Editor's Note: This article will be presented at the Spring Research Conference, June 18–20, in Roanoke, Virginia.

Markov Random Fields in Pattern Recognition for Semiconductor Manufacturing

Michael Baron

Programs in Mathematical Sciences
University of Texas at Dallas
Richardson, TX 75083-0688
(*mbaron@utdallas.edu*)

Choudur K. Lakshminarayan and Zhenwu Chen

Texas Instruments Inc.
Dallas, TX 75265

Under the most general conditions of an anisotropic Markov random field, we model the two-dimensional spatial distribution of microchips on a silicon wafer. The proposed model improves on its predecessors as it stipulates the spatial correlation of different strengths in all eight directions. Its canonical parameters represent the intensity of failures, main effects, and interactions of neighboring chips. Explicit forms of conditional distributions are derived, and maximum pseudo-likelihood estimates of canonical parameters are obtained. This numerical characteristic summarizes general patterns of clusters of failing chips on a wafer, capturing their size, shape, direction, density, and thickness. It is used to classify incoming wafers to known root-cause categories by matching them to the closest pattern.

KEY WORDS: Artificial neural network; Clique; Exponential family; Gibbs sampler; Maximum pseudo-likelihood; Neighborhood.

Our objective in this article is to model the pattern of failing chips on a silicon wafer, find its numerical (perhaps, multidimensional) characteristic, and recognize patterns of failing clusters of chips captured by this numerical characteristic. We then use an artificial neural network to classify incoming produced wafers and match them to the known root causes by finding the closest known pattern.

To achieve this goal, we propose a probabilistic model that describes the joint distribution of chips on a wafer, taking into account their mutual spatial dependence. Here we invoke the idea of Markov dependence according to which the probability of failure for any chip depends on the quality of other chips on a wafer in such a way that closer chips have a stronger effect on it than spatially distant ones. A vast amount of literature has been devoted to deriving a stochastic model of failing chips; see Stapper (1975, 1989), Stapper, Armstrong, and Saji (1983), Ketchen (1985), Hemmert (1981), Meyer and Pradhan (1989), Shier (1988), Taam and Hamada (1993), and others. However, no realistic and mathematically tractable joint distribution had been derived until Longtin, Wein, and Welsch (1996) proposed the use of Markov random fields to model the inner 16 mm of wafers. They considered an isotropic model with only horizontal and vertical interactions of chips on a wafer, mentioning its anisotropic extension. Apart from nonstationary effects, this model does not capture any diagonal patterns. On the other hand, a noticeable portion of clusters of defective chips in the current production are stretched in directions other than horizontal and vertical. To resolve this problem, we propose an *anisotropic* Markov-random-field model to capture interactions in all the directions on the surface and to explain the complicated two-dimensional correlation structure of failures.

The assumption of a Markov random field means that every chip is considered along with the neighborhood of surrounding chips so that the probability of its failure depends on the remaining chips only through its neighborhood. In other words, given the condition of neighboring chips, the remaining part of a wafer does not add information about the probability of chip failure. Any probability measure on a wafer whose conditional distributions define a neighborhood structure is a Markov random field (see Besag 1974; Cressie 1991).

The model is defined as follows. Using Cartesian coordinates on a wafer, a Bernoulli variable X_{ij} for the (i, j)th chip is assigned, where $X_{ij} = 1$ if the chip fails, $X_{ij} = 0$ otherwise. If $\mathcal{Q}_{i,j}$ is a neighborhood of the (i, j)th chip, the Markov-field assumption is equivalent to

$$\mathbf{P}\{X_{ij} = 1 | X_{kl}, (k, l) \neq (i, j)\} = \mathbf{P}\{X_{ij} = 1 | X_{kl}, (k, l) \in \mathcal{Q}_{ij}\}.$$

A common scheme involves nearest neighbors, where the chips located immediately above, below, to the left, and to the right of the given chip constitute its neighborhood. This leads to an *autologistic model* (Besag 1974), essentially used by Longtin et al. (1996) to model failures. Although relatively simple and tractable, this scheme fails to model any patterns other than those parallel to the sides of chips. Thus, only very basic and primitive shapes of clusters can be modeled.

An extension of the autologistic model is the *king-move neighborhood scheme*. In addition to the four adjacent chips, we let the neighborhood include the other four chips that are diagonal neighbors of a given chip so that

$$\mathcal{Q}_{ij} = \{(k, l) \neq (i, j) : |k - i| \leq 1,\ |l - j| \leq 1\}.$$

The resulting Markov-random-field model does not belong to the class of automodels, and its distribution has a more complicated form.

The appropriateness of a homogeneous model for the distribution of chips on a wafer may be argued, especially because of the edge effects. However, this model reflects the typical

spatial clustering that dominates other effects for most of the wafers (see Longtin et al. 1996, sec. 4.2). In the current production, the edge chips are not used at all, so they should not be included in the model.

The article is organized as follows. In Section 1, under the conditions of the Hammersley–Clifford theorem (Besag 1974), we obtain a closed form of conditional distributions of X_{ij}, given the neighborhood. The joint distribution is derived as well, but it is known to involve a normalizing constant whose closed form cannot be derived in general (Cressie 1991). Due to this constant, no standard estimation procedures can be used. However, this joint distribution can be embedded in a multiparameter exponential family whose canonical parameters have meaningful interpretations from the factorial-design-of-experiments point of view. They are estimated in Section 2 by the method of maximum pseudo-likelihood. A Gibbs sampler is then used to generate wafers with the same canonical parameters to be compared with the real ones. Section 3 contains analysis of real data in which we apply the proposed algorithm to a sample set of 388 wafers representing all the known root causes. Goodness of fit is analyzed in Section 4. Implementation of the proposed classification scheme by means of an artificial neural network is discussed in Section 5.

