Modeling and Forecasting of Defect-Limited Yield in Semiconductor Manufacturing

Michael Baron, Asya Takken, Emmanuel Yashchin, and Mary Lanzerotti, *Senior Member, IEEE*

Abstract—A detailed cause-and-effect stochastic model is developed to relate the type, size, location, and frequency of observed defects to the final yield in IC manufacturing. The model is estimated on real data sets with a large portion of unclassified defects and uninspected layers, and in presence of clustering of defects. Results of this analysis are used for evaluating kill ratios and effects of different factors, identifying the most dangerous cases and the most probable causes of failures, forecasting the yield, and designing optimal yield-enhancement strategies.

Index Terms—Clustered defects, defect characteristics, diagnostics, EM algorithm, incomplete defect data, kill ratio, multilayer structures, process characterization, wafer inspection, yield estimation.

I. INTRODUCTION

THE main objective of this paper is to build a cause-and-effect model explaining the patterns of failing chips in terms of observable defects. Fitting such a model to training data sets allows further factorial analysis, such as: estimation and comparison of effects of different factors, detection of influential cases and the most probable causes of failures, etc. Given a detailed estimated model, forecasting of the yield at any time during the manufacturing cycle becomes straightforward, and also, accurate yield predictions can be made for future modifications of the production process, resulting in the optimal choice of yield-enhancing strategies.

A number of models for failing chips on a wafer have been proposed, concentrating on modeling the total number of failures ([7], [8], [18]–[20], [22], [26], and others), spatial dependence of failing chips on a wafer [2], [6], [9], [11], [23], modeling the yield per each produced layer [24], [25], modeling the yield based on critical area summary curves [5], [12], [15], [24], and defect counts per each defect type [13], [14], [16], [17].

Comparing with its predecessors, the model proposed here incorporates detailed information on the observed defects, in order to predict and explain the final yield. Defect types (codes), sizes, frequencies, and locations (layers or operations) play the role of covariates. The study involving nearly 1 000 lots and

millions of chips of different grades and designs showed significance of each mentioned factor. Further, the standard chi-square analysis of the log-likelihood showed significance of *interaction terms* between the defect type and the layer where the defect occurred. To avoid over-parameterization of the model, similarly composed layers were combined into groups, and only interactions between defect types and groups of layers were included.

As a result, the proposed model contains *a large number of parameters*: effect of each defect type and effect of each layer, interactions, defect frequencies, and also, effects of other causes. Apparently, some failed chips had no defects on any of their layers. Such chips were killed by causes other than observable defects. The corresponding effect is lot-specific because all wafers in a lot receive a similar treatment (unless it is decided to split a lot).

For the efficient computation of multi-dimensional parameter estimates, an *expectation-maximization (EM) algorithm* with some modifications is used. Designed to handle large-dimensional estimation problems, the EM algorithm [10], [27] is an iterative numerical estimation routine. During each cycle of this algorithm, a small portion of parameters is estimated by a maximum-likelihood method (M-step). The remaining parameters are then estimated by the corresponding conditional expectations, updating each time the estimates obtained during the previous iteration (E-step).

The introduced modification of the EM algorithm is essentially an extra step during each cycle that accelerates the numerical routine and prevents its convergence to possible local extrema. As shown in [1], the new step can only improve the algorithm's performance.

It is important to notice that a bulk of information gathered on chips is always missing. For reasons of economy, the vast majority of detected defects is unclassified. In addition, only selected layers are inspected on each wafer. No usable information is available for the remaining uninspected layers, although they may certainly contain fatal defects, that is, chip killers. Nevertheless, a carefully applied formula of total probability allows to include all the collected pieces of information (for example, sizes and locations of unclassified defects) into the likelihood.

As mentioned in [17], due to extremely complex designs and delicate technology, the defect information gathered on chips is not perfectly clean and not perfectly reliable. Realizing this issue, the estimation routine is accompanied by the *diagnostics* module aimed to assess the goodness of fit and to detect probable outliers and influential single chips and whole wafers.

After cycles of data cleaning, parameter estimation, and diagnostics, one obtains a set of parameter estimates that explains

---

the impact of various factors on the final yield. The estimated model is then used:

— to forecast the yield at any time during the production process;

— to evaluate *kill ratios*, that is, probabilities for defects of certain types on certain layers to be chip killers;

— to compute and compare effects and risks associated with each defect type and each layer;

— to identify the most dangerous combinations of defect type, size, and location;

— to find, on low-yield wafers, the most probable sources of chip killers and the most probable causes of failures;

— to compare the influence of layers;

— to evaluate significance of other causes, besides the visible defects;

— to predict results of any yield-improving modification of the manufacturing process, in terms of the expected change in the final yield;

— to determine optimal strategies to increase the yield.

Some of the problems outlined above (e.g., yield forecasting and estimation of kill ratios) can be addressed by taking advantage of the critical area computations [21] that are provided by a number of vendors. Such computations are typically based on simulated defects and they do not require any defect data, provided one has information about the defect size distributions. Information of this type can be obtained either from theoretical considerations or empirically; for example, histogram density estimates of the size distributions were used in our study.

Such estimation is especially useful in the product design phase, when very little data is typically available. In this phase, one can typically also take advantage of the knowledge gained from data analyses corresponding to other (for example, upstream) products. In the manufacturing phase, however, one can take full advantage of the information contained in the vast data streams coming from the tools to obtain data driven estimates of the quantities of interest. In this process, one can still incorporate knowledge learned from other products, for example by imposing appropriate model constraints, as illustrated in the next section. Among other things, the data-driven approach presents an opportunity to 1) validate practical relevance of the assumptions made in the critical area computations and 2) gain better understanding of differences between observed yield trends of various products that are not directly explained by critical area comparisons.

The paper is organized as follows. The stochastic model relating observable defects and chip failures is developed in Section II. Parameter estimation and model diagnostics tools are presented in Section III. Development of the proposed methods for yield forecasting and yield improvement is discussed in Section IV. Section V describes a large case study where the proposed methodology is applied to a certain type of wafers produced at IBM. Similar results and a similar prediction power were obtained for many other lots. In a few exceptions, where the model failed to provide a reasonable yield forecast, we were always able to find the assignable cause. Results and conclusions are summarized in Section VI.

II. A CAUSE-AND-EFFECT RELATION BETWEEN DEFECTS AND FAILURES

We start building the likelihood from a single defect. Suppose a defect of type j and size s occurred on layer l of chip i. What is the probability for the chip to survive this defect? A number of competing models is proposed and compared in [1]. According to our experiments, the model with

$$P\{\text{a chip survives}\} = \exp\{-r(j)a(l)g(s)\} \qquad (1)$$

where $r(j)$, $a(l)$, and $g(s)$ are effects of the defect type, layer, and size, dominated and provided the best fit and the most accurate forecasts. Among others, the function $x = g(s) = \log(1 + s)$ provided the best fit. Hence, we use it here and below.

It should generally be recommended to have *a bank of plausible models* that can be compared for each new mode of production, so that the model with the best fit can always be chosen. The methodology described below can be used for each proposed model, and only minor changes in the computer code will be necessary.

Further, the *interaction terms*, or the *joint effects* of a defect type and a group of layers, were found significant and significantly improving the fit, based on a large number of lots spanning several types of production. Layers are grouped by similarity into brightfield, darkfield, light metal, and other groups. Such a grouping of layers appears sufficient to improve the accuracy of yield prediction and the overall fit, at the same time avoiding undesired over-parameterization. In the sequel, we will not change the introduced notation and will let the index j run through the set of observed pairs of defect types and groups of layers. Only a portion of such pairs actually occurs.

