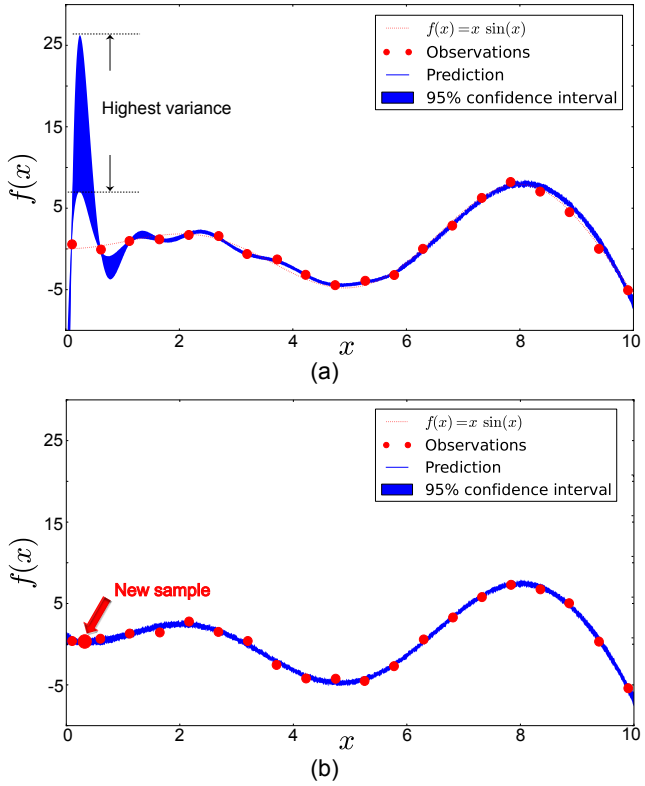


Fig. 3: Example of progressive sampling with prediction at (a) iteration i and (b) iteration $i+1$

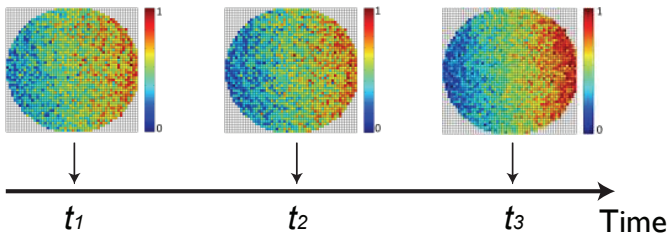


Fig. 4: Time-dependent spatial variation

across all wafers in a lot, could be more accurate than individual models learned from the same samples for each wafer. An advantage of using Gaussian process regression is the ability to apply a Gaussian process over any arbitrary index set. Thus far, we have been describing a Gaussian process implementation that estimates wafer-level measurements over a 2D Cartesian plane; however, we are free to use any other field. As noted above, we expect wafer-level spatial measurements to exhibit time dependence in manufacturing. To accommodate this in our Gaussian process model, we can simply update our coordinates from $\mathbf{x} = [x, y]$ to include a time feature t:

$$\mathbf{x} = [x, y, t]$$

Applying Gaussian process regression over this space will result in a model that takes time dependent variation into account. In Figure 4, we illustrate the concept of time-dependent wafer-level spatial variation. Evidently, the spatial variation exhibited by each of these three consecutively produced wafers is not only similar but also exhibits a temporal trend (i.e. a linear gradient that intensifies over time).

C. Spatio-temporal GP w. Progressive Sampling (GP-ST-PS)

Combining the two new enhancements to wafer-level correlation modeling is also a plausible direction. Specifically, it is possible that the accuracy of a spatio-temporal model can be improved by employing progressive sampling. Indeed, the two enhancements are orthogonal, in the sense that progressive sampling seeks to provide a training set which better reflects the underlying statistics of the problem, while spatio-temporal modeling explores the correlation which exists both within a wafer (spatial variation) and across wafers (time-dependent variation), to establish a more comprehensive model. In order to assist GP in learning time-dependent variation, we modify the sampling strategy such that the initial portion of the sampling budget covers common locations over all the wafers in a lot. Then, the progressive sampling method is applied to the entire lot as a whole for the remaining portion of the sampling budget. Again, our conjecture is that wafers in a lot have both spatial and temporal correlations which can be effectively learned by a spatio-temporal GP model, hence we train one GP model for all the wafers in a lot. While by combining these two models we can improve the prediction accuracy, this would require multiple insertion of wafers which might not be practical. In the next section, we discuss practical considerations of the proposed models.