1. THE DISTRIBUTION MODEL

Based on the assumptions of a Markov random field with the defined above eight-unit neighborhood structure, here we derive the general form of the joint distribution of chips on a wafer.

Let $s = (i, j)$ denote coordinates of a chip, $t(s) = (i, j+1)$ denote the chip above s, and similarly, $b(s)$, $l(s)$, and $r(s)$ denote the chips below, to the left, and to the right of it, respectively. Further, let $rt(s)$ be the diagonally neighboring chip, located at the right top corner of s, and similarly, $rb(s)$, $lt(s)$ and $lb(s)$ (see Fig. 1).

A set of chips in which every two chips are neighbors is called a *clique*. For example, $\{s, t(s), r(s)\}$ is a clique, but $\{s, t(s), b(s)\}$ is not. The eight-unit neighborhood scheme provides 10 types of cliques, $\alpha, \lambda, \ldots, \kappa$ (Fig. 2). These 10 cliques determine 10 canonical parameters of the joint distribution. In the rest of the article, we will unambiguously use the same notation for cliques and for the corresponding parameters.

Let $\text{logit}(p) = \log(p) - \log(1-p)$. For every chip s and its subneighborhood $\mathcal{Q}'_s \subset \mathcal{Q}_s$, define *conditional logits* $\alpha_s(\mathcal{Q}'_s) = \text{logit}\big(\mathbf{P}\{X_s = 1 | X_t = 1 \text{ for } t \in \mathcal{Q}'_s, X_t = 0 \text{ for } t \in \mathcal{Q}_s \cap \mathcal{Q}'_s\}\big)$. We assume that a wafer is *stochastically homogeneous*; that is, $\alpha_s(\mathcal{Q}'_s)$ depends on the form of $\mathcal{Q}'_s$ but is independent of the location of a chip s. Without this assumption, the number of parameters exceeds the number of chips on a wafer so that parameters are no longer estimable. Then, let

$$\alpha = \alpha(\emptyset);$$
$$\lambda = \alpha(r) - \alpha = \alpha(l) - \alpha; \qquad \mu = \alpha(t) - \alpha = \alpha(b) - \alpha;$$
$$\nu = \alpha(rb) - \alpha = \alpha(lt) - \alpha;$$
$$\pi = \alpha(rt) - \alpha = \alpha(lb) - \alpha;$$

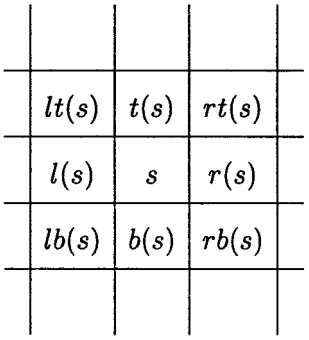

Figure 1. Chip s and its Neighborhood $\mathcal{Q}_s$.

$$\begin{aligned}
\eta &= \alpha(l,t) - \lambda - \mu - \alpha = \alpha(r,rt) - \lambda - \pi - \alpha \\
&= \alpha(b,lb) - \mu - \pi - \alpha; \\
\zeta &= \alpha(r,b) - \lambda - \mu - \alpha = \alpha(l,lb) - \lambda - \pi - \alpha \\
&= \alpha(t,rt) - \mu - \pi - \alpha; \\
\theta &= \alpha(l,b) - \lambda - \mu - \alpha = \alpha(r,rb) - \lambda - \nu - \alpha \\
&= \alpha(t,lt) - \mu - \nu - \alpha; \\
\iota &= \alpha(r,t) - \lambda - \mu - \alpha = \alpha(l,lt) - \lambda - \nu - \alpha \\
&= \alpha(b,rb) - \mu - \nu\alpha; \\
\kappa &= \alpha(r,t,rt) - \eta - \zeta - \iota - \lambda - \mu - \pi - \alpha \\
&= \alpha(l,t,lt) - \eta - \theta - \iota - \lambda - \mu - \nu - \alpha \\
&= \alpha(r,b,rb) - \theta - \zeta - \iota - \lambda - \mu - \nu - \alpha \\
&= \alpha(l,b,lb) - \eta - \zeta - \theta - \lambda - \mu - \pi - \alpha. \qquad (1)
\end{aligned}$$

Validity of these equalities is proved in Theorem 1. Here and in the following, the same letters $\alpha, \ldots, \kappa$ denote the cliques and the corresponding parameters.

Clearly α is the logit of a conditional probability of a failure given that all surrounding chips are good; μ is the *additional logit* due to the failure of one chip below or above (b or t); λ is the additional logit due to the failure of l or r; ν and π are additional logits due to the failure of diagonally adjacent chips. Furthermore, η is the additional logit due to the *simultaneous* failure of the chips above and to the left of s, and similarly, ζ, θ, and ι are additional logits due to the simultaneous failure of the other pairs of neighbors. Finally, κ represents the additional logit due to the simultaneous failure of three neighbors so that all four chips form a clique.

Using the terminology of general linear models, λ, μ, ν, and π are main effects of neighboring chips on a given chip; η, ζ, θ, and ι are second-order interactions; and κ is the third-order interaction. Hence, our choice of the eight-unit neighborhood scheme reflects the neglect of interaction effects of order higher than κ.