Typically, an experienced engineer will be able to generate the structure of the model (including interactions) with minimal effort. In this process, one can frequently take advantage of the critical area computations; one immediate benefit would be confirming the relevance of a particular function $g(s)$ chosen for (1). Furthermore, one could impose an additional structure to accommodate various types of products: for example, based on the critical area analysis one may impose a restriction that the layer effects for products A and B should satisfy a relationship $a(A) = v + a(B)$, where some of the components of the vector v are presumed known.

First, consider an idealized situation where all layers are inspected on all wafers, and all the detected defects are **classified**. Then, assuming independence of effects, as in [24] and [25], and including the effect of other causes $b(m)$ for lot m, the survival probability for chip i of lot m, $m = 1, \ldots, M$, is

$$\phi_i = P\{\text{chip } i \text{ survives all its defects}\}$$
$$= e^{-b(m)} \prod_{l=1}^{L} \prod_{j=1}^{J} \prod_{k \in C_{ijl}} e^{-r(j)a(l)x_k}. \qquad (2)$$

Here, J is the number of known defect types, L is the number of layers, and k runs through the set C_{ijl} of all defects on layer l of chip i that are classified to defect type j. For any defect k, $x_k = \log(1 + s_k)$ is the chosen transformation of the defect

size s_k. Similarly, the type of this defect and the layer on which it occurred will be denoted by j_k and l_k.

Next, let ξ_i equal 1 for a functioning and 0 for a failing chip i. Under the independence of ξ_i, which only means that each chip failure is caused by its own defects or other causes, but not by the condition of other chips, the following likelihood is constructed,

$$\mathcal{L}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{r}) = \prod_{m=1}^{M} \prod_{w\in m} \prod_{i\in w} \phi_i^{\xi_i}(1-\phi_i)^{1-\xi_i}. \qquad (3)$$

Notation $w \in m$ is used to mark all wafers w belonging to lot m. Similarly, $i \in w$ means a chips i that lies on wafer w.

However, the situation is different due to a large number of **unclassified defects**. For such defects, only the size s and location (layer or operation) l are known, but the defect type is unknown. By the formula of total probability, a chip survives an unclassified defect k of (transformed) size x_k with probability

$$P\{\xi = 1|x_k\} = \sum_{j=1}^{J} P\{j_k = j|x_k\}P\{\xi = 1|j_k = j, x_k\} \qquad (4)$$

Computing (4), the conditional survival probabilities, given $j_k = j$, are obtained from (1), and probabilities of different defect types are computed using the Bayes rule,

$$P\{j_k = j|x_k = x\} = \frac{P\{j_k = j\}\pi(x|j_k = j)}{\sum_{j'} P\{j_k = j'\}\pi(x|j_k = j')}$$

where

$$P(j_k = j) = \frac{\lambda(j, l_k, m_k)}{\lambda(l_k, m_k)}$$

is the corresponding proportion of defect frequencies, with $\lambda(l_k, m_k) = \sum_j \lambda(j, l_k, m_k)$, being the expected number of defects per chip on level l_k of lot m_k, and

$$\pi_j(x) = \pi(x|j_k = j)$$

is the distribution of (transformed) defect sizes which differs from one defect type to another.

Given a large database of classified defects, we estimate the distributions π_j nonparametrically by computing histogram density estimates ([4]); however, a parametric approach is, in principle, also possible. Dependence of the size distribution on the defect type was evident.

As a result, the probability of surviving an unclassified defect is now expressed as

$$P\{\xi = 1|x_k\} = \frac{\sum_j \lambda(j, l_k, m_k)\pi_j(x_k)e^{-r(j)a(l_k)x_k}}{\sum_j \lambda(j, l_k, m_k)\pi_j(x_k)}. \qquad (5)$$

The other source of missing information relates to the practice of selective inspection schemes that leave a large number of

uninspected layers. Most of the wafers have only a small fraction of their layers inspected. At the same time, an uninspected layer may contain fatal defects that cause the chip failures and affect the final yield.

The effect of uninspected layers can also be included into the likelihood through the formula of total probability. Since all the information on such layers is hidden, expectations should be taken over the number of defects of each type as well as their sizes. A Poisson $(\lambda(j, l, m))$ number of defects N_{ijlm} of type j on layer l of chip i of lot m is assumed. Then,

$$P\{\xi = 1|\text{uninspected layer}l\}$$
$$= \prod_j P\{\text{all type } j \text{ defects on layer } l \text{ are not fatal}\}$$
$$= \prod_j \mathbf{E}^{N_{ijlm}}(P\{\text{a defect of type } j \text{ is not fatal}\})^{N_{ijlm}}$$
$$= \prod_j \exp\{-\lambda(j, l, m)(1 - \psi_{jl})\} \qquad (6)$$

where the expectation $\mathbf{E}^{N_{ijlm}}$ is taken with respect to the number of type j defects on level l of chip i of lot m, and

$$\psi_{jl} = \mathbf{E}_j^x e^{-r(j)a(l)x} = \int e^{-r(j)a(l)t} d\pi_j(t) \qquad (7)$$

is the moment generating function of size x, by defect type and layer.

Notice that (7) is the probability of surviving a defect j (of a random size) on layer l. Thus, $(1 - \psi_{jl})$ represents **the kill ratio**, the probability for such a defect to kill a chip, which is the quantity of primary interest to practitioners.

As we notice, probabilities (5) and (6) of surviving unclassified defects and uninspected layers contain unknown defect frequencies $\boldsymbol{\lambda} = \{\lambda(j, l, m)\}$ that are now included as parameters into the overall likelihood. The latter is now expressed as a product of the likelihood of all the observed defects, the conditional likelihood of classified defects given all the observed defects, and the conditional likelihood of chip failures given all classified and unclassified observed defects. For simplicity, we write events "occurrence of the observed (classified, unclassified) defects" and "occurrence of the observed failures" briefly as "defects" and "failures."

As a result, the overall likelihood is now expressed as a product of three components:

$$\mathcal{L}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{r},\boldsymbol{\lambda})$$
$$= \mathcal{L}\{\text{defects}\}\,\mathcal{L}\{\text{classified defects}|\text{defects}\}$$
$$\times \mathcal{L}\{\text{failures}|\text{defects}\}$$
$$= \prod_m \prod_{w\in m} \prod_{l\in L_w} \prod_{i\in w} e^{-\lambda(l,m)}\frac{\lambda^{N_{il}}(l,m)}{N_{il}!}$$
$$\times \prod_j \left(\frac{\lambda(j, l, m)}{\lambda(l, m)}\right)^{d_{ijl}} \times \phi_i^{\xi_i}(1-\phi_i)^{1-\xi_i} \qquad (8)$$

where $N_{il} = \sum_j N_{ijlm}$ is the number of all defects on layer l of chip i (of lot m), and $d_{ijl} = |C_{ijl}|$ is the number of classified type j defects on layer l of chip i.

In (8), ϕ_i is the chip i survival probability, updated from the "idealized" version (2), so that

$$\log \phi_i = -b(m) - \sum_{l\in L_w}\sum_{k\in C_{ijl}} r(j_k)a(l)x_k$$
$$+ \sum_{l\in L_w}\sum_{k\in U_{il}} \log \frac{\sum_j \lambda(j,l)\pi_j(x_k)e^{-r(j)a(l)x_k}}{\sum_j \lambda(j,l)\pi_j(x_k)}$$
$$- \sum_{l\notin L_w}\sum_j \lambda(j,l)(1 - \psi_{jl}) \qquad (9)$$

where U_{il} denotes the set of unclassified defects on layer l of chip i, and L_w is the set of inspected layers on wafer w.

