

Low-Cost Analog/RF IC Testing Through Combined Intra- and Inter-Die Correlation Models

Ke Huang

San Diego State University

Nathan Kupp

Yale University

Constantinos Xanthopoulos and Yiorgos Makris

The University of Texas at Dallas

John M. Carulli, Jr.

Texas Instruments

Editor's notes:

Test cost has historically been a significant fraction of overall analog chip cost. This problem is being attacked on two fronts: by exploiting correlations between specification measurements and lower cost alternative tests, and by wafer-level spatial prediction methods. The work in this paper combines the two approaches and demonstrates the effectiveness on commercial analog devices.

—Kenneth M. Butler, Texas Instruments

Toward reducing the excessive cost of specification testing in analog/RF circuits, a statistical intra-die correlation approach, also known as “alternate test,” has been proposed for accurately testing analog/RF devices without explicitly measuring costly performances. The underlying

■ DURING MANUFACTURING OF analog/RF integrated circuits (ICs), every fabricated circuit is tested against its design specifications in order to identify devices that do not function properly due to manufacturing defects or excessive process variations. However, expensive automated test equipment together with lengthy setup and test times for analog/RF ICs result in excessive test costs, which comprise a large percentage of the overall cost of producing these devices. Thus, test cost reduction for analog/RF ICs has been a topic of ongoing interest to the semiconductor manufacturing industry.

principle is to approximate these performances through correlation models based solely on low-cost measurements [1]. Alternatively, as described in [2], a machine-learning-based approach can be employed in order to learn classification boundaries which separate passing and failing populations in a multidimensional space of inexpensive measurements. Along the same line, a defect-based test approach based on On-chip RF Built-in Tests (ORBiTs) was proposed in [3] to reduce test cost in analog/RF devices.

The aforementioned test methods leverage die-level correlations in order to reduce test cost. Specifically, if \mathbf{m}_{alt} denotes the low-cost measurement vector and m_{per} denotes the performances to be predicted, they use a set of training samples to learn the correlation function between \mathbf{m}_{alt} and m_{per} , denoted by f_1 , and thereby to predict performances on

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/MDAT.2014.2361721

Date of publication: 07 October 2014; date of current version:

16 January 2015.

new devices with \mathbf{m}'_{alt} , as $\hat{m}_{\text{per}} = f_1(\mathbf{m}'_{\text{alt}})$. While this approach achieves dramatic test cost reduction, it comes at the cost of increased test escapes (TE) and yield loss (YL).

Recently, the use of statistical correlation toward reducing test cost based on wafer-level spatial correlation modeling has also attracted interest. In this case, costly specification tests are not completely eliminated. Instead, they are only performed on a sparse subset of die on each wafer and, subsequently, used to build a spatial model f_2 , which is then used to predict performances at unobserved die locations: $\hat{m}_{\text{per}} = f_2(\mathbf{x})$, where \mathbf{x} denotes the wafer's Cartesian coordinate $\mathbf{x} = [x, y]$. Along these lines, the expectation–maximization (EM) algorithm was used in [4] to estimate spatial wafer measurements. The virtual probe (VP) approach [5] modeled spatial variation via a discrete cosine transform (DCT), which projects spatial statistics into the frequency domain. Similarly, Liu [6] 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. As recently shown in [7], using such GP models can dramatically improve both prediction accuracy and computational time, as compared to the VP model. In case the spatial variation exhibits discontinuous effects caused by multisite test environments, the approach proposed in [8] can be used to solve the spatial variation discontinuity problem by partitioning the wafer into distinct regions and training spatial correlation models in each individual region.

In this work, we investigate the potential of combining these two statistical approaches, expecting that the performance prediction accuracy of a joint correlation model will surpass the accuracy of its constituents. The proposed methodology, which is introduced in the Proposed Approach section, relies on a combined model for predicting performances, which incorporates the predictive power of both intra-die and inter-die correlations. In addition, we also introduce a screening step in order to verify that each of the constituent correlations truly exists. Indeed, not all performances are adequately predictable by both approaches. In fact, using a poor prediction model in the mix may not only fail to improve the accuracy of the combined model, but it might actually hurt it. Therefore, assessing effectiveness of the constituent models prior to combining

them not only saves computation time but also safeguards the quality of the combined prediction model. The proposed approach is experimentally assessed in the Experimental Results section, using industrial semiconductor manufacturing data from an RF device. The reported results corroborate our expectation that when an RF measurement is deemed predictable by both models, a combined prediction model will significantly improve the prediction accuracy over the constituent models.