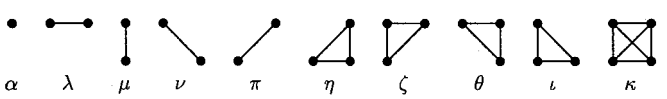

Figure 2. Ten Types of Cliques Under the Eight-Unit Neighborhood Scheme.

Corresponding to the 10 parameters, we introduce the following 10 statistics computed from a wafer:

$$\begin{aligned}
N_\alpha &= \sum X_{ij}, \quad N_\lambda = \sum X_{ij}X_{i+1,j},\\
N_\mu &= \sum X_{ij}X_{i,j+1}, \quad N_\nu = \sum X_{ij}X_{i+1,j-1},\\
N_\pi &= \sum X_{ij}X_{i+1,j+1}, \quad N_\eta = \sum X_{ij}X_{i,j+1}X_{i-1,j},\\
N_\zeta &= \sum X_{ij}X_{i,j-1}X_{i+1,j}, \quad N_\theta = \sum X_{ij}X_{i,j-1}X_{i-1,j},\\
N_\iota &= \sum X_{ij}X_{i,j+1}X_{i+1,j},\\
N_\kappa &= \sum X_{ij}X_{i,j+1}X_{i+1,j}X_{i+1,j+1}. \qquad (2)
\end{aligned}$$

Summations in (2) are taken over all chips on the whole wafer whose neighbors also belong to the wafer. Each of these statistics represents the number of cliques of the corresponding type in Figure 2, consisting entirely of defective chips. According to Theorem 1, this set of 10 statistics is a sufficient statistic for $(\alpha, \ldots, \kappa)$ if a so-called *positivity condition* holds for a wafer. The positivity condition means that wafers with no defective chips are theoretically possible; that is,

$$\mathbf{P}(\mathbf{0}) = \mathbf{P}\{X_s = 0 \text{ for every } s\} > 0. \qquad (3)$$

Consider a wafer $\mathbf{X}$ and a fixed chip s on it, and set $X_{s'} = 0$ for all chips s' outside of $\mathcal{Q}_s$. Then, for every clique $\mathcal{C} \in \{\alpha, \ldots, \kappa\}$, let $N_\mathcal{C}(s, X_s, \mathcal{Q}_s)$ be the resulting value of $N_\mathcal{C}$. It represents the number of interactions of each type inside the neighborhood $\mathcal{Q}_s$. Also, let $\Delta_\mathcal{C}(s, X_s, \mathcal{Q}_s) = N_\mathcal{C}(s, 1 - X_s, \mathcal{Q}_s) - N_\mathcal{C}(s, X_s, \mathcal{Q}_s)$. For a defective chip s, $\Delta_\mathcal{C}(s, X_s, \mathcal{Q}_s)$ is the negative of the number of type $\mathcal{C}$ cliques involving chip s. For a good chip s, $\Delta_\mathcal{C}(s, X_s, \mathcal{Q}_s)$ is the number of type $\mathcal{C}$ cliques obtained by replacing s with a defective chip.

Theorem 1. Suppose that the positivity condition (3) holds for a wafer. Then (a) equalities in (1) are valid, (b) the joint distribution of all chips X_s on a wafer has the form

$$\begin{aligned}
\mathbf{P}(\boldsymbol{X}) = \mathbf{P}(\mathbf{0}) \exp\{&\alpha N_\alpha + \lambda N_\lambda + \mu N_\mu + \nu N_\nu\\
&+ \pi N_\pi + \eta N_\eta + \zeta N_\zeta + \theta N_\theta + \iota N_\iota + \kappa N_\kappa\}, \qquad (4)
\end{aligned}$$

(c) conditional distributions of X_s have the form

$$\mathbf{P}\{X_s | \mathcal{Q}_s\} = \frac{1}{1 + \exp\{\alpha\Delta_\alpha(s, X_s) + \cdots + \kappa\Delta_\kappa(s, X_s)\}}, \qquad (5)$$

and (d) the set of 10 statistics (2) is a sufficient statistic.

2. PARAMETER ESTIMATION AND PATTERN RECOGNITION (THEORY)

Theorem 1 gives the canonical form of the joint distribution function of all chips on a wafer. By a sufficiency principle, estimation procedures should be based on $N_\alpha, \ldots, N_\kappa$. However, Expression (4) contains a normalizing constant,

$$\mathbf{P}(\mathbf{0}) = \left(\sum_W \exp\{\alpha N_\alpha + \lambda N_\lambda + \cdots + \kappa N_\kappa\}\right)^{-1},$$

which is the probability of having a wafer with no defective chips, or the proportion of 100%-good wafers. The summation is taken over all $2^{|W|}$ possible combinations of good and failed chips on a wafer W, and it has no closed form that can be used for practical computations. Therefore, only conditional distributions (5) can be used for the estimation of canonical parameters. In such situations, one uses the *maximum pseudo-likelihood estimator* of $\boldsymbol{\Theta} = (\alpha, \lambda, \mu, \nu, \pi, \eta, \zeta, \theta, \iota, \kappa)$,

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}} &= \arg\max_{\boldsymbol{\Theta}} \prod_{s \in W_0} \mathbf{P}\{X_s | \mathcal{Q}_s\}\\
&= \arg\min_{\boldsymbol{\Theta}} \prod_{s \in W_0} \left(1 + \exp\left\{\sum_\mathcal{C} \mathcal{C}\Delta_\mathcal{C}(s, X_s, \mathcal{Q}_s)\right\}\right). \qquad (6)
\end{aligned}$$

The set W_0 is usually chosen in such a way that all chips in it are conditionally independent of each other given the remaining chips ($W \cap W_0$). Under the king-move neighborhood scheme, the largest such set consists of all pixels with both even coordinates, $W_0 = \{(i, j) | (i/2, j/2) \in W\}$. This justifies the multiplication of conditional probabilities in (6).