D. Considerations for Production Test Deployment

Spatial correlation models have primarily been seeking to achieve test-cost / test-time reduction at probe, by landing on a limited set of die locations and then using the correlation models to predict the performances of the unobserved die. The progressive sampling method outlined in Section III-A adds a complication to this flow, since the wafer will need to remain on the probe equipment while the models are constructed and applied, and samples will need to be obtained iteratively through multiple passes. This iterative procedure may become overly time-consuming and may eventually counterbalance the time reduction achieved by the spatial correlation model. Clearly, this trade-off between test time and prediction accuracy needs to be considered in order to decide whether progressive sampling makes sense in an HVM test.

Spatio-temporal modeling, on the other hand, as described in Section III-B, brings about a different challenge. Specifically, after measurements on a sparse sample of die are obtained at probe, the wafer will leave the probe equipment without the ability to make decisions on the unobserved locations. Indeed, one will have to wait until all wafers have passed through the probe equipment, so that the obtained samples can be used to build the spatio-temporal model, which will then be used to predict the unobserved measurements for all wafers in the lot. While the technical capabilities for coding this in modern ATE are available, if the overall confidence of the model is low and the rest of the locations on each wafer need to be explicitly measured, then the wafers will have to be subjected to a second test insertion. Therefore, the use of spatio-temporal models should be sparingly deployed for measurements that tend to exhibit strong correlation.

Evidently, in the case where spatio-temporal GP models are combined with progressive sampling both of these considerations will have to be taken into account. Overall, individually applying either spatio-temporal modeling or progressive sampling results in practical and effective solutions for pass/fail label determination. However, the combined model may not be practical to deploy for production testing in an HVM environment where throughput is crucial, since it requires multiple test insertions. Nevertheless, the accuracy improvement that the combined model brings about can be leveraged in a different application, which we outline next.

IV. High Volume Manufacturing Yield Estimation

Prior to commencing High Volume Manufacturing (HVM), die samples from early silicon wafers are obtained and subjected to thorough characterization. The objective of such characterization is manyfold and includes post-silicon design validation as well as HVM performance and yield estimation with better accuracy than what pre-silicon Monte Carlo simulations may offer [13]–[15]. Indeed, such tasks are crucial as they rely on a few die in order to make both manufacturing and marketing decisions that affect the production lifetime

of a device. In reality, only a small number of engineering wafers are available for such analysis and only a small number of die are measured from each such wafer. Typically, such early silicon is only repeated a few times over a fairly short period of time prior to HVM. However, analysis of such die is extensive and is done in dedicated characterization labs without the time and throughput constraints of production testing, which were outlined in Section III-D. This setting lends itself naturally to the proposed GP-ST-PS method, as it is possible to progressively sample die across available wafers and repeat the process every time new engineering wafers are received, in order to obtain a coherent picture of the performance distribution and expected yield of the product.

Essentially, given a small set of observations over a few wafers, our objective is to infer the probability density function (PDF) of the entire population from the limited observations available. In general, density estimation approaches are categorized into parametric and non-parametric. In parametric methods, assumptions about the form of density are made, while in the non-parametric no such assumptions are made. Herein, we use the latter, since we have no information regarding the form of the density function. Two options are explored, namely a simple histogram-based estimation method and a more advanced kernel density estimation method.

A. Histogram with Random Sampling

A simple and straightforward method for density estimation from a small set of samples is a histogram. To create a histogram, the range of n observations is divided into bins and the number of samples which fall in each bin is counted. The probability density is, then, estimated as:

$$f'(\mathbf{x}) = \frac{n_b}{n * h} \; for \; \mathbf{l_b} \leq \mathbf{x} \leq \mathbf{u_b} \tag{5}$$

where $\mathbf{n_b}$ is the number of observations in bin $\mathbf{b}$, $\mathbf{u_b}$ and $\mathbf{l_b}$ are the upper and lower limits of bin $\mathbf{b}$, and $\mathbf{h} = \mathbf{u_b} - \mathbf{l_b}$.

Evidently, the accuracy of this estimate depends on the choice of the random sample and, more specifically, by how accurately it reflects the statistics underlying the observed phenomenon. In the context of our problem, the randomly chosen die may be limited in representing the statistical distribution across the available wafers, let alone across the entire HVM production, especially since no notion of spatial or temporal correlation is captured in a histogram.

B. Histogram with GP-ST-PS

To address this limitation, we can improve the quality of the sample that is used to generate the histogram by employing the proposed spatio-temporal Gaussian process method with progressive sampling (GP-ST-PS). Instead of estimating the histogram using a random set of n samples from few available wafers, we employ the GP-ST-PS method to intelligently select these n samples in order to build an accurate spatio-temporal GP model which can subsequently be used to accurately predict the values of all unobserved die locations across the available wafers. In this sense, we increase the number and utility of available points (i.e. both observed and predicted) for building the histogram and estimating the probability density function, without increasing the cost of sampling. Assuming that the GP-ST-PS method provides accurate estimates by leveraging the spatio-temporal correlation across the available wafers, the new histogram should yield a more accurate density function than the one obtained from only the n observed die samples.