Proposed approach

Overview

In this section, we describe in detail the proposed methodology for combining die-level (i.e., alternate test) and wafer-level (i.e., spatial correlation) models for analog/RF test cost reduction. Figure 1 illustrates an overview of the methodology, which consists of two stages, namely training and testing. Let $\hat{m}_{\text{per}1}$ and $\hat{m}_{\text{per}2}$ denote the estimated performance by die-level model f_1 and wafer-level model f_2 , respectively. During the training stage, we first learn f_1 and f_2 as explained in the previous section and as depicted in Figure 1. Then, we assign a weight w_i to the i th model, $i \in \{1, 2\}$, by solving the following optimization problem:

$$\text{minimize}_{\mathbf{w}} \|\mathbf{m} - \mathbf{w} \cdot \mathbf{f}^T\| \quad (1)$$

where $\|\cdot\|$ denotes the $L2$ -norm, \mathbf{w} denotes the weight vector of correlation functions $\mathbf{w} = [w_1, w_2]$, \mathbf{m} denotes the measurement vector of n samples used to assign the weights, and \mathbf{f} denotes the vector of considered correlation models $\mathbf{f} = [f_1, f_2]$. In this work, we propose to solve the optimization problem in (1) using the ordinary least squares (OLS) method to learn the optimal weight vector $\hat{\mathbf{w}}$. Once $\hat{\mathbf{w}}$ is learned, we can readily predict values of performances for untested die locations, as shown in the lower part of Figure 1

$$\hat{m}_{\text{com}} = \mathbf{w} \cdot \mathbf{f}^T \quad (2)$$

where \hat{m}_{com} denotes the predicted performances by the combined model. Thus, in order to obtain the combined estimation on a particular die using (2), we need to perform: 1) low-cost measurements on the same die; and 2) specification tests on a sparse set of other die on the same wafer. The approach shown in (1) and (2) is, in essence, a regression combination approach, which assigns a weight to

each considered model according to its performance, and predicts the final outcome using the estimated weights. Various methods to address this issue have been proposed, including Bayesian model averaging and weighting based on bootstrap or perturbation. However, combining models may not always provide improvement in the prediction accuracy, especially when a subset of models have rather poor performances. To avoid this problem, we adopt the method proposed in [9], which introduces a model screening step. Such screening narrows down the list of candidate models, not only for saving computation time but also for removing poor models that would hurt the combined estimator.

Model screening

As described previously, introducing models with poor performance in the combined estimator could degrade the quality of the final prediction. Thus, it is important to evaluate the performance of each model before the combined estimator is constructed and applied. Yuan and Yang [9] proposed a model screening step, which narrows down the list of candidate models in the combined estimator. In this section, we present the details of model screening for each considered model.

Screening of alternate test model. The alternate test model consists of predicting a high-cost specification test m_{per} using low-cost alternate measurements \mathbf{m}_{alt} . As shown in [10], to assess the effectiveness of the considered model, we can use a holdout set S_1 of n_1 samples containing the specification test and the alternate measurement vector: $S_1 = \{\mathbf{m}_{\text{alt}}^{(i)}, m_{\text{per}}^{(i)}\}, i = 1, \dots, n_1$. Then, we split S_1 into