However, especially for wafers with the smaller number of chips, one may consider the *generalized maximum pseudo-likelihood estimator* by taking $W_0 = W$. Although the chips in W_0 are no longer conditionally independent in this case, Comets (1992) proved consistency of $\widehat{\boldsymbol{\Theta}}$.

The resulting estimators $\hat{\alpha}, \hat{\lambda}, \ldots, \hat{\kappa}$ give the desired numerical description of patterns of defective chips on a wafer. Their interpretation is as follows. The "pure" failure rate is described by $\hat{\alpha}$. The probability that a chip fails due to any reason other than the spatial interaction with other chips is then estimated by $e^{\hat{\alpha}}/(1 + e^{\hat{\alpha}})$. In applications to semiconductor manufacturing, α is often found to be negative, which means that the effect of clustering dominates "pure" failures. Indeed, usually most of the defective chips on a wafer belong to some clusters, with only very few isolated ones—that is, with very few defective chips surrounded by only good chips. For such a wafer,

$$\mathbf{P}\{X_s = 1 | X_t = 0 \text{ for all } t \in \mathcal{Q}_s\} \approx 0$$

and

$$\mathbf{P}\{X_s = 0 | X_t = 0 \text{ for all } t \in \mathcal{Q}_s\} \approx 1,$$

which makes α negative and considerably large in absolute value.

Parameter estimates corresponding to two-unit cliques, $\hat{\lambda}, \hat{\mu}, \hat{\nu}$, and $\hat{\pi}$, describe shapes, sizes, thicknesses, and main directions of clusters of failed chips. High values of all four parameters correspond to a strong interaction between the adjacent chips, which results in large clusters. If these estimates are small, it is more likely to have small clusters scattered across the wafer.

Comparison of $\hat{\lambda}$ and $\hat{\mu}$, $\hat{\nu}$, and $\hat{\pi}$, shows the typical shape of clusters. If $\hat{\lambda} \approx \hat{\mu}$ and $\hat{\nu} \approx \hat{\pi}$, the clusters will be close to circular. However, if, say, $\hat{\lambda}$ is significantly greater than $\hat{\mu}$, it implies that percolation is stronger in the horizontal direction than in vertical, and the clusters tend to be narrow, stretched in a horizontal direction.

The larger in each pair $(\hat{\lambda}, \hat{\mu})$ and $(\hat{\nu}, \hat{\pi})$ shows the typical direction of clusters. More precisely, if $u = \max\{\hat{\lambda}, \hat{\mu}\}$, $v = \max\{\hat{\nu}, \hat{\pi}\}$, then the most typical direction can be "estimated" by a vector

$$\mathbf{z} = \frac{e^u \mathbf{z}_1}{1 + e^u} + \frac{e^v \mathbf{z}_2}{1 + e^v},$$

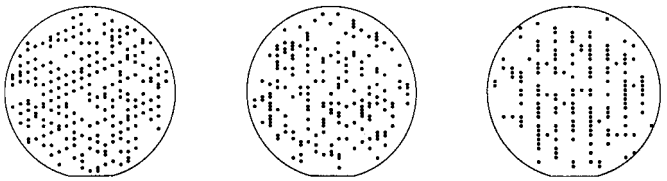

Figure 3. Three Wafers Generated by the Gibbs Sampler. The wafer on the left has strong diagonal interactions; parameters of the other two wafers are estimated from lots V and VIII, Figure 5, with different degrees of vertical interaction.

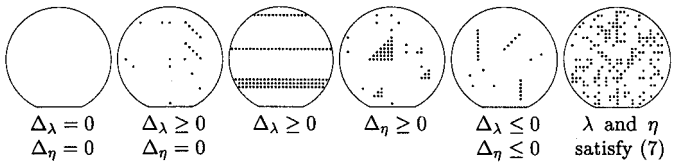

Figure 4. Patterns of Defective Chips Illustrating Cases 1, 2, and 3.

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are unit vectors on the plane, making the angle of 45°, with the directions determined by the cliques corresponding to u and v. These scenarios are illustrated by three wafers in Figure 3, generated by the *Gibbs sampler*. The joint distribution of the first wafer is a Markov random field with large parameters ν and π, which determine the dominating diagonal patterns. The vertical interaction dominates on the second wafer. This wafer was generated to be a Markov random field with the same parameters as were estimated from lot V depicted in Figure 5, Section 3. Parameters of the third wafer are estimated from lot VIII in Figure 5 with an even stronger vertical interaction. The only pairs of neighboring defective chips in lot VIII are vertical neighbors, so the maximum pseudo-likelihood estimates of all parameters except α and λ equal $-\infty$ and κ is not estimable (see Sec. 3, Cases 2 and 3). Thus, all the clusters on the generated wafer are stretched in the vertical direction, similar to the wafers in lot VIII.

3. PARAMETER ESTIMATION AND PATTERN RECOGNITION (PRACTICE)

The maximum pseudo-likelihood estimators obtained from (6) are not always unique and finite. Depending on the following conditions, different algorithms should be used.