C. Kernel Density Estimation

While the histogram method is easy to implement and straightforward to interpret, it has several disadvantages. For example, the choice in the number and distribution of bins dominantly affects the estimation. Furthermore, the tails of the distribution, which have a very low likelihood of being represented in the sample, will remain insufficiently modeled. To address these issues, the concept of kernel density estimation (KDE) has been proposed for density estimation and enhanced synthetic population generation. In order to generate a large volume of synthetic data which accurately reflects the distribution of measurements in high-volume production from a small number of observations, we use a non-parametric KDE method [8]. This method relies on the estimation of the densities $f(\vec{t})$, using the available observations $\vec{t_i}$, $i = 1, \cdots, M$, where M is the number of available samples used to build the density. We do not make any assumption regarding its parametric form (e.g. normal). Instead, we use non-parametric KDE, which allows the observations to speak for themselves. The kernel density estimate is defined as [16]

$$\hat{f}(\vec{t}) = \frac{1}{M \times h^d} \sum_{i=1}^{M} K_e(\frac{1}{h}(\vec{t} - \vec{t_i})) \tag{6}$$

where h is a parameter called bandwidth, $d = n_m$ is the dimension of $\vec{t}$, and $K_e(m)$ is the Epanechnikov kernel

$$K_e(m) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)(1 - m^T m) & \text{if} \quad m^T m < 1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and $c_d = 2\pi^{d/2}/(d \cdot \Gamma(d/2))$ is the volume of the unit d-dimensional sphere. The kernel density estimate can be interpreted as the normalized sum of a set of identical kernels centered on the available observations for the 1-dimensional case. The bandwidth h corresponds to the distance between the center of the kernel and the kernel's edge, where the kernel's density becomes zero.

In this work, we use an adaptive version of (6). In particular, we allow the bandwidth h to vary from one observation $\vec{t_i}$ to another, allowing larger bandwidths for the observations that lie at the tails of the distribution. The adaptive kernel density estimate is defined as [16]

$$\hat{f}_\alpha(\vec{t}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{(h \cdot \lambda_i)^d} K_e(\frac{1}{h \cdot \lambda_i}(\vec{t} - \vec{t_i})) \tag{8}$$

where the local bandwidth factors λ_i are defined as

$$\lambda_i = \{\hat{f}(\vec{t_i})/g\}^{-\alpha}, \tag{9}$$

$\hat{f}(\vec{t}_i)$ is the pilot density estimate given in (6), g is the geometric mean

$$\log g = M^{-1} \sum_{i=1}^{M} \log \hat{f}(\vec{t}_i) \tag{10}$$

and α is a parameter which controls the local bandwidths. The larger the α, the larger the measurement space where the density $\hat{f}(\vec{t})$ is nonzero.

Once the probability density $\hat{f}(\vec{t})$ is estimated, we can sample $\hat{f}(\vec{t})$ to generate a large synthetic data set $s_n = \{\vec{t}_1, \ldots, \vec{t}_{M'}\}, M' \gg M$, which will better reflect the distribution tails.

Of course, in the context of our problem, KDE can be used either with the initial n die samples as the starting point (i.e. KDE with Random Sampling), or with the enhanced sample as predicted by the GP-ST-PS method (KDE with GP-ST-PS Sampling). The expectation here is that GP-ST-PS method will provide a better starting point for KDE, which will ultimately lead to a more accurate estimation of the probability density.

V. Experimental Results

We now evaluate the effectiveness of the proposed methods using a 65nm analog/RF device[1] currently in HVM production. Our dataset comprises a total of 300 time-stamped wafers, grouped in lots of 8-12 wafers, and produced over a period of 6 months. Each wafer has approximately 5500 devices, on which 39 parametric probe test measurements are collected. Our experiments seek to (i) quantify the accuracy improvement achieved by progressive sampling and spatio-temporal modeling in predicting probe tests using wafer-level spatial correlation, and (ii) assess the ability of the enhanced model in predicting HVM yield based on measurements from a small number of early production wafers.

A. Accuracy Improvement of Enhanced Model

In all of the experiments described below we sample 10% of the die locations on each wafer for training the correlation models, which are subsequently used to predict the values on the remaining 90% of non-sampled die locations. The predicted values are then compared to the actual values that we have in our dataset, in order to calculate and report the prediction error of each of the compared models. In the case of the baseline spatial correlation model, which is used as a reference point, the 10% sample is randomly selected across the wafer. In the case of progressive sampling, this 10% sampling budget is reached in multiple steps, where the selection of samples added in each step is guided by the models constructed using the samples of all previous steps. In the case of the spatio-temporal models, each lot of wafers is treated holistically. Specifically, we use the 10% sample of (randomly or progressively selected) die from each wafer in the lot to construct one spatio-temporal correlation model for the entire lot, which is subsequently used to predict the values on the remaining 90% of die locations on each of the wafers in the lot. In summary, the four prediction models that we compare in our experiment are the following:

- **Gaussian Process (GP):** Given a wafer, randomly select 10% of the die on this wafer, measure the parameter of interest and train a Gaussian Process model to predict this parameter as a function of die coordinates on the wafer. Then, use this model to predict this parameter for the remaining 90% of die on the wafer.
- **GP with Progressive Sampling (GP-PS):** Given a wafer, select die locations using the progressive sampling algorithm (Algorithm 1) in increments of 2.5%. In each iteration, use all available samples to identify the die locations where model confidence is low, in order to select the next die sample increment. Once the overall sampling budget of 10% is reached (i.e. after 4 iterations), use the final model to predict this parameter for the remaining 90% of die on the wafer.
- **GP with Spatio-temporal Features (GP-ST):** Given a lot of wafers, randomly select 10% of die from each wafer and train one spatio-temporal Gaussian Process model (i.e. a model that expresses a parameter as a function of die coordinates as well as the temporal order of a wafer within its lot) for the entire lot. Then use this spatio-temporal model to predict this parameter for the remaining 90% of die on all wafers in the lot.
- **GP with Spatio-temporal Features & Progressive Sampling (GP-ST-PS):** Given a lot of wafers, select die locations on each wafer using the progressive sampling algorithm in increments of 2.5% per lot. In the first increment, the same random locations are chosen on each wafer, in order to help the model capture time-dependent correlation. Subsequent sample increments are chosen across all wafers in the lot based on the prediction confidence level of the spatio-temporal correlation model constructed using all previous samples from all the wafers. Once the overall sampling budget of 10% is reached (i.e. after 4 iterations), use the final spatio-temporal correlation model to predict this parameter for the remaining 90% of die on all wafers in the lot.

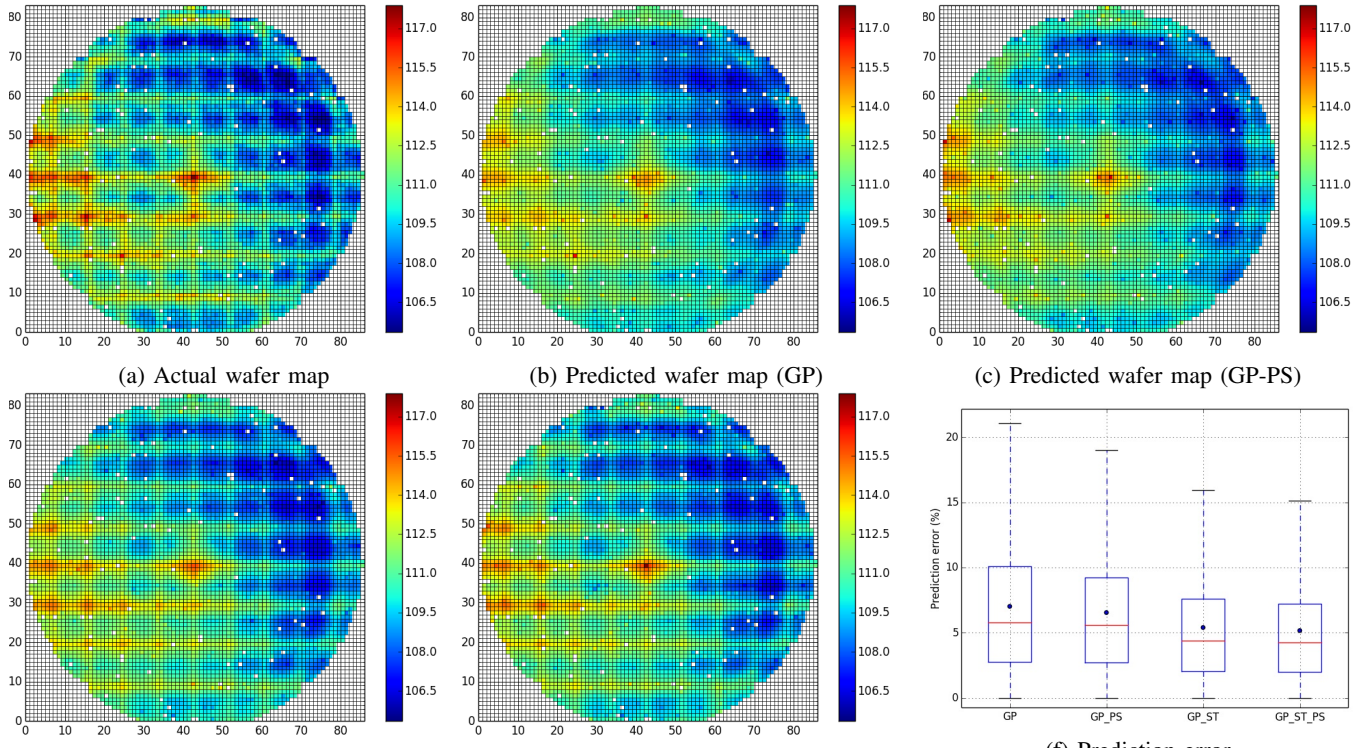
In order to assess the accuracy of each model, we compare the values for the predicted die locations (90%) to the actual probe test outcomes and we capture the discrepancy using the absolute percentile error metric:

$$\epsilon = \frac{|\mathbf{t}' - \mathbf{t}|}{\text{Specification Range}} \tag{11}$$

where $\mathbf{t}$ is the probe test outcome for a die, $\mathbf{t}'$ is the corresponding predicted value for that die and *Specification Range* is the range of that measurement across the wafer after outlier removal using a $\pm 3\sigma$ filter.

Figure 5 presents the impact of the introduced correlation model enhancement methods on a randomly chosen parameter (i.e. measurement 16) among the 39 probe-tests in our dataset, on one randomly chosen wafer. Figure 5(a) shows the actual wafer map while Figures 5(b)-(e) show the predicted wafer

[1]Details regarding the device cannot be released due to an NDA under which this dataset has been provided to us.



(a) Actual wafer map (b) Predicted wafer map (GP) (c) Predicted wafer map (GP-PS)

(d) Predicted wafer map (GP-ST) (e) Predicted wafer map (GP-ST-PS) (f) Prediction error

Fig. 5: Actual and predicted wafer maps and prediction error for measurement 16

maps using each of the four methods, i.e. GP, GP-PS, GP-ST and GP-ST-PS, respectively. Even though the differences are subtle, they are still visible through the wafer maps, where it may be observed that the predictions by GP-PS, GP-ST, and GP-ST-PS are better than the prediction of the original GP and become progressively more accurate with respect to the actual wafer map. The box-plot of Figure 5(f) quantifies this improvement by reporting the prediction error for each of the four models. The $\mathbf{y-axis}$ shows the percentile prediction error, with the black dot on each bar representing the mean of the prediction error across all die on the wafer.

Figure 6 shows the prediction error of the four methods for measurement 16 across all 300 wafers. The mean error is $\mathbf{7.3}\%$, $\mathbf{6.9}\%$, $\mathbf{5.9}\%$ and $\mathbf{5.6}\%$ for GP, GP-PS, GP-ST and GP-ST-PS, respectively. We note that the proposed models not only generate a lower prediction error in the test data, but also result in tighter error bars than the baseline model, which indicates that the variance of error is also smaller.

Finally, Figure 7 summarizes the mean error of the four methods for all 39 measurements across all 300 wafers. As may be observed, the enhancements proposed herein consistently result in lower prediction error for all measurements. In order to demonstrate the average improvement over the original GP model, we use the difference in mean relative error (MRE) as defined below:

$$\delta\text{-MRE} = \left| \frac{\text{GP Error} - \text{Proposed Method Error}}{\text{GP Error}} \times 100 \right|$$

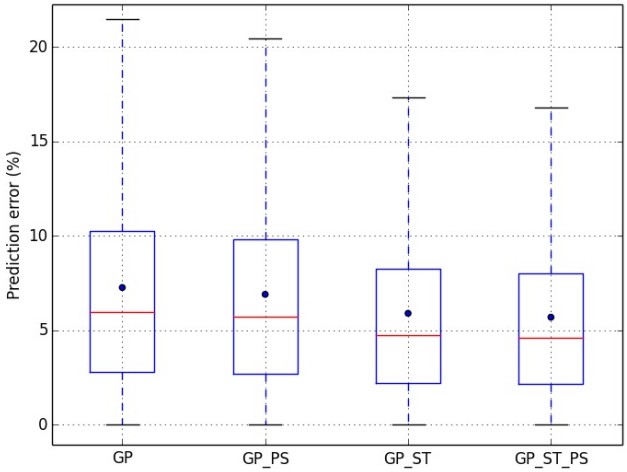


Fig. 6: Measurement 16 prediction error over all wafers

Accordingly, we calculate the δ-MRE between GP and GP-PS, GP-ST, and GP-ST-PS as $\mathbf{2.6}\%$, $\mathbf{11}\%$, and $\mathbf{16}\%$, respectively.