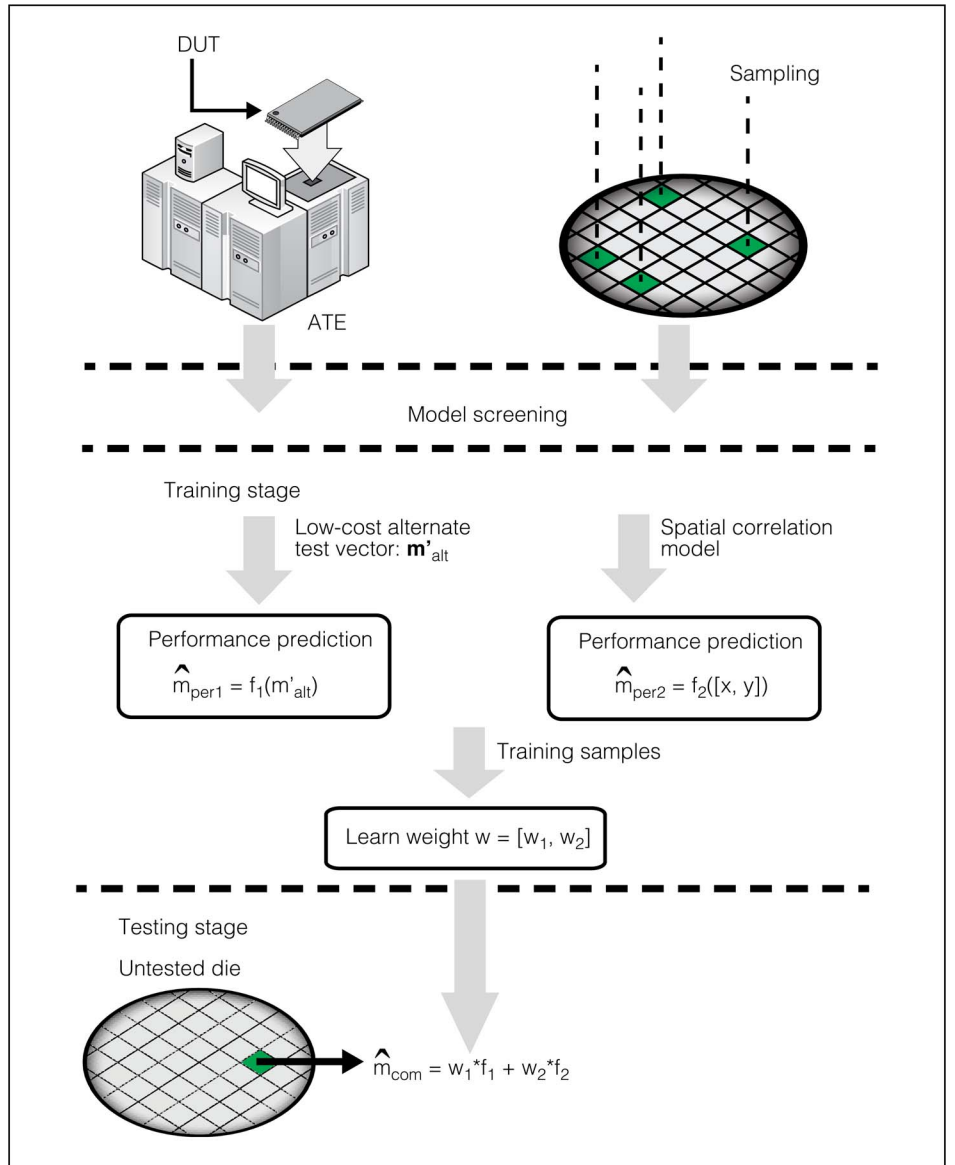


Figure 1. Proposed approach: combined die- and wafer-level correlation models.

two equal sets S_{1t} and S_{1v} uniformly at random. The regression model for alternate test $f_1(\mathbf{m}_{\text{alt}})$ is trained using S_{1t} and validated using S_{1v} . Finally, we compute the normalized root mean square (RMS) error in the validation set S_{1v}

$$\varepsilon_v = \frac{\sqrt{\frac{\sum_{i=1}^{n_1} (\hat{m}_{\text{per}_i} - m_{\text{per}_i})^2}{\frac{n_1}{2}}}}{r_{\text{per}}} \quad (3)$$

where r_{per} denotes the variation range of m_{per} in S_1 , defined as $r_{\text{per}} = \max(m_{\text{per}}) - \min(m_{\text{per}})$, and \hat{m}_{per_i}

is the estimated performance value for the i th device in $S_{1\nu}$. We set a threshold value θ_1 for the considered performance m_{per} , such that the alternate test model is considered to be poor if $\epsilon_v > \theta_1$. Poor models are then screened out in the combined estimator. Note that, in case of process shifts, alternate test model screening can be repeated periodically to monitor the fitness of the model. An n -fold cross validation can also be employed to accurately assess the error incurred by applying the alternate test model and to select θ_1 for a target error.