This survival probability consists of four parts representing four sources of failing chips. It reflects the fact that in order to function, a chip needs to survive non-defect causes (the first term), all classified defects on it (second term), all unclassified defects (the third term), and all the uninspected layers (fourth term). Then, (9) represents a detailed *cause-and-effect relation* between defects and chip failures.

III. MODEL FITTING: ESTIMATION, GOODNESS OF FIT, AND DIAGNOSTICS

This section proposes the parameter estimation and model adequacy evaluation routines. In practice, one would apply this scheme to most recent training data sets, first, to update the parameter estimates that are used for effect comparison and yield prediction, and second, to test whether the chosen model continues to be adequate for the current production.

A. Estimation

Given the explicit form of the likelihood (8), it is the first impression that maximum likelihood estimation is natural and straightforward. However, the problem has a very high dimension. Indeed, besides tens of defect type effects $r(j)$, tens of layer effects $a(l)$, and hundreds of lot-specific other causes effects $b(m)$, one has to estimate the defect frequencies $\lambda(j, l, m)$ for $j = 1, \ldots, J, l = 1, \ldots, L, m = 1, \ldots, M$. Because of the latter, the total number of parameters often approaches 100 000, immediately making the "brute-force" optimization of the likelihood computationally infeasible.

On a side note, let us mention that estimation of defect frequencies is the problem of its own keen interest. Reduction of the frequencies $\lambda(j, l)$ of the most dangerous defects is a viable yield increasing strategy. On the other hand, a strategy directed towards elimination of the most dangerous defects may not be yield efficient if these defects have negligibly low frequencies. For this reason, in addition to kill ratios, practitioners often consider *weighted defect densities*

$$\lambda(j, l)P\{\text{defect of type } j \text{ on layer } l \text{ is fatal}\}$$

that evaluate effects of defect types taking into account both their probabilities to kill and their frequencies.

The *Expectation—Maximization (EM) algorithm* ([10], [27]) offers an iterative computational method that converges to the maximum likelihood estimator (see [3]). It allows to split the set of parameters into two groups and estimate each group separately during each iteration, by an *M-step* and an *E-step*, conditionally on the other group. Each M-step represents usual maximum likelihood estimation and involves a moderate group of parameters that can be handled by the chosen optimization routine. During each E-step, the remaining parameters are treated as missing values. Being such, they are estimated by conditional expectations given their old values, obtained from the previous cycle, and the refined first group of parameters. A large number of parameters can be re-estimated by means of the E-step. In view of this, a natural split for model (8) is to estimate effects of defect types, layers, and other causes during the M-step, and to estimate the defect frequencies during the E-step. Explicit formulae for the initial defect frequencies and estimated effects and their iterative recomputation during E-steps and M-steps are given in [1]. In addition, a new directional step is introduced in [1] that coordinates between parameter estimates obtained at different steps and accelerates the entire estimation routine without sacrificing its accuracy. The M-step is the most computer intensive in the entire scheme.

Initial approximation of parameter estimates plays a rather important role here. A good choice of the starting point accelerates the algorithm and prevents it to converge to a local extremum. In this case study, a meaningful initial point was computed by putting defect frequencies of different types λ_{jl} proportional to the observed numbers of classified defects d_{jl}, estimating the total frequency λ_l for $l = 1, \ldots, L$ based on all lots where this layer had been inspected, and for a rough first approximation, replacing all defect sizes x_k by their average $\bar{x}$, defect type effects and layer effects by constants, $r_j = r$ and $a_l = 1$. Then the initial value of r can be obtained by the method of moments, equating the actually observed yield and the yield computed from these approximate values.

B. Assessment of the Goodness of Fit

Repeatedly applying the described three steps until the convergence criterion is met (which is inevitable because each cycle improves the likelihood by at least ε), we obtain the set of parameter estimates $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{\lambda}})$. How good are these estimates, and how adequate is our model? From a practical standpoint, how useful is it for yield prediction, evaluation of kill ratios, designing new strategies, and other objectives? We propose two general goodness-of-fit assessment tools.

1. Standard goodness-of-fit tests (e.g., the chi-square test) compare the expected and observed counts. In our case, it is a comparison of

$$\hat{Y}_m = \begin{array}{l}\text{actual number}\\\text{of functioning chips}\end{array} = \sum_{i\in m}\xi_i$$

and

$$Y_m = \begin{array}{l}\text{predicted number}\\\text{of functioning chips}\end{array} = \sum_{i\in m}\mathbf{E}_{\phi_i}(\xi_i) = \sum_{i\in m}\phi_i$$

for each lot or each wafer. One can evaluate the closeness of $\hat{Y}_m$ to Y_m by the standard chi-square statistic, or even by the correlation coefficient between Y and

$\hat{Y}$. Along with the graph of actual and predicted *yield* $(y_m, \hat{y}_m) = I_m^{-1}(Y_m, \hat{Y}_m)$ for $m = 1, \ldots, M$, I_m being the total number of chips on lot m, it provides a simple illustration of the predictive power.

2. Even if the actual and predicted yields above are found close to each other, it may happen that the right yield was predicted by wrong reasons. For example, one could (at least, theoretically) predict a high yield on actually failed chips and a low yield on good functioning chips that in combination returned a prediction close to Y.

Therefore, it is elucidating to compute predicted yield separately for good (functioning) chips and for bad (failing) chips. The actual numbers of good and bad chips are different, therefore, the only fair comparison is based on *proportional yields*

$$\hat{y}_g = \hat{P}\{\text{predicted good}|\text{actually good}\}$$
$$= \frac{\sum_i \phi_i \xi_i}{\sum_i \xi_i}$$

and

$$\hat{y}_b = \hat{P}\{\text{predicted good}|\text{actually bad}\}$$
$$= \frac{\sum_i \phi_i (1 - \xi_i)}{\sum_i (1 - \xi_i)}.$$

The model obtains predictions $\hat{y}_g$ and $\hat{y}_b$ from the observed defects only, without seeing the actual failures. Ideally, we would certainly wish to predict a 100% yield on good chips and a 0% yield on bad chips. However, this is not possible, according to [1]. One reason for this is existence of pairs of chips with an identical defect situation, whereas one chip in a pair is good and the other is bad. Moreover, there are failed chips that are exposed to other causes only and chips that survived not only the other causes but also many defects on different layers. According to our experience, achieving an estimate that matches the observed level of yield, and a ratio $\hat{y}_g/\hat{y}_b \approx 2$ generally indicates an adequate fit.

C. Clustering and Rare Defects

In practical implementations it is important to take into account a number of special properties of defects that require adjustments of the modeling and estimation procedures. In this section, we discuss two properties of this type: a tendency of defects to cluster and defect rarity.

1) Clustering: Consider a chip with, say, 8 000 detected defects. Under any plausible model, the probability for such a chip to fail is practically 1. However, it is not unusual to see such a chip that is recorded as *good*! Proportion of such chips is low, but keeping them in the overall likelihood without any correction has a strong influence on final results. The model tries to explain their yield, and the only way this yield can be positive is when each defect type seen on a chip so many times has zero effect.

Investigating situations involving thousands to hundreds of defects observed on the same chip, we found that the vast majority of these defects occur on the same layer and belong to the same defect type or remain unclassified. Also, in all such

cases, these defects were marked as *clustered*. Survival of chips containing such clusters indicates that *the effect of a cluster of defects is weaker than the combined effect of the same number of individual defects*. Clusters of different sizes, from just a few defects to thousands of defects, are registered on a large portion of chips, thus such chips cannot be ignored or deleted.

There are different ways the effect of a cluster can be modeled. A plausible method is to treat a cluster as a single defect whose size (before the transformation $x = g(s)$ is applied) equals the sum of their individual sizes. In all the cases such modeling provided better fit comparing with the scheme that treats chips with unusually high number of defects as outliers, deletes them, and applies the uncorrected model to the remaining chips.