Based on the above results, we observe the following:

- Enhancing the original Gaussian Process based wafer-level spatial correlation method with temporal inter-wafer information and progressive sampling results in a notable improvement in the accuracy of the prediction model.
- The improvement obtained by spatio-temporal modeling is significantly higher than the improvement obtained by progressive sampling. This is attributed to the fact that the

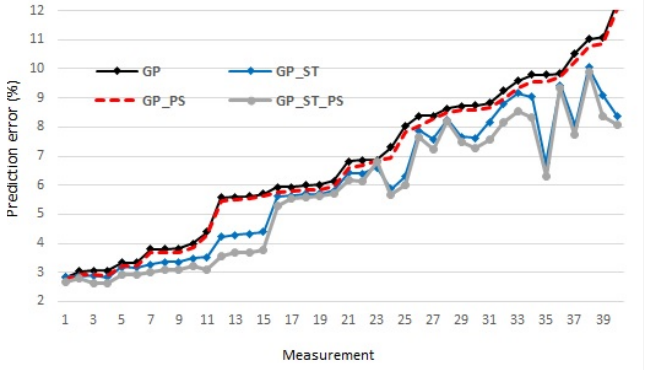


Fig. 7: Prediction error for 39 measurements over all wafers

locations sampled on each wafer in a lot are different; therefore, since wafers in the same lot are expected to be affected by the same systematic variation sources, sampled locations in a wafer carry valuable information regarding the same locations on other wafers, which were not included in the sample. Thereby, the collective statistics for the entire wafer are significantly improved.