Screening of spatial model. The underlying question of assessing the spatial model is whether a systematic spatial correlation exists, or whether it is dominated by random noise. In the latter case, the performance cannot be predicted by a spatial correlation model. In [10], existence of systematic spatial correlation is assessed by computing the Pearson's correlation coefficient between two adjacent wafers, assuming that systematic spatial patterns remain similar across adjacently manufactured wafers. However, this assumption may not hold true in the presence of outlier wafers or process shifts. Figure 2 shows two wafer maps of a measurement, chosen from the same manufacturing line. As may be observed, the systematic spatial variation may exhibit a radial shift across wafers. Even though it is still present, assessing Pearson's correlation coefficient between these wafers may lead to an erroneous decision that no spatial correlation exists on the wafer.

In order to provide a model fitness assessment on a per-wafer basis, we propose herein an alternative method for verifying existence of systematic spatial patterns. This method, which relies only on the training samples used to learn the spatial correlation model on each wafer, resembles and is inspired by the location averaging method proposed in [11] for

screening defective die on a wafer. Since the model fitness assessment is performed on a per-wafer basis, this method is robust to process shifts and outlier wafers. In particular, for a given wafer to be tested, let the set $S_2 = \{(m_{\text{per}}|\mathbf{x})_1, \dots, (m_{\text{per}}|\mathbf{x})_\nu\}$ denote the samples used to train the spatial correlation model, where ν is the number of training samples and \mathbf{x} is the vector of die coordinates. Also, let $S_i' = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_n}\}$ denote the neighboring die locations of the i th die sample in S_2 , where N_n is the total number of available die locations near the i th die sample.

We train the spatial correlation model f_2 using samples in S_2 , and use f_2 to predict the values in S_i' for $i = 1, \dots, \nu$. Finally, we compute the average of predicted values of the die locations near the i th die sample in S_2

$$\tilde{m}_{\text{per}_i} = \frac{1}{N_n} \sum_{j=1}^{N_n} \hat{m}_j \quad (4)$$

where \hat{m}_{per_i} denotes the average of predicted values of the die locations near the i th die sample in S_2 and \hat{m}_j denotes the predicted j th neighborhood value. Once the average value is calculated, we can then compute the difference between the averaged location value and the actual value

$$\epsilon_i = m_{\text{per}_i} - \hat{m}_{\text{per}_i} \quad (5)$$

where ϵ_i denotes the difference between the averaged location value and the actual value of the i th die sample in S_2 . Finally, we compute the averaged error across all samples in S_2

$$\epsilon_s = \frac{1}{\nu} \sum_{j=1}^{\nu} \epsilon_j. \quad (6)$$

Similarly to alternate test screening, we can set a threshold value θ_2 for the considered i th measurement m_{per_i} , such that the spatial correlation model is deemed to be poor if $\epsilon_s > \theta_2$. Figure 3a and b shows examples of spatial correlation model fitness assessment for a nonspatially correlated measurement and a spatially correlated measurement. It can be readily observed that the nonspatially correlated measurement has predicted neighborhood values which significantly differ from the sample value, as shown in Figure 3a. This difference will result in a large ϵ_i value in (5). On the other hand, the spatially correlated measurement has neighborhood values which are very close to the sample value, as shown

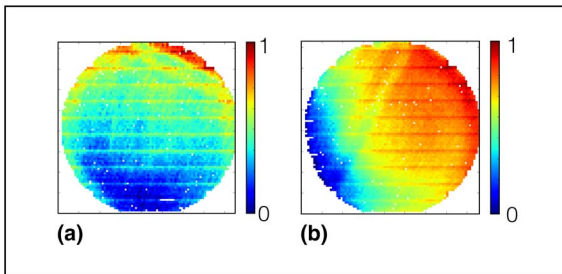


Figure 2. Shift of systematic spatial patterns across wafers.