Case 1 (minimax condition). We define this condition as

$$\max_C \min_{s \in W_0} \Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) < 0 \quad \text{and} \quad \min_C \max_{s \in W_0} \Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) > 0. \quad (7)$$

Condition (7) is sufficient for the existence of finite maximum pseudo-likelihood estimators. If it holds, then

$$\prod_{s \in W_0} \left(1 + \exp\left\{\sum_{\mathcal{C}} \mathcal{C} \Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s)\right\}\right) \to +\infty$$

as one or several parameters $\mathcal{C}$ tend to $\pm\infty$. Usually, this condition is not satisfied by the wafers with only a few defective chips or by the wafers with clearly marked patterns. For example, (7) is not satisfied by the lot VIII in Figure 5. These wafers have no defective cliques except the cliques of types α and μ. Therefore, one has $N_{\mathcal{C}} = 0$ for all cliques $\mathcal{C}$ except α and μ. For these cliques, $\Delta_{\mathcal{C}} \in \{0, 1\}$, which violates the first part of (7).

Case 2 (inestimable effect). If (7) does not hold and $\Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) = 0$ for all cliques of a certain type and all chips $s \in W_0$, then the maximum pseudo-likelihood estimator of $\mathcal{C}$ is not unique. From (6), $\widehat{\mathcal{C}}$ is any real number. Indeed, suppose that $\mathcal{C}$ is the additional logit due to a simultaneous failure of one or several neighbors and $\Delta_{\mathcal{C}} \equiv 0$ on a wafer. Then such a simultaneous failure of neighbors does not occur on a given wafer, which makes the effect of such failure inestimable. The other parameters, which satisfy (7), are still estimated in the usual manner, according to (6).

Case 3 (perfect pattern). It is also possible that $\Delta_{\mathcal{C}} \geq 0$ ($\Delta_{\mathcal{C}} \leq 0$) everywhere on the wafer, with $\Delta_{\mathcal{C}} > 0$ ($\Delta_{\mathcal{C}} < 0$) for at least one chip. If $\Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) \geq 0$ for all $s \in W_0$ and some clique type $\mathcal{C}$, then the formal minimization in (6) results in $\widehat{\mathcal{C}} = -\infty$, and if $\Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) \leq 0$ for all $s \in W_0$, then $\widehat{\mathcal{C}} = +\infty$.

Indeed, if $\Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) \geq 0$ for all s, then $N_{\mathcal{C}} = 0$. On such a wafer, a chip s belonging to a clique $\mathcal{C}$ never fails due to a simultaneous failure of the rest of $\mathcal{C}$. Moreover, if the rest of $\mathcal{C}$ fails, chip s has to be good. Hence, the corresponding conditional probability of a failure is 0, and both $\mathcal{C}$ and $\widehat{\mathcal{C}}$ equal $-\infty$. Similarly, if $\Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) \leq 0$ for all s, then $\mathcal{C} = \widehat{\mathcal{C}} = +\infty$. On such a wafer, a chip belonging to a clique $\mathcal{C}$ fails with probability 1 if the rest of $\mathcal{C}$ fails.

Thus, in the case of a perfect pattern, estimates of some parameters equal $\pm\infty$, which dominates the other effects. Still, the other parameters can be estimated according to (6) by minimizing the product of conditional probabilities over the set $W_0' = \{s \in W_0 : \Delta_{\mathcal{C}}(s, X_s, \mathcal{Q}_s) = 0 \text{ for all } \mathcal{C} \text{ such that } |\widehat{\mathcal{C}}| = \infty\}$.

Cases 2 and 3 are not unusual in wafer manufacturing. Fortunately, they are easy to recognize. A simple rule is given in Theorem 2 in terms of an *immediate subclique* of $\mathcal{C}$, which we define to be a cluster lacking exactly one chip to become a clique of type $\mathcal{C}$. The proof follows trivially from the preceding discussion.

Theorem 2. For a clique $\mathcal{C} \in \{\alpha, \ldots, \kappa\}$, one has (a) Case 2, if and only if $N_{\mathcal{C}'} = 0$ for all immediate subcliques of $\mathcal{C}$; (b) Case 3 with $\Delta_{\mathcal{C}} \geq 0$, if and only if $N_{\mathcal{C}} = 0$ and $N_{\mathcal{C}'} > 0$ for some immediate subclique of $\mathcal{C}$; (c) Case 3 with $\Delta_{\mathcal{C}} \leq 0$, if and only if there are no defective immediate subcliques of $\mathcal{C}$, which are not parts of a defective clique $\mathcal{C}$; and (d) Case 1, if and only if (a), (b), and (c) do not hold.

Six wafers shown in Figure 4 are computer-generated to illustrate these cases for the cliques λ and η.

Figure 5 illustrates 162 production wafers from eight lots with eight different root causes. For these wafers and all the training and validation wafers, the root causes have been determined by direct in-depth engineering analysis. Failed chips are depicted in black. Some patterns can be noticed by visual inspection, but others require more sensitive methods such as those proposed here. Although visual inspection is still practiced, it is no longer sufficient due to the volume of wafers produced every day.