2) Rare Defects: A considerable number of defect types are seen rather rarely, say, 1–5 times per 10 000 chips. Even if their effect on the quality of a chip is strong, they do not affect the remaining vast majority of chips, and therefore, the final yield is not affected by such "rare" defect types.

Since for some chips rare defects appear to be the cause of failure, they cannot be simply ignored. Other defects would then be classified as chip killers, introducing a bias in parameter estimates. In view of their small effect on the final yield, all such defect types can be combined, so that only their average effect and their cumulative frequency is estimated. This reduces the overall number of estimated parameters, leading not only to acceleration of the numerical routine, but also to a higher predictive power of the model.

D. Diagnostics and Outlier Detection

Besides situations described in the previous section, large data sets may contain chips, wafers, or even lots, that "do not belong there" and should be treated as outliers. Several outlier detection methods are proposed in this section.

Typically, a certain portion of data can be deleted from the study immediately by means of a simple inspection. This includes wafers with approximately zero yield, chips with hundreds of non-clustered defects, inspected layers with no detected defects on an entire lot (paradoxically!), etc. Such cases are usually results of various errors in data records.

After this inevitable data cleaning, we identify "suspicious" wafers and lots by answering the following two questions:

1) How different would the results be if obtained from this wafer (lot) only?
2) How different would the results be if obtained *without* this wafer (lot)?

To address these questions, we use four types of diagnostics.

1) Likelihood-Based Diagnostics: The general model (8) is estimated separately for each lot m, providing its own maximum value $\mathcal{L}_m(\theta^{(m)})$ of the likelihood of this lot. It is always higher than the value of $\mathcal{L}_m(\theta)$ based on the global estimators θ obtained by means of the modified EM algorithm described above. If the difference is significant, it means that the global estimators do not fit to lot m; its likelihood can be increased significantly given its own parameters.

To find a fair measure of significance, we notice that the test statistic

$$2 \sum_{m=1}^M \log \mathcal{L}_m(\theta_m) - 2 \log \mathcal{L}(\theta) \qquad (10)$$

has asymptotically χ^2 distribution with $(J+L)(M-1)$ degrees of freedom, if there are no differences between the lots. Here J, L, and M represent the number of defect types, layers, and lots, respectively. If a lot is an outlier, and the global parameters don't fit it, one should expect its two-log-likelihood to increase approximately by a χ^2 variable with $(J+L)(1-1/M)$ degrees of freedom. Exceeding the critical value of χ^2_α automatically puts a lot into the list of suspicious ones.

Further, the test statistic (10) consists of the sum of differences, by wafer, and thus, the largest summand points to the most outlying wafer that is responsible for the large difference.

Remark: Even for large data sets, it is typically feasible to conduct separate estimation for each lot. In the process of such estimation, the global parameter estimates θ are very helpful, because they can serve as initial values in estimation for each lot. Notice that the purpose of this analysis is to find significant deviations from the global results. If a lot is well explained by the global model, such a deviation is small, and starting from θ, the algorithm will converge quickly.

2) Parameter-Based Diagnostics: Continuing the likelihood-based diagnostics, one can compare the lot-specific estimates θ_m with the global estimates θ. Under the null hypothesis, or in the absence of outliers, vectors θ_m, obtained from similar lots with the same number of inspected layers, are i.i.d., and their mean vector and covariance matrix can be estimated by standard methods. This provides a null multivariate normal distribution (according to the asymptotic normality of maximum likelihood estimators), against which the differences $(\theta_m - \theta)$ can be compared.

3) Prediction-Based Diagnostics: A different way to identify lots that do not agree with the global model is to analyze the yield predictions for each lot. A good model should separate the predicted yield on good and on bad chips $\hat{y}_g$ and $\hat{y}_b$, and the difference between $\hat{y}_g$ and $\hat{y}_b$ should be positive. Otherwise, the model appears not to have a good predictive power on such a lot. Then it should be deleted from the estimation routine, as long as the parameter estimates are used for prediction on future lots. Typically, however, most of these lots appear to be already deleted by the two diagnostics tools described above.

4) Cross-Validation: In cross-validation, each lot is deleted, one at a time, and the model parameters are estimated without it. Again, the global parameter estimates θ can be used as initial values for the EM-algorithm. Then, predictions are made on the deleted lot and compared with its actual yield. Large differences between predicted and actual yield indicate possible outliers.

All the proposed methods can be applied either to lots or to wafers. The latter is recommended for wafers with sufficiently many inspected layers. We have seen a number of cases where an outlying lot was classified as such only due to one outlying wafer on it. Applying the diagnostics tools on a wafer level would in general delete fewer units. On the other hand, wafers

with only one or two layers inspected may provide a rather small sample of defects. Inference made on such a wafer separately from other wafers in diagnostics tools 1–3 will not be reliable.

After the "suspicious" lots (wafers) are identified, it is very useful and strongly recommended to conduct an in-depth analysis of each case. In our application, it resulted in a number of "discoveries". Almost every critical case was attributed to a special cause, such as error in data collection and data recording, spontaneous change in the sensitivity of the inspection instrument, scrapped wafer or reworked layer.

IV. APPLICATIONS: FORECASTING YIELD AND IDENTIFYING MAIN CAUSES OF FAILURES

The described cycle of data cleaning, model fitting, model diagnostics, outlier detection, and most likely, model refitting and refinement results in a estimated model and a set of parameter estimates. This section concerns immediate applications of this analysis, usable information and interpretation that can be drawn from it.

1) Yield Forecasting: One of the obvious practical by-products of our modeling is the possibility to predict the yield for each lot and each wafer. Indeed,

$$\phi_i = \mathbf{E}\{\xi_i|\text{observed defects occur on chip } i)\}$$

is the expected yield, or number of good chips, out of one chip i. Then

$$\hat{Y}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \phi_i \qquad (11)$$

is the expected number of good chips that can be computed for any set of chips $\mathcal{I}$, which may be a wafer, a lot, or a number of lots. Thus $\hat{Y}_{\mathcal{I}}/|\mathcal{I}|$ serves as the yield forecast, and it is based on the observed defects on $\mathcal{I}$ and the parameter estimates obtained from the training data.

Using this method, the yield can be predicted at the end of the production cycle, after all layers have been processed but before the final testing. It can also be used to predict the yield at earlier stages of the production line. Layers that have not been processed are then treated as uninspected. If a low yield is predicted on some wafer or lot, a decision may be made about terminating its production at an early stage, or a layer at fault can be reworked.

2) Kill Ratios and Weighted Defect Densities: Next, we will derive the kill ratios and related probabilities, such as the probability for a certain defect to be fatal, the probability for a certain defect type or layer to contain the chip killer, etc.

According to (1), the *kill ratio*, or the probability that a type j defect on layer l is fatal, is

$$p_{jl}^{\text{KR}} = 1 - \mathbf{E}_j^x \exp\{-r(j)a(l)x\} = 1 - \psi_{jl}. \qquad (12)$$

Further, the probability for an arbitrary type j defect to be fatal is

$$p_{j.}^{\text{KR}} = \sum_l \frac{\lambda(j,l)}{\lambda(j)} p_{jl}^{\text{KR}} \qquad (13)$$

by the formula of total probability. This is a *proportion of type j defects that appear fatal* for a chip. Analysis of large data sets showed that estimator (13) is more accurate than the kill ratios computed from defect counts only ([13], [16]). Similarly, one computes the probability for a defect on layer l to be a chip killer,

$$p_{.l}^{\mathrm{KR}} = \sum_j \frac{\lambda(j,l)}{\lambda(l)} p_{jl}^{\mathrm{KR}} \tag{14}$$

which is a *proportion of fatal defects on layer l*. Finally,

$$p_{..}^{\mathrm{KR}} = \sum_j \sum_l \frac{\lambda(j,l)}{\lambda} p_{jl}^{\mathrm{KR}} \tag{15}$$

is *the overall proportion of fatal defects*. Generalizations used in $p_{.l}^{\mathrm{KR}}$, $p_{j.}^{\mathrm{KR}}$, and $p_{..}^{\mathrm{KR}}$, are helpful to compare kill ratios across defect types and across layers. One estimates these probabilities by replacing unknown ψ_{jl} and $\lambda(j,l)$ with their estimates (Section III).