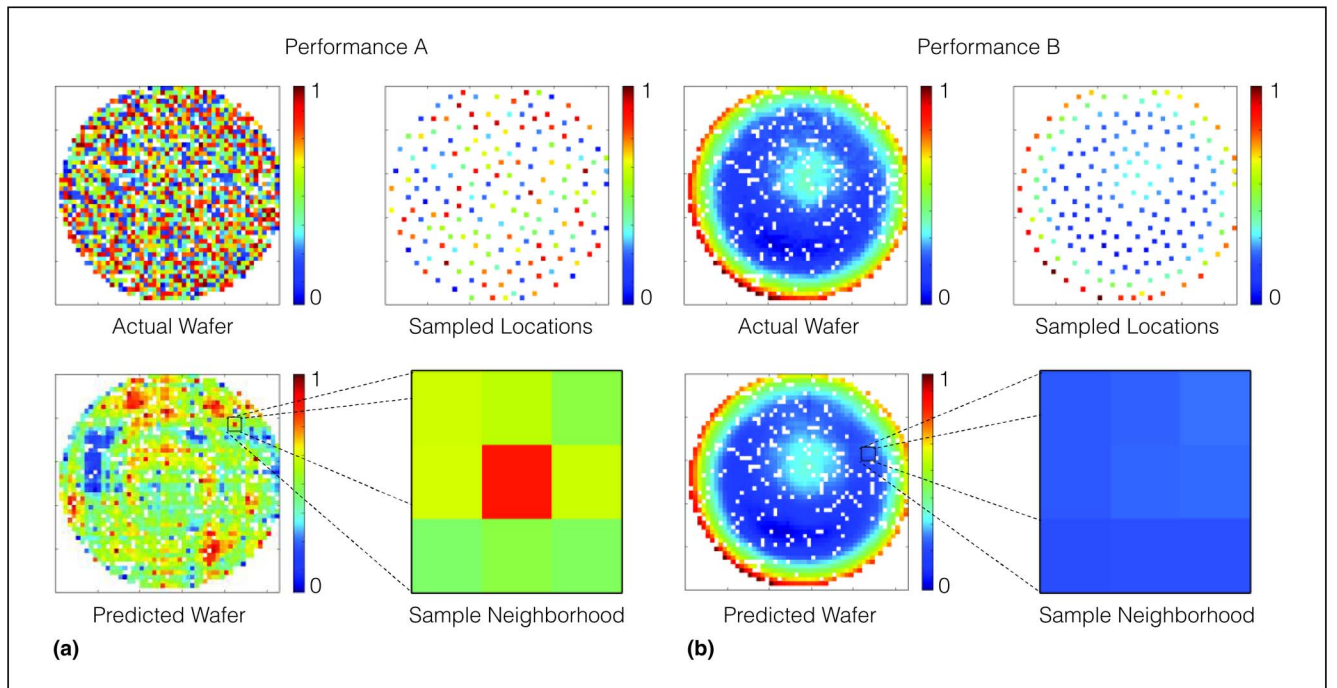


Figure 3. (a) Nonspatially correlated and (b) spatially correlated performance measurements.

in Figure 3b. Poor models are then screened out in the combined estimator as before.

We note that θ_1 and θ_2 can be chosen based on the acceptable levels of TE and YL, as we further discuss in the Experimental Results section. If a lower defective parts per million (dppm) value is required, θ_1 and θ_2 will be set lower accordingly to screen out more poor predictors, in order to provide the desired prediction accuracy. We can also use an n -fold cross-validation method to determine the values of θ_1 and θ_2 . Finally, we note that the combined estimator, as well as the model screening approach, is independent of the underlying models; in other words, the proposed approach can be applied to combine any alternate test and spatial correlation models.

Experimental results

The proposed method is evaluated on high-volume manufacturing test data from an RF transceiver built in 65-nm technology. Our data set has a total of 291 wafers, each of which has approximately 2000 dies characterized by 224 low-cost ORBiTs obtained via on-chip sensors, and 101 high-cost RF specification tests. Both low-cost and RF tests are performed at probe level. Thus, a statistically significant data set of approximately 582 000 devices, each with 325 measurements, is used in our case study.

Prediction with alternate test model

As described previously, introducing models with poor performance in the combined estimator could degrade the quality of the final prediction. Thus, it is important to evaluate the performance of each model before the combined estimator is constructed and applied. Yuan and Yang [9] proposed a model screening step, which narrows down the list of candidate models in the combined estimator. In this section, we present the details of model screening for each considered model. In this work, we use least angle regression (LARS) [12] to construct alternate test models for each considered specification: $\hat{m}_{\text{per}} = f_1(\mathbf{m}_{\text{alt}})$. The LARS regression model automatically chooses a subset of variables in vector \mathbf{m}_{alt} which are most correlated with m_{per} in order to build the regression model f_1 . Thus, this approach allows us to handle high-dimensional data in building the regression model without needing to perform a feature selection or dimensionality reduction analysis, which is very appropriate in our case since the dimension of \mathbf{m}_{alt} is 224. Details of the LARS regression model can be found in [12].