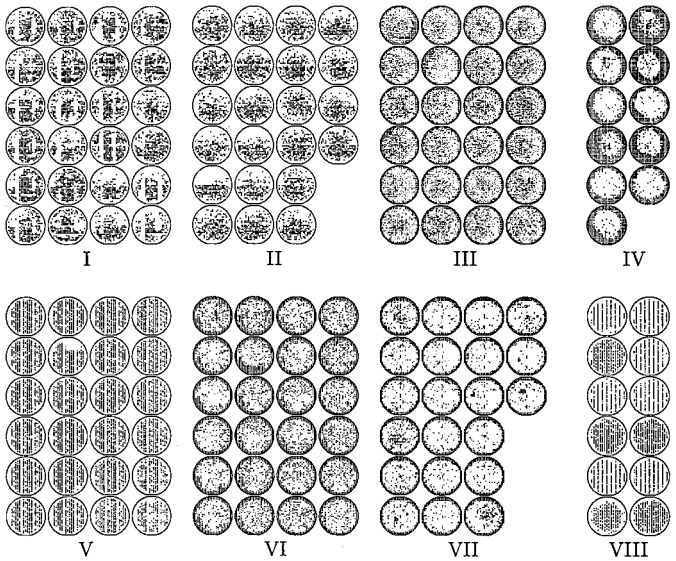

Figure 5. Wafer Maps Classified Into Eight Groups According to Root Causes.

Parameters of the Markov random field are estimated for each wafer according to (6), with the use of (5) to evaluate $\mathbf{P}\{X_s|\Omega_s\}$ for the mentioned set of sites s that are not neighbors of each other. Results are summarized in Table 1. Wafers with root causes I through VII satisfy the minimax condition, so all their parameters are estimable. Wafers from lot VIII have a perfect pattern—namely, a strong positive vertical correlation. Only α and μ have finite maximum pseudo-likelihood estimates for these wafers, and the parameter κ is not estimable because the wafers contain no defective immediate subcliques of κ. Negative values of $\hat{\alpha}$ imply that the effect of clustering dominates the "pure failure" of chips on a wafer.

4. MODEL COMPARISON AND THE GOODNESS OF FIT

In this section, we compare the performance of our estimates with the automodels used by Longtin et al. (1996). Our eight-unit neighborhood model has seven additional parameters, $\nu, \pi, \eta, \zeta, \theta, \iota$, and κ. The measure of improvement is therefore twice the difference of maximum log-pseudo-likelihood functions, which under the hypothesis of nonsignificance follows approximately the chi-squared distribution with 7 df (see Wilks 1938). Results of the corresponding χ^2 tests for 388 wafers are summarized in Table 2. The third row represents the proportion of wafers in which the extended model is significantly better (under the given level α) than the four-unit anisotropic model. Notice that given the set of chips $W \cap W_0$ (Sec. 2), the remaining chips are conditionally independent. Therefore, classical results about the limiting distribution of the likelihood ratio test statistic apply to the pseudo-likelihood ratios.

Results show strong significance of the extended model. For example, if the hypothesis is tested under the significance level of .005, then in 93.3% of all 388 wafers this hypothesis is rejected (instead of only .5% of wafers expected under the hypothesis).

Comparison of possible extensions of the automodel with respect to the Akaike information criterion shows that no model can be claimed to be the best for all the wafers. However, there is strong evidence of a better performance of high-order models. Therefore, for classification purposes, and especially because of the possibility of new patterns occurring in future, it is recommended to use the model with all 10 parameters.

Three wafers generated by the Gibbs sampler and depicted in Figure 3 also show how the parameter estimates capture and reflect the patterns of failed chips and clusters on a wafer.

5. IMPLEMENTATION: ARTIFICIAL NEURAL NETWORK

5.1 Back-Propagation Algorithm

An immediate application of the proposed model and the estimation scheme is classification of the produced wafers according to the known root causes. Maximum pseudo-likelihood estimators of 10 canonical parameters (call them signatures) are used as explanatory variables for the classification scheme. It is not feasible to find a mathematical expression relating these inputs and outputs (probable root causes), so an artificial neural network is used to establish the relationship. Training the neural network corresponds to finding the appropriate set of weights in an approximate relationship between the input–output pairs. We use the *back-propagation* method which usually achieves this objective given a sufficient number of training samples and correctly chosen input–output pairs. A detailed discussion of the back-propagation algorithm was given by Freeman and Skapura (1991).

Figure 6 is a schematic of the organization of the databases required for classification based on estimated values of canonical parameters of our model. Databases of failure types are the key to the implementation of the classification program. A file consisting of a lot number, wafer numbers, vector of estimated parameters, and corresponding failure mode is compiled.

Table 1. Maximum Pseudo-Likelihood Estimates for the Given Eight Root Causes

Root causes	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\nu}$	$\hat{\pi}$	$\hat{\eta}$	$\hat{\zeta}$	$\hat{\theta}$	$\hat{\iota}$	$\hat{\kappa}$
I	−2.52	1.15	1.34	.21	.19	.43	−.39	−1.30	−.22	.83
II	−2.46	1.06	1.34	.27	.69	−.66	−.04	−.12	−.08	−.28
III	−1.71	.51	.22	.26	.29	−.15	−.09	.10	.06	.49
IV	−3.39	.38	1.29	.11	−.31	2.82	.32	1.72	−.08	−2.82
V	−3.09	.77	1.65	−.09	1.47	−.99	−.50	.10	.16	−.38
VI	3.06	−.89	−1.54	−.61	−1.23	1.83	.41	.23	1.05	−.82
VII	−2.59	.64	.08	−.11	−.88	.85	1.52	.45	−.07	.28
VIII	−5.35	−∞	1.88	−∞	−∞	−∞	−∞	−∞	−∞	Undef.