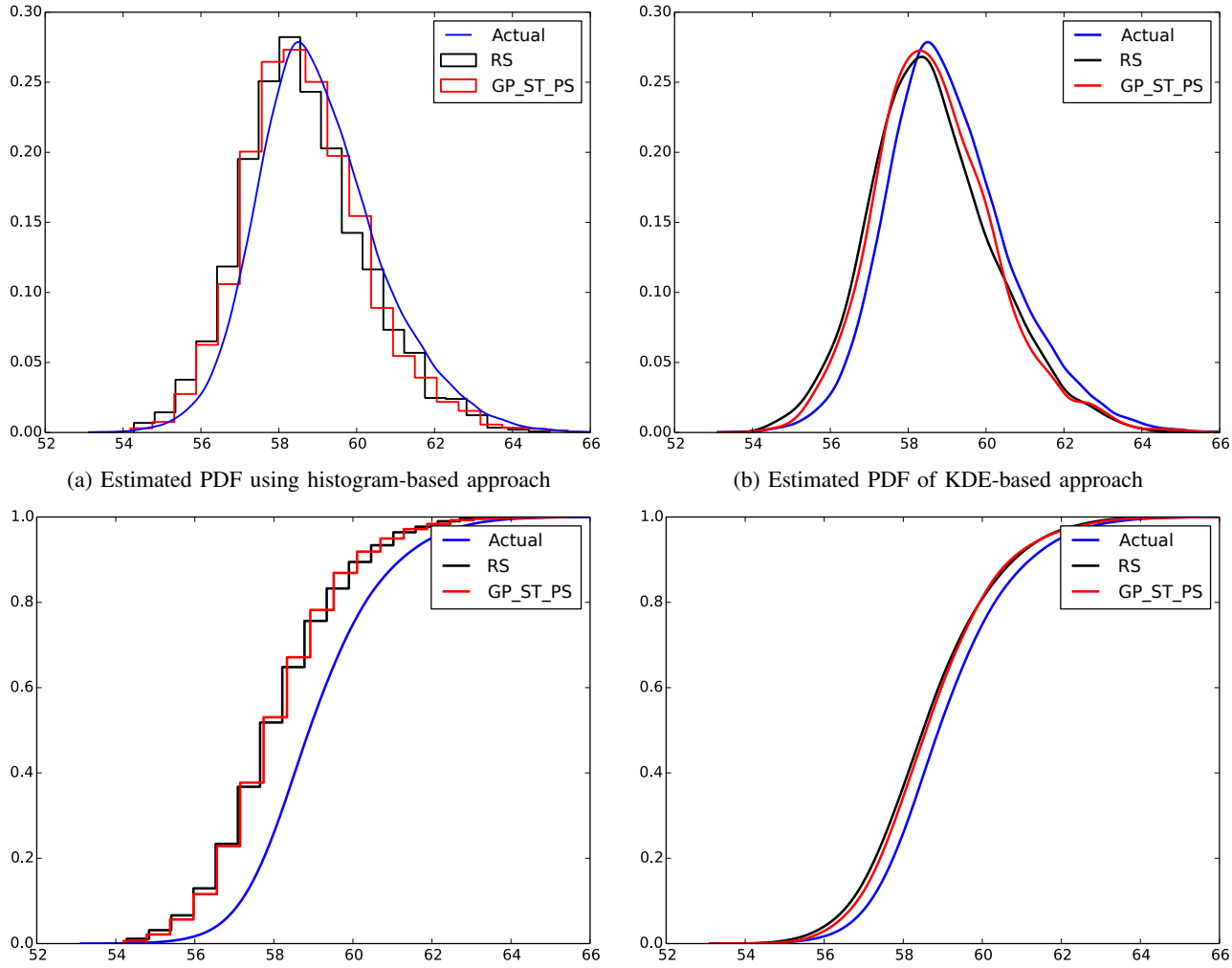
B. HVM Yield Estimation

In the experiments described below, which seek to demonstrate the effectiveness of the proposed method in estimating HVM yield from a few early silicon wafers, we use the first wafer from each of the first $\mathbf{5}$ lots -chronologically- in our dataset, as our early silicon samples. Our sampling budget remains $\mathbf{10}\%$ of the die locations as in the previous experiments. Using these samples, we estimate the probability density function (PDF), the cumulative density function (CDF), and the yield for each of the 39 probe tests in our dataset using the methods described in Section IV. For the entire dataset of the remaining $\mathbf{295}$ wafers, we also calculate the actual density and the actual yield for each of the 39 probe tests. For the purpose of this study, since the specification limits for each probe test were not disclosed to us, the actual yield is computed by setting the lower and upper specification limits at the $[\mathbf{L}, \mathbf{U}] = \pm\mathbf{3}\sigma$ range of each measurement across the 295 HVM wafers. The same limits are also used to predict the yield from the estimated PDF/CDF and compare to the actual yield, so that the accuracy of the proposed methods can be assessed. The actual data and the four estimation methods considered in our experiment are summarized below:

- **Actual:** All die of all HVM wafers (295 wafers) are used to compute the actual yield and density for each of the 39 probe tests.
- **Histogram with Random Sampling (Hist-RS):** Given the 5 early silicon wafers, randomly select 10% of the die from each wafer and create a histogram with 20 uniformly distributed bins. The percentage of sampled die across these 5 wafers that falls in the $[\mathbf{L}, \mathbf{U}]$ range will reflect the yield for each probe test.
- **Histogram with Spatio-temporal GP and Progressive Sampling (Hist-GP-ST-PS):** Given the 5 early silicon wafers, select die locations on every wafer using the progressive sampling algorithm in increments of 2.5%. Ensure that the same die locations are picked on each wafer in the first iteration and guide the selection of subsequent iterations using the prediction confidence level of a spatio-temporal GP model built using all previously selected samples. Once the 10% sample budget is reached (i.e. after 4 iterations) use the final GP-ST-PS model to predict the parameter of interest for the unobserved 90% die across the 5 early wafers. Finally, use both the sampled and the predicted die to create a histogram with 20 uniformly distributed bins. The percentage of all die across these 5 wafers that falls in the $[\mathbf{L}, \mathbf{U}]$ range will reflect the yield for each probe test.
- **KDE with Random Sampling (KDE-RS):** Given the 5 early silicon wafers, randomly select 10% of the die from each wafer and estimate the density of these samples using KDE. Then sample the estimated distribution in order to generate one million synthetic die instances. The percentage of synthetic die instances that falls in the $[\mathbf{L}, \mathbf{U}]$ range will reflect the yield for each probe test.
- **KDE with Spatio-temporal GP and Progressive Sampling (KDE-GP-ST-PS):** Given the 5 early silicon wafers, select die locations on every wafer using the progressive sampling algorithm in increments of 2.5% until the 10% sampling budget is reached, following the same procedure described above for Hist-GP-ST-PS. Then use the final GP-ST-PS model to predict the parameter of interest for the unobserved 90% die across the 5 early wafers. Subsequently, use all sampled and predicted values to estimate the density of the distribution using KDE. Then sample the estimated distribution in order to generate one million synthetic die instances. The percentage of synthetic die instances that falls in the $[\mathbf{L}, \mathbf{U}]$ range will reflect the yield for each probe test.

To evaluate the effectiveness of these methods, we first compare the estimated distributions. Figure 8(a) shows the actual PDF and the estimated PDFs using the Hist-RS and the Hist-GP-ST-PS methods for a randomly selected among the 39 probe-tests (i.e. measurement 24). Similarly, Figure 8(b) shows the actual PDF and the estimated PDFs using the KDE-RS and the KDE-GP-ST-PS methods for the same measurement. Although the differences are subtle, one can still observe that the GP-ST-PS method provides a more accurate sample and results in a better estimation both with the simple histogram method and with the advanced KDE method. As expected, one can also observe that KDE is a very powerful method for estimating the actual distribution. Nevertheless, it still benefits from having a better starting point, as provided by the GP-ST-PS method.