To perform screening of the alternate test models, as described in the Screening of Alternate Test Model section, we generate the following data set: we choose the first wafer in our data set as the holdout

set S_1 in order to assess the correlation between \mathbf{m}_{alt} and m_{per} : $S_1 = \{\mathbf{m}_{\text{alt}}^{(i)}, m_{\text{per}}^{(i)}\}$, $i = 1, \dots, n_1$, where $n_1 \approx 2000$. Then, S_1 is split into training set S_{1t} and validation set S_{1v} for the purpose of model screening. We choose $\theta_1 = 12\%$ such that the TE and YL are approximately 1000 ppm across our data set. The prediction error ϵ_v for each of the performances in S_1 is computed using (3). As a result of applying alternate model screening, 52 RF specification tests have ϵ_v smaller than θ_1 in our study. These measurements successfully pass model screening and are, therefore, forwarded to the combined model.

Prediction with spatial model

In this work, wafer-level spatial correlation models are built using the GP-based approach described in [7] with a 10% die sample size. Note that, in manufacturing, the possible sample sites are affected by the layout of the probe card in multisite environments, and the initial samples chosen to learn the spatial correlation models may not be available in each site. This issue can be solved by choosing samples in each individual site through domain-specific knowledge. In other words, we first define different sites on the wafer from domain-specific knowledge about multisite environments, and then we randomly select samples in each site to train the spatial correlation models on the whole wafer. The density at which we apply measurements on each wafer controls the test cost and test error, generating a tradeoff curve between the two. The optimal number of samples is chosen according to the test cost reduction plan and the maximum test error that can be tolerated. In screening out poor spatial models (i.e., performances

that are not spatially correlated), the procedure described in the Screening of Spatial Model section is employed. Once again, we choose $\theta_2 = 8\%$ such that the TE and YL are approximately 1000 ppm across our data set. As a result of applying alternate model screening, 46 RF specification tests have ϵ_s smaller than θ_2 in our study. These measurements successfully pass model screening and are, therefore, forwarded to the combined model as before.

Prediction with combined model

The final combined model is used to predict performances for which both alternate test and spatial correlation models pass the respective screening step. In our experiment, the intersection of the two sets (of cardinality 52 and 46, respectively) contains 34 performances. For each of them, we use, again, the first wafer as the holdout set in order to assign the weight w_l for the l th prediction model computed by (1), and (2) to predict the performances for the unobserved die locations on other wafers. For new wafers to be tested, low-cost alternate tests are taken on each die location, while specification tests are performed on 10% of the available die locations randomly sampled across the wafer, in order to train the spatial correlation model using GP as described in the Introduction. Figure 4 shows the RMS prediction error sorted in ascending order using all three methods, obtained by predicting on the remaining 90% die locations for all wafers. As may be observed, the combined model consistently outperforms, or at least performs equally well as the best of the individual models. Figure 5a and b shows the prediction plots of measurements 85 and 86 for the same devices in the validation set, using all three models. As may be observed by simple visual inspection, the combined model indeed provides the best prediction results.

The second and fifth columns of Table 1 also show this comparison in terms of RMS prediction error. Once again, it can be observed that the combined model consistently provides the best prediction results, which justifies our choice to combine the two predictive models.

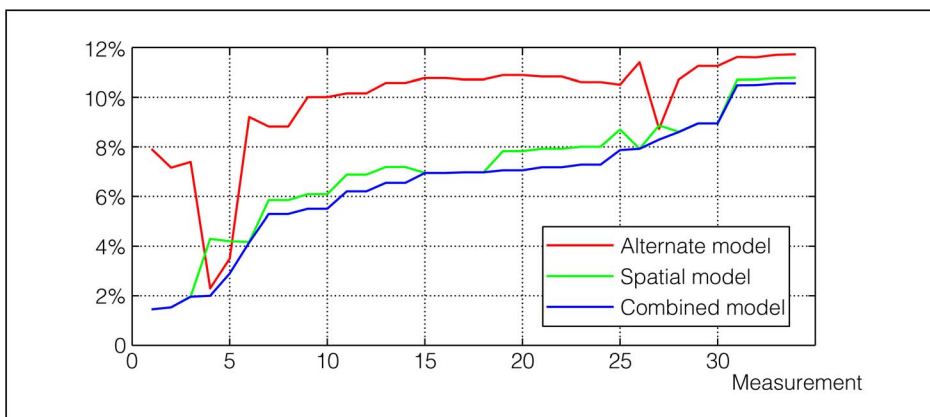


Figure 4. Comparison of RMS prediction error for measurements of passing screening of both models.