Table 2. The Chi-Squared Analysis of the Maximum Log-Pseudo-Likelihood Functions

Significance level, α	.100	.050	.025	.010	.005
Critical values, χ^2_α, 7 df	12.017	14.067	16.013	18.475	20.278
Proportion of rejected hypotheses	.982	.977	.969	.951	.933

The setup of the algorithm essentially follows the guidelines recommended by Freeman and Skapura (1991). A multilayered feed-forward neural network with one hidden layer is chosen for training with the following characteristics:

1. The input layer consists of 10 elements, corresponding to the 10 estimated parameters of the model.
2. The data are scaled such that the domain of any given estimated parameter is the unit interval (0,1).
3. The training data consist of approximately 500 estimated patterns per root cause. Twenty percent of the training set was utilized for the validation set.
4. A network with 10 input layer nodes, 6 hidden layer nodes, and 8 output layer nodes was used. This architecture yielded the best performance. The chosen number of hidden layer units is one-third of the sum of input and output layer units, as recommended in the literature.
5. The hidden layer and output layer weights are generated from a Gaussian distribution with mean 0 and variance .2.
6. All bias units are set to 0 in the entire network.
7. The eight output classes are set to the identity matrix 8×8, the largest element (1) replaced by .9, and the smallest element (0) replaced by .1 according to Freeman and Skapura (1991).
8. The training set is used to train preliminary network architectures.
9. The validation set is used to identify the network with the lowest error sums of squares.
10. The network is trained using the pattern-by-pattern approach. The patterns are selected randomly from each root-cause category to eliminate any biases during learning.

5.2 Test Results

Estimated shape patterns from the eight root causes saved for validation of the neural network were submitted pattern by pattern. A total of 675 patterns of possible 780 were correctly classified, yielding a classification rate of 87%. The program is implemented on-line, with a graphical user interface residing on top of the database for quick and speedy analysis of semiconductor data.

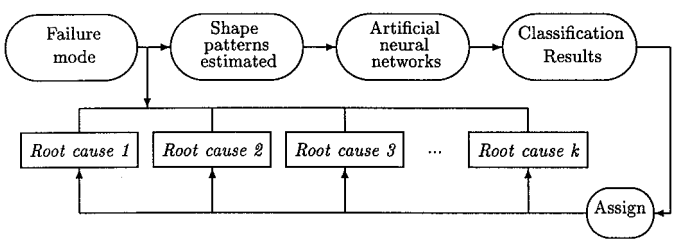

Figure 6. Pattern Recognition Program Layout.

6. SUMMARY AND CONCLUSIONS

A Markov random-field model with the king-move neighborhood structure is fitted to binary valued integrated circuit failures as the simplest and the lowest-rank anisotropic model explaining spatial interaction between the chips in all the directions. Analysis of 388 production wafers proves that this extension of earlier models is rather significant, providing a significantly better fit and classifying failures adequately.

Ten canonical parameters of the model characterized as the pure failure rate, pairwise, and higher-order interactions are estimated by the maximum pseudo-likelihood method. The obtained estimates are used as inputs for an artificial neural network, which classifies wafers by matching them to the root causes.

To implement the proposed scheme, the user must build a database consisting of a plethora of failure patterns with associated root causes. Graphical user interfaces atop databases of failures and a routine for statistical estimation in an automated fashion is a key to successful operation of advanced statistical methodologies in a dynamic manufacturing plant environment.

Implementation of this algorithm has reaped many benefits in a manufacturing environment. It aids the practicing engineer to identify potential root causes and associates specific failure patterns to probable root causes. The proposed classifier replaces visual inspection of wafer maps that is usually utilized in the semiconductor industry, guaranteeing incomparably faster inspection and a higher rate of correctly classified wafers.

ACKNOWLEDGMENTS

We thank the editor K. Kafadar, the associate editor, and two anonymous referees for their insightful comments, which led to an improved version of this article. Research of the first author is supported by the National Science Foundation grant 9818959 and the grant sponsored by Semiconductor Technical Council at Texas Instruments, Inc.

APPENDIX: PROOF OF THEOREM 1

According to the Hammersley–Clifford theorem (Besag 1974) the log-likelihood ratio of $\mathbf{P}(\boldsymbol{X})$ and $\mathbf{P}(\mathbf{0})$ has the form

$$\log \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\mathbf{0})} = \sum_{\mathcal{C}} G_{\mathcal{C}}(\mathcal{C}) \prod_{s \in \mathcal{C}} X_s, \tag{A.1}$$

where the sum is taken over all cliques $\mathcal{C}$ on a wafer. In the case of Bernoulli variables, if a clique contains at least one good chip, then the corresponding product in (A.1) is 0; otherwise it equals 1. The number of nonzero products is determined by $N_\alpha, \ldots, N_\kappa$. Hence, by homogeneity,

$$\log \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\mathbf{0})} = N_\alpha G_\alpha + N_\lambda G_\lambda + \cdots + N_\kappa G_\kappa. \tag{A.2}$$

Functions G can be found from the factorization theorem (Besag 1974; Cressie 1991), which relates the probabilities of two wafers $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\boldsymbol{Y})} = \prod_{i=1}^{n} \frac{\mathbf{P}\{X_i | X_j, j < i; Y_k, k > i\}}{\mathbf{P}\{Y_i | X_j, j < i; Y_k, k > i\}}, \tag{A.3}$$

where all chips are enumerated from 1 to n by a one-dimensional index i. The definition of a Markov random field allows restricting both conditions in the right side of (A.3) to the neighborhood of chip i.