In order to quantitatively compare the quality of estimation, we use the Kolmogorov-Smirnov (KS) test [17] as a goodness-of-fit metric. KS test is a nonparametric test for one-dimensional probability distributions that can be used to compare a sample to a reference. In KS, the comparison metric is the maximum distance between the CDF of the estimated



(a) Estimated PDF using histogram-based approach

(b) Estimated PDF of KDE-based approach

(c) CDF of histogram-based estimation approach

(d) CDF of KDE-based estimation approach

Fig. 8: PDF and CDF estimated by histogram-based and KDE methods for measurement 24

density and the actual CDF. A smaller distance (i.e. closer to 0) indicates a better fit between the real distribution and the estimated one. Figure 8(c) shows the actual CDF and the estimated CDFs using the Hist-RS and the Hist-GP-ST-PS methods for measurement 24, while Figure 8(d) shows the actual CDF and the estimated CDFs using the KDE-RS and the KDE-GP-ST-PS methods for the same measurement. Based on these CDFs, in Table I we compute the KS metric for each of the four predicted CDFs in contrast to the actual. The results corroborate our claim that the information added by GP-ST-PS helps in better estimating the actual distribution, both for the histogram-based and for the KDE-based method.

Finally, we compare the yield estimated by each of the four methods to the actual HVM yield and we compute the corresponding yield estimation error as the absolute difference between the two. Figure 9 shows the yield estimation error for each of the four aforementioned methods for each of the 39 probe tests, as a percentage on the $\mathbf{y-axis}$. Additionally, Table II shows the average yield error over all 39 measurements. Once again, the results confirm the effectiveness of enhancing the initial sample using the GP-ST-PS method towards improving the accuracy of HVM yield estimation.

TABLE I: KS metric of estimated CDFs for measurement 24

	Histogram (RS)	Histogram (GP-ST-PS)	KDE (RS)	KDE (GP-ST-PS)
Distance	0.25	0.23	0.11	0.087

VI. Conclusion

Progressive sampling and inclusion of spatio-temporal features constitute powerful enhancements to the popular Gaussian process wafer-level spatial correlation modeling method. As demonstrated using a comprehensive dataset from an industrial device, their combination leads to much more accurate predictions of parametric measurements from an intelligently assembled sample of die across the wafers of a lot, than the original method. Furthermore, these enhanced models can be used in a post-silicon validation environment to estimate

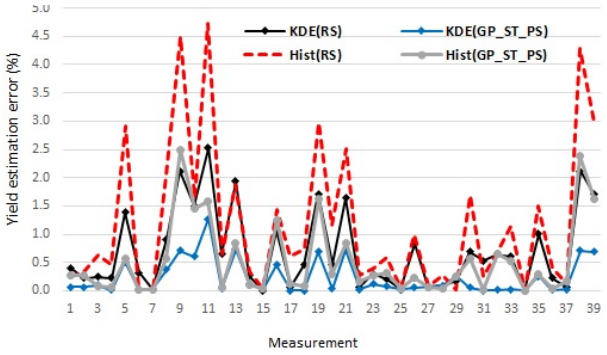


Fig. 9: Yield estimation error for 39 measurements

TABLE II: Average error of HVM yield estimation

	Histogram (RS)	Histogram (GP-ST-PS)	KDE (RS)	KDE (GP-ST-PS)
Yield error	1.16%	0.63%	0.61%	0.21%

the HVM yield of a device from a small number of early silicon wafers. Indeed, as our experiments corroborate, the correlation models constructed through the proposed methods reflect better the overall statistics of HVM production than a random sampling of die from the available early silicon wafers, either in isolation or in conjunction with previously proposed powerful synthetic population enhancement methods.

VII. Acknowledgement

This research has been partially supported by the Semiconductor Research Corporation (SRC) Task 1836.131 and a generous gift by Intel Corporation.

References

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

[2] 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.

[3] W. Zhang, X. Li, T. 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.

[4] 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.

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

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

[7] K. Huang, N. Kupp, J.M. Carulli, and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," in *Design, Automation & Test in Europe Conference*, 2013, pp. 553 – 558.

[8] H.-G. Stratigopoulos, S. Mir, and A. Bounceur, "Evaluation of analog/RF test measurements at the design stage," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 4, pp. 582–590, 2009.

[9] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[10] M.A. Aizerman, E.A. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[12] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., 1998.

[13] F. Liu, E. Acar, and S. Ozev, "Test yield estimation for analog/RF circuits over multiple correlated measurements," in *IEEE International Test Conference*, 2007, pp. 1–10.

[14] S. Sunter and N. Nagi, "Test metrics for analog parametric faults," in *IEEE VLSI Test Symposium*, 1999, pp. 226–234.

[15] R.O. Topaloglu, "Early, accurate and fast yield estimation through monte carlo-alternative probabilistic behavioral analog system simulations," in *IEEE VLSI Test Symposium*, 2006, pp. 137–142.

[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, 1986.

[17] J. Durbin, *Distribution Theory for Tests based on Sample Distribution Function*, vol. 9, SIAM, 1973.