Expressions (13)–(15) are based on the products $\lambda(j,l) p_{jl}^{\mathrm{KR}}$ that are called *weighted defect densities* by practitioners. Combining the probability to kill with the defect frequency, these quantities are often used to measure the adverse effect of each group of defects on the final yield.

3) Causes of Failures: In this section, we consider chips that are known to have failed. Can the model pinpoint the cause of their failure?

The probability for a failed chip to be killed by a type j defect on layer l is

$$p_{jl}^{\mathrm{CF}} = \frac{p_{jl}^{\mathrm{GKR}}}{1 - e^{-b}(1 - q)} \tag{16}$$

where $p_{jl}^{\mathrm{GKR}} = 1 - \exp\{-\lambda(j,l)(1 - \psi_{jl})\}$ is *the group kill ratio*, i.e., the probability of *at least one chip killer among all type j defects on layer l*, and $p_{.l}^{\mathrm{GKR}} = 1 - \prod_l \prod_j (1 - p_{jl}^{\mathrm{GKR}})$ is the probability of *at least one fatal defect* on a chip, which is also the probability for a chip to fail not (or not only) because of other causes ([1]).

Similarly, *the probability for a failed chip to be killed by a type j defect is*

$$p_{j.}^{\mathrm{CF}} = \frac{p_{j.}^{\mathrm{GKR}}}{1 - e^{-b}(1 - p_{..}^{\mathrm{GKR}})} \tag{17}$$

the probability for a failed chip to be killed by some defect on layer l is

$$p_{.l}^{\mathrm{CF}} = \frac{p_{.l}^{\mathrm{GKR}}}{1 - e^{-b}(1 - p_{..}^{\mathrm{GKR}})} \tag{18}$$

and *the probability for a failed chip to be killed by some defect* (and not, or not only, by other causes) is

$$p_{..}^{\mathrm{CF}} = \frac{p_{..}^{\mathrm{GKR}}}{1 - e^{-b}(1 - p_{..}^{\mathrm{GKR}})} \tag{19}$$

where $p_{j.}^{\mathrm{GKR}} = 1 - \prod_l(1 - p_{jl}^{\mathrm{GKR}})$ is the probability of *at least one chip killer on a failed chip among all type j defects*, and $p_{.l}^{\mathrm{GKR}} = 1 - \prod_j(1 - p_{jl}^{\mathrm{GKR}})$ is the probability of *at least one fatal defect on layer l*.

Also, $p_{..}^{\mathrm{CF}}$ represents *the proportion of failed chips that contain fatal observable defects*, and $(1 - p_{..}^{\mathrm{CF}})$ is *the proportion of failed chips that are killed by other causes*. For the derivation of these expressions, see [1].

4) Single Cause of Failure: Based on the probabilities computed above, should defects with the highest chance of being fatal be regarded to as most dangerous, and should the reduction of such defects be given the highest priority?

Noticeably, even having such a defect on some layer, a failed chip may have been killed by other defects. In view of this, for example, it would be wrong to attribute the proportion of $p_{j.}^{\mathrm{CF}}$ of failed chips to defects of type j only. And if so, then what proportion of failed chips is due to defects of type j?

Here we compute probabilities for a group of defects on a failed chip *and nothing else* to cause its failure. Thus, such probabilities are also proportions of failed chips that can be attributed to the considered defects *only*, making such a group a *single cause of failure*.

The probability that a chip failed due to defects of type j on layer l *only* is

$$p_{jl}^{\mathrm{SCF}} = \frac{p_{jl}^{\mathrm{CF}} - p_{jl}^{\mathrm{GKR}}}{1 - p_{jl}^{\mathrm{GKR}}}. \tag{20}$$

Similarly, the probability that *all* fatal defects on a failed chip are of type j is

$$p_{j.}^{\mathrm{SCF}} = \frac{p_{j.}^{\mathrm{CF}} - p_{j.}^{\mathrm{GKR}}}{1 - p_{j.}^{\mathrm{GKR}}} \tag{21}$$

the probability that the chip failed due to defects on layer l *only*,

$$p_{.l}^{\mathrm{SCF}} = \frac{p_{.l}^{\mathrm{CF}} - p_{.l}^{\mathrm{GKR}}}{1 - p_{.l}^{\mathrm{GKR}}} \tag{22}$$

and the probability that *a failed chip was killed by defects but not by other causes*,

$$p_{..}^{\mathrm{SCF}} = e^{-b} p_{..}^{\mathrm{CF}}. \tag{23}$$

Notice that the sum of probabilities in each (20)–(22) is less than 1. Each equation deals with probabilities of mutually exclusive but not exhaustive events because several types of defects or several layers may be at fault for the chip failure.

5) Influential Layers: If an unusually low or unusually high yield is predicted for some lot, a simple method can be proposed to find the main reason of the unusual prediction. Typically, the reason for an unusual prediction is an unusual situation that occurred on some inspected layer (of course, no surprise can be found on uninspected layers). It is then straightforward to evaluate the contribution of each layer into prediction.

Moving sequentially through all the inspected layers on a lot or a wafer, consider one inspected layer at a time. Recompute predicted yield $\hat{y}_m$ with this layer being hidden, or uninspected.

This operation does not require additional computer code or much computer time. As a result, we obtain the difference in predicted yield which shows how much of the yield is lost or gained due to the defects observed on this layer.

Not only does this measure the influence of each inspected layer on the final yield, but it also points practitioners in a good direction in their efforts to improve yield.

6) Sensitivity Analysis: When choosing an efficient yield improving strategy, one would be interested to predict the changes such a strategy will bring in the final yield. For example, if one manages to reduce the number of type j defects by 10%, how strongly will this affect the yield? Is it worth the effort, and is there a more efficient strategy? In other words, how sensitive is the yield to certain changes in frequencies or sizes of defects?

These questions can be answered by recomputing the predicted yield with altered parameters. One obtains that *sensitivity of yield to the frequency of type j defects on layer l* is measured by

$$e_{jl} = -\lambda(j,l)(1 - \psi_{jl}). \tag{24}$$

This formula should be used as follows: an $100\alpha\%$ reduction in the number of type j defects on layer l results in the yield improvement by $\alpha e_{jl}\%$. For example, a 5% reduction in the number of type j defects on layer l results in a yield improvement by $0.05e_{jl}\%$.

Next, suppose that a certain modification of the production process can reduce the size s of type j defects on layer l by $100\beta\%$. As a result, *the proportional change in yield due to a small reduction of sizes of type j defects on layer l* is

$$f_{ij} = \lambda(j,l)r(j)a(l)\mathbf{E}_j^x \exp\{-r(j)a(l)x\}s/(s+1)$$

For relatively large sizes s, it can be approximated by

$$f_{ij} \approx \lambda(j,l)r(j)a(l)\psi_{jl}. \tag{25}$$

Then, a $100\beta\%$ reduction in sizes of type j defects on layer l result in approximately $100f_{jl}\beta\%$ improvement in the final yield.