We put $\boldsymbol{Y} = \mathbf{0}$ and let $\boldsymbol{X}$ be a wafer with a single defective chip s. From (A.2) and (A.3), we obtain

$$\log \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\mathbf{0})} = G_\alpha = \log \frac{\mathbf{P}\{1 | X_{s'} = 0,\ s' \neq s\}}{\mathbf{P}\{0 | X_{s'} = 0,\ s' \neq s\}} = \alpha.$$

Next, let $\boldsymbol{X}$ be a wafer with only two defective chips, s and $r(s)$. Then, from (A.2) and (A.3),

$$\begin{aligned} \log \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\mathbf{0})} &= 2G_\alpha + G_\lambda \\ &= \log \frac{\mathbf{P}\{1 | X_{r(s)} = 1, X_{s'} = 0, s' \neq s, s' \neq r(s)\}}{\mathbf{P}\{0 | X_{r(s)} = 1, X_{s'} = 0, s' \neq s, s' \neq r(s)\}} \\ &\quad + \log \frac{\mathbf{P}\{1 | X_{s'} = 0,\ s' \neq s\}}{\mathbf{P}\{0 | X_{s'} = 0,\ s' \neq s\}} = \alpha(r) + \alpha = \lambda + 2\alpha, \end{aligned}$$

from where $G_\lambda = \lambda$. Similarly, taking $\boldsymbol{X}$ to be a wafer with all good chips except one clique and applying (A.2) and (A.3), we obtain $G_{\mathcal{C}} = \mathcal{C}$ for all types of cliques $\mathcal{C}$, from α to κ. Along with (A.2), this proves (A.3).

The order in which the chips are enumerated in (A.3) is arbitrary. Computing the log-likelihood ratio with different orders produces alternative expressions for functions G and parameters in (1) and proves assertion (a).

Next, (4) defines a 10-dimensional *standard exponential family* of distributions (Brown 1986, chap. 1) with the canonical parameter $(\alpha, \ldots, \kappa)$. Therefore, assertion (d) follows directly from (b).

Since $\mathbf{P}\{X_s | \mathcal{Q}_s\}$ does not depend on the chips outside of $\mathcal{Q}_s$, without loss of generality, let $X_{s'} = 0$ for all s' outside of $\mathcal{Q}_s$. In this case

$$\mathbf{P}(\boldsymbol{X}) = \mathbf{P}(\mathbf{0}) \exp\{\alpha N_\alpha(s, X_s) + \cdots + \kappa N_\kappa(s, X_s)\}. \tag{A.4}$$

Hence, with $\boldsymbol{X}^{(s)}$ denoting the array of all chips except X_s,

$$\begin{aligned} \mathbf{P}\{X_s | \mathcal{Q}_s\} &= \mathbf{P}\{X_s | \boldsymbol{X}^{(s)}\} \\ &= \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\boldsymbol{X}^{(s)})} = \frac{\mathbf{P}(\boldsymbol{X})}{\mathbf{P}(\boldsymbol{X}^{(s)}, 0) + \mathbf{P}(\boldsymbol{X}^{(s)}, 1)}. \end{aligned} \tag{A.5}$$

All three probabilities in the right side of (A.5) can be evaluated using (A.4), and after simplification (c) follows.

[Received April 1998. Revised August 2000.]

REFERENCES

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Korean Statistical Society*, 2, 192–236.

Brown, L. D. (1986), *Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory* (Lecture Notes–Monograph Series, Vol. 9), Hayward, CA: Institute of Mathematical Statistics.

Comets, F. (1992), "On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice," *The Annals of Statistics*, 20, 455–468.

Cressie, N. A. C. (1991), *Statistics for Spatial Data*, New York: Wiley.

Freeman, J. A., and Skapura, D. M. (1991), *Neural Networks, Algorithms, Applications, and Programming Techniques* (Computation and Neural Systems Series), Reading MA: Addison–Wesley.

Hemmert, R. S. (1981), "Poisson Process and Integrated Circuit Yield Prediction," *Solid-State Electronics*, 24, 511–515.

Ketchen, M. B. (1985), "Point Defect Yield Model for Wafer Scale Integration," *IEEE Circuits and Devices Magazine*, 1, 24–34.

Longtin, M. D., Wein, L. M., and Welsch, R. E. (1996), "Sequential Screening in Semiconuctor Manufacturing, I: Exploiting Spatial Dependence," *Operations Research*, 44, 173–195.

Meyer, F. J., and Pradhan, D. K. (1989), "Modeling Defect Spatial Distribution," *IEEE Transactions on Computers*, 38, 538–546.

Shier, J. (1988), "A Statistical Model for Integrated-Circuit Yield With Clustered Flaws," *IEEE Transactions on Electronic Devices*, 35, 524–525.

Stapper, C. H. (1975), "On a Composite Model to the IC Yield Problem," *IEEE Journal of Solid-State Circuits*, ED-20, 655–657.

——— (1989), "Large Area Fault Clusters and Fault Tolerance in VLSI Circuits: A Review," *IBM Journal of Research and Development*, 33, 162–173.

Stapper, C. H., Armstrong, F. M., and Saji, K. (1983), "Integrated Circuit Yield Statistics," *Proceedings of IEEE*, 71, 453–470.

Taam, W., and Hamada, M. (1993), "Detecting Spatial Effects From Factorial Experiments: An Application from Integrated-Circuit Manufacturing," *Technometrics*, 35, 149–160.

Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62.