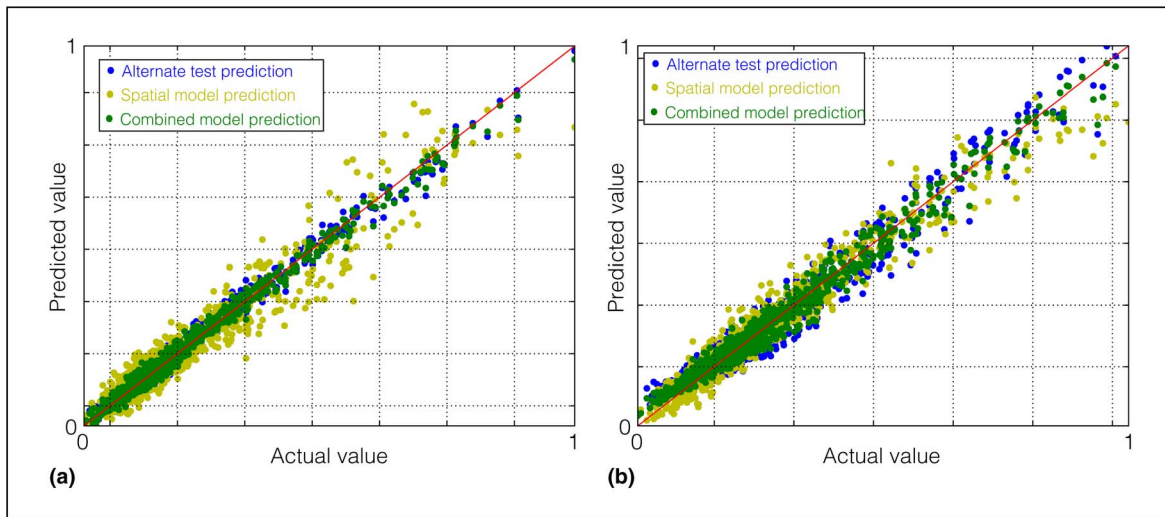


Figure 5. Prediction plot by different models for (a) measurement 85, and (b) measurement 86.

Test escapes and yield loss improvement

To further elucidate the benefits of our approach, it is also worthwhile to compute the TE and YL incurred by applying the predictive model. For a particular measurement, let the indicator functions $I_1^{(i)}/I_2^{(i)}$ be equal to “1” if the predicted value of the i th die location passes/fails its specification, while the actual value fails/passes the specification, and let $I_1^{(i)}/I_2^{(i)}$ be equal to “0” otherwise. Then, the overall TE and YL are defined as

$$\widehat{\text{TE}} = \frac{1}{N} \sum_{i=1}^N I_1^{(i)} \quad (7)$$

$$\widehat{\text{YL}} = \frac{1}{N} \sum_{i=1}^N I_2^{(i)} \quad (8)$$

where N denotes the total number of predicted die locations across all wafers. In the third, fourth, sixth, and seventh columns of Table 1, we compare TE and YL of measurements 85 and 86. Evidently, the proposed approach achieves a significant TE and YL improvement as compared to the individual models, thereby justifying the use of a combined model in predicting high cost performances.

Discussion

One might point out that the proposed approach incurs additional cost in predicting performances, as compared to traditional alternate test, since it requires that specification tests are performed on a subset of die on each wafer. In reality, however, even

in a traditional alternate test setting, a certain level of specification tests are also needed to ensure integrity of the models in the presence of process shifts, to deal with die for which performances cannot be predicted with high confidence (two-tier test [2]), and to monitor the process. With this in mind, and taking into account the achieved TE and YL reduction, which might give the extra nudge needed to meet test quality goals, we believe that the added cost is justifiable from a test economics point of view.