V. PRACTICAL USE AND A CASE STUDY

Having a set of estimated parameters, a practitioner can routinely predict the yield on new finished or unfinished lots and identify the most influential layers responsible for high or low predictions. This part of the scheme is not computationally intensive; it can be conducted on demand or be completely automated.

Regular diagnostics of the model and assessment of the goodness of fit is certainly highly recommended. When insufficient fit is detected (or, say, at least once a month), parameters of the model should be re-estimated. This part of the proposed scheme is the most time consuming, as it involves estimation of a large number of parameters from large data sets. Depending on the size of the problem, the desired level of accuracy and the quality of initial approximation, the proposed EM algorithm may take from several hours to a couple of days on a 2.4 mHz personal computer with 1 GHz of RAM that was used in our analysis.

Generally, execution time can be greatly reduced by using previous parameter estimates as initial approximation for the new set of iterations.

Tracking all major changes in parameter estimates appears very informative. Not only does it show the trend of the overall yield, but also the results of implemented strategies reducing or eliminating certain defects.

To illustrate the proposed methodology, numerical results are reported for a certain type of wafers produced at IBM in a stream of consecutive lots. The data set is real and so are the reported results, although products, time frames, defect types and layers cannot be named for reasons of commercial confidentiality. In this report, defect types are coded as D1, D2, ..., and layers are L1, L2, ...

Size of the Problem: This data set, of moderate size, included 145 wafers from $M = 25$ lots, with a total of 284 796 defects observed on $L = 24$ layers of 33 634 chips. Each wafer contained between 430 and 469 tested chips. Among the observed defects, 19 473 were classified to $J = 65$ defect type—layer combinations, and 265 323 were unclassified.

Only a few layers were inspected on each wafer. Only one layer was inspected on 50 wafers out of 145, between two and five layers on 55 wafers, the other 40 wafers had at least 6 layers inspected. Thus, most of the information about defects was treated as missing data by the proposed methods.

Estimation, Outlier Detection, and Goodness-of-Fit: After nine cycles of the modified EM algorithm (each cycle taking approximately 30 minutes of CPU), the negative log-likelihood of the observed defects and chip failures $|\ln \mathcal{L}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{r}, \boldsymbol{\lambda})|$ reduced from 406 254 to 220 335 and was not changing by more than $\varepsilon = 0.0001$ per cycle, thus meeting the convergence criterion. The main reduction, from 406 254 to 220 637, occurred during the first 2 iterations, after which the parameter estimates were slightly refined.

Searching for possible outliers, similar analysis was conducted, in turn, for lot 1, then for lot 2, etc. The estimates obtained from the entire data set were used as starting values, therefore, for most of the individual lots the algorithm converged quickly. For each lot, the difference in $|2 \ln \mathcal{L}|$ (corresponding to the contribution of this lot to (10)) was compared against a chi-square distribution with $(J+L)(1-1/M) = 85.4$ degrees of freedom. On four lots, the critical value of 107.5 at the 5% level of significance was exceeded, and for the same lots the parameter estimates differed significantly from the estimates obtained from the entire set. The seven most extreme wafers were found to be responsible for a significant increase in $|2 \ln \mathcal{L}|$. We then proceeded to exclude these wafers, re-estimate the parameters and establish that there were no goodness-of-fit violations in accordance with the criteria of Section III.

As a result, the predicted proportional yield was $\hat{y} = 0.2552$, whereas the actual proportion of good chips in this data set was $y = 0.2542$. One should expect very close results here because prediction and estimation was conducted on the same data set. Subsequent application of the model to a new testing sample of 27 lots confirmed that the model has indeed predictive value. Despite a usual situation with unclassified defects and uninspected layers on the new lots, the correlation between predicted and actual yield is 0.487 (Fig. 1).
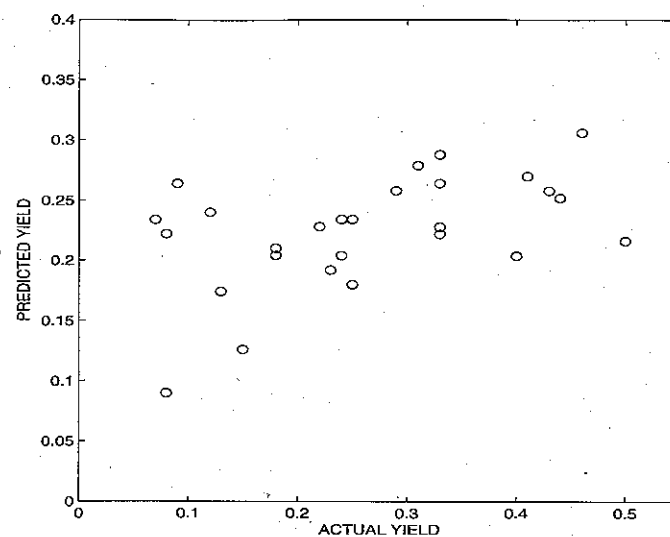
Fig. 1. Yield prediction for the testing sample of 27 lots.

The defects-based model distinguished between good and bad chips. Having only the information about the observed defects, it predicted a proportional yield of $\hat{y}_g = 31.87\%$ among chips that were actually good, and $\hat{y}_b = 23.35\%$ among chips that actually failed. Although we wished for a rule-of-thumb $2:1$ ratio between $\hat{y}_g$ and $\hat{y}_b$, one should take into account that very few layers were inspected on most of the wafers. Yield prediction on these wafers was based on only a small observed portion of actually occurring defects.

Influential Layers: Unusual defect situation on some of the inspected layers was found responsible for for unusual predictions on 4 wafers, all within the same lot. An 18% to 19% of proportional yield was lost due to an unusually high concentration of defects on layer L1, and on one wafer, an 8% loss was due to layer L2. On the other hand, unusually small number of defects on layers L3 and L4 caused a 7% proportional increase in our yield predictions. These unusual defect-layer situations were then reviewed.

Practical Conclusions: Based on the obtained parameter estimates, it was found that wafers of this type should provide a 24.7% yield. Sensitivity analysis shows that this yield can be increased by 1.26%, if the frequency of all types of defects on all levels is simultaneously reduced by 1%. For individual defect types, the biggest yield increase (of 0.25%) should be expected if the frequency of type D14 defects is reduced by 1%.

Also, a 0.86% increase of the yield should be expected from an across the board 1% reduction of defect sizes of all types. This figure can be computed separately for each defect type, each layer, or each defect-layer combination.

Further, the average number of defects (observed and unobserved) per chip was estimated as 7.55, among which 11.1% are predicted to be fatal. Out of these, the most dangerous defect types were D2, D7, D12, D14, and D20, whose probability to kill a chip is over 0.5 when they occur on certain layers.

Among the chips that are known to have failed, D14 was the major cause of failures, as almost 30% of bad chips failed due to this defect type. It was also found that 75.1% of failures were caused by some (observed or unobserved) defects,

and only 24.9% were exclusively due to other causes. The most vulnerable layers were L5 and L6 which accommodated 23.1% and 12.7% of fatal defects, respectively.

Such numerical results are helpful in choosing an *optimal strategy* for improvement in defect-limited yield. For example, defects D7 and D12 are among the most dangerous ones, however, they occur on a chip with low probabilities of 0.019 and 0.028, respectively, and therefore, they were rarely found as causing failures. Based on the analysis illustrated above, one could rather choose to direct efforts towards reducing type D11 defects that occur at a high rate of 1.8 defects per chip, causing 11.3% of all failures. Successful implementation of such defects reducing strategy can be expected to result in a certain yield improvement, predictable by the sensitivity analysis.