COMBINING ALTERNATE TEST with spatial correlation modeling holds great promise in further reducing test cost in analog/RF ICs without compromising test quality. Such merging, however, needs to be carefully orchestrated, taking into account that poorly performing constituents may jeopardize the effectiveness of the joint predictor and should, therefore, be screened out. Experimental results using test data from high-volume manufacturing assert that the joint prediction model proposed herein achieves lower RMS prediction error and, by extension, reduces TE and YL.

Table 1 Prediction outcome for measurements 85 and 86.

	Measurement 85			Measurement 86		
	RMS	TE	YL	RMS	TE	YL
Alternate model	2.3%	515	115	3.5%	430	83
Spatial model	4.3%	589	347	4.2%	563	334
Combined model	2%	401	68	2.9%	343	54

■ References

- [1] P. N. Variyam, S. Cherubal, and A. Chatterjee, "Prediction of analog performance parameters using fast transient testing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 3, pp. 349–361, Mar. 2002.
- [2] H.-G. Stratigopoulos and Y. Makris, "Error moderation in low-cost machine-learning-based analog/RF testing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 2, pp. 339–351, Feb. 2008.
- [3] D. Mannath et al., "Structural approach for built-in tests in RF devices," in *Proc. Int. Test Conf.*, 2010, DOI: 10.1109/TEST.2010.5699241.
- [4] S. Reda and S. R. Nassif, "Accurate spatial estimation and decomposition techniques for variability characterization," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 3, pp. 345–357, Aug. 2010.
- [5] W. Zhang et al., "Virtual probe: A statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 12, pp. 1814–1827, Dec. 2011.
- [6] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proc. Design Autom. Conf.*, 2007, pp. 817–822.
- [7] N. Kupp, K. Huang, J. Carulli, and Y. Makris, "Spatial correlation modeling for probe test cost reduction," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, 2012, pp. 23–29.
- [8] K. Huang, N. Kupp, J. Carulli, and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," in *Proc. Design Autom. Test Eur. Conf.*, 2013, pp. 553–558.
- [9] Z. Yuan and Y. Yang, "Combining linear regression models: When and how?" *J. Amer. Stat. Assoc.*, vol. 100, no. 472, pp. 1202–1214, 2005.
- [10] K. Huang, N. Kupp, J. Carulli, and Y. Makris, "On combining alternate test with spatial correlation modeling in analog/RF ICs," in *Proc. IEEE Eur. Test Symp.*, 2013, DOI: 10.1109/ETS.2013.6569358.
- [11] W. R. Daasch, J. McNames, R. Madge, and K. Cota, "Neighborhood selection for I_{DDQ} outlier screening at wafer sort," *IEEE Design Test Comput.*, vol. 19, no. 5, pp. 74–81, Sep./Oct. 2002.
- [12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

Ke Huang is an Assistant Professor in the Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA, USA. His research interests include application of data mining and machine learning in analog/RF testing and hardware security. Huang has a PhD in electrical engineering from the University of Grenoble, Grenoble, France. He is a member of the IEEE.

Nathan Kupp has a PhD in electrical engineering from Yale University, New Haven, CT, USA. His research interests include the application of machine learning and statistical learning theory to problems in semiconductor manufacturing and test. He is a member of the IEEE.

Constantinos Xanthopoulos is currently working toward a PhD in computer engineering at The University of Texas at Dallas, Richardson, TX, USA. His research interests focus on the application of statistical learning theory to problems in analog and RF test. He is a student member of the IEEE.

Yiorgos Makris is a Professor of Electrical Engineering at The University of Texas at Dallas, Richardson, TX, USA. His main research interests lie in the application of machine learning and statistical analysis toward developing reliable and trusted integrated circuits, with particular emphasis in the analog/RF domain. Makris has a PhD in computer science and engineering from the University of California, San Diego, La Jolla, CA, USA. He is a senior member of the IEEE.

John M. Carulli, Jr. is a Distinguished Member of the Technical Staff at Texas Instruments, Dallas, TX, USA. His research interests include data mining, adaptive test methods, product reliability modeling, performance modeling, and security. Carulli has an MSEE from the University of Vermont, Burlington, VT, USA. He is a senior member of the IEEE.

■ Direct questions and comments about this article to Yiorgos Makris, The University of Texas at Dallas, Richardson, TX 75080 USA; yiorgos.makris@utdallas.edu.