The model was applied to subsequent product stream, typically producing a correlation between observed and predicted yields in the range $(0.3, 0.5)$, depending on the number of inspected layers. We also observed groups of lots where the correlation would weaken to levels below 0.1 or disappear completely; in such cases, we were typically able to trace prediction failure to anomalous conditions, data quality issues or individual wafers. Generally, only yields for lots where relatively "dirty" layers were observed could be predicted with an actionable level of confidence. Lots for which the inspected layers appeared "normal" or relatively clean accounted for most of large prediction errors because of the low fraction of inspected layers. We expect the accuracy of yield prediction to improve when the information on observed defects is combined with other data collected on the wafers, such as electrical inline tests. Of special value, however, appeared to be the ability of our model to generate a list of suspected defect type-layer combinations, and in rare cases direct the attention of the engineers to previously unknown yield detractors.

VI. SUMMARY AND CONCLUDING REMARKS

In general, one computes (11) to forecast the yield under the current conditions. Then, (12)–(23) show the most dangerous defects and defect type-layer combinations, the most probable causes of failures, and influential layers that had the highest impact on the final yield. When choosing an efficient strategy, modifying the production process and resulting in reduced numbers or reduced sizes of such defects, one uses (24)–(25) to predict possible outcomes. In practice, the expected gain from such a strategy will then be weighed against its expected costs, and based on this balance, a business decision regarding its implementation will be made.

The described EM algorithm appears to be numerically stable: we did not see a single case where it converged to a provably local maximum. We suspect that the underlying reason is that the log-likelihood function corresponding to (8) is pseudo-concave—but we have no proof of it. One of the most time-consuming and challenging issues in the estimation process is making sure that it is not undermined by the data quality or stability issues. For example, a decision by an engineering team to rename a defect, a decision to increase the sensitivity of an optical scanner for selected layers, or a decision to change the reporting procedure for re-worked layers can easily remain unnoticed and lead to loss of predictive

power. Furthermore, often there are reasons to suspect that the process of defect frequencies exhibits temporal changes. In such cases, defect frequencies for older lots can be mostly viewed as nuisance parameters whose value is primarily in strengthening inference on $\{a(l)\}, \{r(j)\}, \{b(m)\}$, and possibly other parameters. Of primary interest in terms of defect frequencies are more recent lots. Defect frequencies estimated from these lots can be filtered (using, for example, a process of exponential smoothing) to produce estimates of the current defect frequencies. Obtaining filters that are provably powerful and robust is an important open problem, especially in light of high percentage of missing data. The estimated current frequencies can then be used to predict yields for partially processed lots. Providing an ability to detect and re-work (if possible) lots that are predicted to have a low yield is one of the most important benefits offered by the proposed model.

The approach described in this paper can be further developed in several directions. Of special importance might be extensions related to use of various types of regularization that could potentially further reduce prediction error and provide better understanding of the sources of variability, especially in relation to defect intensities. One possible approach in this respect could take into account prior information about the variability in the framework of Bayesian analysis.
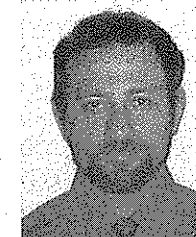
REFERENCES

[1] M. Baron, A. Takken, E. Yashchin, and M. Lanzerotti, Factorial Analysis and Forecasting of Integrated-Circuit Yield. Yorktown Heights, NY, IBM Res. Rep. RC #23386, 2004.

[2] M. Baron, C. K. Lakshminarayan, and Z. Chen, "Markov random fields in pattern recognition for semiconductor manufacturing," *Technometrics*, vol. 43, pp. 66–72, 2001.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[4] L. Devroye and L. Györfi, *Nonparametric Density Estimation. The L_1 View*. New York: Wiley, 1985.

[5] A. V. Ferris-Prabhu, "Modeling the critical area in yield forecast," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 4, pp. 878–880, Aug. 1985.

[6] M. H. Hansen, V. N. Nair, and D. J. Friedman, "Circuit fabrication processes for spatially clustered defects," *J. Amer. Statist. Assoc.*, vol. 39, pp. 241–253, 1997.

[7] R. S. Hemmert, "Poisson process and integrated circuit yield prediction," *Solid-State Electron.*, vol. 24, pp. 511–515, 1981.

[8] M. B. Ketchen, "Point defect yield model for wafer scale integration," *IEEE Circuits and Devices Mag.*, vol. 1, pp. 24–34, Jul. 1985.

[9] M. D. Longtin, L. M. Wein, and R. E. Welsch, "Sequential screening in semiconuctor manufacturing, I: Exploiting spatial dependence," *Oper. Res.*, vol. 44, pp. 173–195, 1996.

[10] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[11] F. J. Meyer and D. K. Pradhan, "Modeling defect spatial distribution," *IEEE Trans. Comput.*, vol. 38, no. 4, pp. 538–546, Apr. 1989.

[12] L. S. Milor, "Yield modeling based on in-line scanner defect sizing and a circuit's critical area," *IEEE Trans. Semicond. Manuf.*, vol. 12, no. 1, pp. 26–35, Feb. 1999.

[13] P. Mullenix, J. Zalnoski, and A. J. Kasten, "Limited yield estimation for visual defect sources," *IEEE Trans. Semicond. Manuf.*, vol. 10, no. 1, pp. 17–23, Feb. 1997.

[14] R. Ott, H. Ollendorf, H. Lammering, T. Hladschik, and W. Haencsh, "An effective method to estimate defect limited yield impact on memory devices," in *Proc. IEEE/SEMI Advanced Semiconductor Manufacturing Conf.*, 1999, pp. 87–91.

[15] E. Papadopoulou and D. T. Lee, "Critical area computation via Voronoi diagrams," *IEEE Trans. Comput.-Aided Design of Integrated Circuits Syst.*, vol. 18, no. 4, pp. 463–474, Apr. 1999.

[16] O. D. Patterson and M. H. Hansen, "The impact of tolerance on kill ratio estimation for memory," *IEEE Trans. Semicond. Manuf.*, vol. 15, no. 4, pp. 404–410, Nov. 2002.

[17] S. L. Riley, "Limitations to estimating yield based on in-line defect measurements," in *Proc. 1999 IEEE Int. Symp. Defect and Fault Tolerance in VLSI Systems*, 1999, pp. 46–54.

[18] J. Shier, "A statistical model for integrated-circuit yield with clustered flaws," *IEEE Trans. Electron Devices*, vol. 35, no. 4, pp. 524–525, Apr. 1988.

[19] C. H. Stapper, "On yield, fault distributions, and clustering of particles," *IBM J. Res. Develop.*, vol. 30, pp. 326–338, 1986.

[20] C. H. Stapper, "Large area fault clusters and fault tolerance in VLSI circuits: A review," *IBM J. Res. Develop.*, vol. 33, pp. 162–173, 1989.

[21] C. H. Stapper and R. J. Rosner, "Integrated circuit yield management and yield analysis: Development and implementation," *IEEE Trans. Semicond. Manuf.*, vol. 8, no. 2, pp. 95–102, May 1995.

[22] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated circuit yield statistics," *Proc. IEEE*, vol. 71, no. 1, pp. 453–470, Jan. 1983.

[23] W. Taam and M. Hamada, "Detecting spatial effects from factorial experiments: An application from integrated-curcuit manufacturing," *Technometrics*, vol. 35, pp. 149–160, 1993.

[24] A. Venkataraman and I. Koren, "Determination of yield bounds prior to routing," in *Proc. 1999 IEEE Int. Symp. Defect and Fault Tolerance in VLSI Systems*, 1999, pp. 4–13.

[25] I. A. Wagner and I. Koren, "An interactive VLSI CAD tool for yield estimation," *IEEE Trans. Semicond. Manuf.*, vol. 8, no. 2, pp. 130–138, May 1995.

[26] R. M. Warner, "Applying a composite model to the IC yield problem," *IEEE J. Solid-State Circuits*, vol. SC-9, no. 3, pp. 86–95, Jun. 1974.

[27] M. Watanabe and K. Yamaguchi, *The EM Algorithm and Related Statistical Models*. New York: Marcel Dekker, 2003.

Michael Baron Dr. Baron is Professor of Statistics at the Department of Mathematical Sciences at the University of Texas at Dallas. His research areas include sequential analysis, change-point problems, Bayesian inference, and applications of Statistics in semiconductor manufacturing, clinical trials, epidemiology, and energy finance. In 2003–04, he joined the IBM T. J. Watson Research Center as an Academic Visitor. M. Baron has a University Diploma in Mathematics from St. Petersburg State University, Russia (1992) and a Ph.D. degree from the University of Maryland (1995). In his turn, he graduated four doctoral students and is working on the fifth one.

Asya Takken received the B.A. degree (*cum laude*) in mathematics from Harvard University, Cambridge, MA, in 1992, and the Ph.D. degree in statistics from Stanford University, Palo Alto, CA, in 2000.

She is currently a member of the Market Analytics team at Cisco. Her areas of application included internet advertising (at Doubleclick 2000–2002) and chip fabrication (at IBM, 2002–2006).

Dr. Takken is a member of the American Statistical Association.

Emmanuel Yashchin received the Diploma in applied mathematics from Vilnius State University (U.S.S.R.) in 1974, the M.Sc. degree in operations research and the D.Sc. degree in statistics from the Technion—Israel Institute of Technology, Haifa, in 1977 and 1981, respectively.

In 1982, he was a Visiting Assistant Professor at the Iowa State University. Since 1983, he is a Research Staff Member in the Department of Mathematical Sciences, IBM Thomas J. Watson Research Center. From 1996 to 2002, he served as the Manager of the Statistics Group in IBM Research. His research interests include Quality Control, Reliability, Statistical Modeling, Risk Analysis and Operations Research.

Dr. Yashchin is an Elected Member of the International Statistical Institute, Fellow of the American Statistical Association, and Senior Member of the American Society for Quality.

Mary Yvonne Lanzerotti (M'97–SM'05) received the A.B. degree (*summa cum laude*) from Harvard University, Cambridge, MA, in 1989, the M.Phil. degree from University of Cambridge in 1991, the M.S. degree in 1994, and the Ph.D. degree in 1997, both from Cornell University, Ithaca, NY, in 1997, all in physics.

She is a Research Staff Member (VLSI Design Department) at the IBM Thomas J. Watson Research Center. She joined IBM in 1996. Her research interests include the design of on-chip interconnections and analysis of timing-critical paths in high-performance microprocessors for IBM pSeries and zSeries eServers.

Dr. Lanzerotti is a member of the IEEE-Solid State Circuits Society, IEEE-Lasers and Electro-Optics Society, IEEE Women in Engineering, American Physical Society, and Phi Beta Kappa.

Automatic Identification of Defect Patterns in Semiconductor Wafer Maps Using Spatial Correlogram and Dynamic Time Warping

Young-Seon Jeong, Seong-Jun Kim, and Myong K. Jeong

Abstract—A wafer map is a graphical illustration of the locations of defective chips on a wafer. Defective chips are likely to exhibit a spatial dependence across the wafer map, which contains useful information on the process of integrated circuit (IC) fabrication. An analysis of wafer map data helps to better understand ongoing process problems. This paper proposes a new methodology in which spatial correlogram is used for the detection of the presence of spatial autocorrelations and for the classification of defect patterns on the wafer map. After the detection of spatial autocorrelation based on our proposed spatial randomness test using spatial correlogram, the dynamic time warping algorithm which provides nonlinear alignments between two sequences to find optimal warping path is adopted for the automatic classification of spatial patterns based on spatial correlogram. We also develop generalized join-count (JC)-based statistics and then propose a procedure to determine the optimal weights of JC-based statistics. The proposed method is illustrated using real-life examples and simulated data sets. The experimental results show that our method is robust to random noise and has a robust performance regardless of defect location and size.

Index Terms—Dynamic time warping, join-count (JC) statistics, spatial autocorrelation, spatial correlogram, wafer map.

I. INTRODUCTION

A WAFER is an elementary unit in semiconductor manufacturing. Several hundred integrated circuits (ICs) are simultaneously fabricated on a single wafer (Fenner *et al.* [14]). After the completion of IC fabrication, each chip is classified as either functional or defective. A wafer map is used to display the locations of defective ICs chips on the wafer. A wafer map is likely to exhibit a spatial dependence across the wafer. As explained in Hansen *et al.* [17], defective chips commonly occur in clusters or display some systematic patterns. Such defect patterns contain useful information about manufacturing process conditions (Cunningham and McKinnon [20]). For example, uneven temperatures or chemical aging lead to spatial cluster on the wafer map. Clusters also can be the result of crystalline nonuniformity, photo-mask misalignment, or particles caused by mechanical vibration. Stepper and/or probe malfunctioning and sawing imperfections also are major causes of repetitive patterns. Material shipping and handling also can leave a scratch on the wafer map (Cunningham and McKinnon [20], Hansen and Tyregod [7], Hansen *et al.* [17], and Taam and Hamada [22]).

The defect patterns represented on the wafer map hold important information that can assist process engineers in their understanding of the ongoing manufacturing processes. Consequently, wafer maps have been widely used in the semiconductor industry for process monitoring and yield enhancement. Chen and Liu [10] and Liu *et al.* [19] developed intelligent systems that use wafer maps and wafer bin maps, respectively, to recognize defect spatial patterns and aid in the diagnosis of causes of failures. They adapted a neural network called an adaptive resonance theory network 1 (ART1) for this purpose. Hsieh and Chen [12] developed an analytical structure made up of a fuzzy rule-based inference system to help identify defect spatial patterns. Tong *et al.* [15] used the multivariate Hotelling T^2 control chart that indexes the number of defects and defect clusters as a way to monitor the wafer manufacturing process. The merit of this method is that it simultaneously monitors the number of defects and the presence of the cluster of defects.

As a wafer gets larger, a spatial inhomogeneity frequently occurs. According to the literature (Bailey and Gatrell [21]), analysis of spatial inhomogeneity is also one of the promising approaches for detecting defective clustering. However, there is very little in the literature about the use of spatial correlogram to analyze defect patterns on the wafer map. This paper proposes a new methodology based on spatial correlogram to detect the presence of spatial autocorrelations and classify defect patterns. This paper is the first attempt to develop a methodology to detect spatial autocorrelation and to classify defect patterns automatically based on a spatial correlogram of a wafer map. After detecting the presence of defect patterns, dynamic time warping (DTW) is adopted to classify defect patterns into one of known patterns automatically. Spatial correlogram based on the proposed method is very robust to random noise, defect location, and defect size on the wafer map.

The remainder of this paper is organized as follows. Section II generalizes a couple of join-count (JC)-based statistics and explores their properties. Section III describes a spatial correlogram and proposes generalized JC-based statistic with optimal weights. Section IV contains a visual illustration that uses simulated and real life examples and presents a new spatial randomness test. In Section V, we present the new automatic defect classification methodology and compare its performance with that of neural network. Section VI presents conclusions and some future research topics.

SPECIAL SECTION ON THE INTERNATIONAL SYMPOSIUM ON SEMICONDUCTOR MANUFACTURING

IEEE TRANSACTIONS ON

SEMICONDUCTOR MANUFACTURING

SPECIAL SECTION ON THE IEEE INTERNATIONAL CONFERENCE ON MICROELECTRONIC TEST STRUCTURES

TSM/21/4/2005

Dallas

IEEE