Sequential Screening in Semiconductor Manufacturing, II: Exploiting Spatial Dependence

Mark D. Longtin, Lawrence M. Wein

and Roy E. Welsch

WP# 3452-92-MSA July, 1992

SEQUENTIAL SCREENING IN SEMICONDUCTOR MANUFACTURING, II: EXPLOITING SPATIAL DEPENDENCE

Mark D. Longtin

Operations Research Center, M.I.T.

Lawrence M. Wein

Sloan School of Management, M.I.T.

and

Roy E. Welsch

Sloan School of Management, M.I.T.

Abstract

This paper addresses the same quality management problem as Ou and Wein (1992), except that here screening is performed at the chip level, rather than at the wafer level. We analyze over 300 wafers from two industrial facilities and use a Markov random field model to capture the spatial clustering of bad chips. Chip screening strategies are proposed that exploit the various types of yield nonuniformities that are detected in the data, such as radial effects, spatial clustering of bad chips, and yield variation by chip location. The numerical results suggest that screening at the chip level is significantly more profitable than screening at the wafer level.

July 1992

SEQUENTIAL SCREENING IN SEMICONDUCTOR MANUFACTURING, II: EXPLOITING SPATIAL DEPENDENCE

Mark D. Longtin

Operations Research Center, M.I.T.

Lawrence M. Wein

Sloan School of Management, M.I.T.

and

Roy E. Welsch

Sloan School of Management, M.I.T.

This paper considers the same problem as Ou and Wein (1992): choose a wafer start rate for the fabrication facility and a sequential screening policy before the electrical testing, or *probing* facility, to maximize the expected revenue from nondefective chips minus the variable fabrication and probing costs, subject to average effective capacity constraints on the fabrication and probing facilities. Readers are referred to the Introduction of Ou and Wein for a motivation of this problem and a detailed discussion of yield modeling. Whereas Ou and Wein screen at the wafer level to exploit lot-to-lot variability, we will screen at the chip level to exploit the detailed spatial (intra-wafer) and temporal (inter-wafer) yield dependencies within a lot of wafers.

There are two reasons why screening at the chip level leads to a much more formidable mathematical problem than screening at the wafer level. First, an underlying yield model at the chip level is required that captures the dependence between chips. Such a model would presumably be more intricate than the yield model used in Ou and Wein, which did not explicitly model the dependence between chips on a given wafer. Also, the space of possible screening policies is much richer at the chip level than at the wafer level. Ou and Wein assume that the wafer yields within a given lot are iid, and the screening decision reduces to the binary choice of probing (we will hereafter use the term *testing* rather than probing) another wafer or discarding the remaining wafers in the lot. In this paper, the decision maker must decide which, if any, of the untested chips on the wafer to test next. Due to the yield

dependence between chips, such a decision could presumably depend on the yield and the position of all previously tested chips on the wafer (or even the lot).

Hence, in contrast to Ou and Wein, we will not attempt to find an optimal solution to the problem posed at the beginning of this paper. Rather, our two goals are to identify the yield nonuniformities within a lot that are most prevalant in industry, and to develop simple screening strategies (unlike Ou and Wein, we will use the terms *strategy* and *policy* interchangeably) that effectively exploit these nonuniformities. We analyze 41 *wafer maps* (see the Appendix for some typical maps, which graphically depict the good and bad chips on a wafer) from a *development fab* that principally develops new products, and 275 wafer maps, which represent six lots of wafers, from a *production fab* that employs more established technology. Exploratory data analysis shows a strong radial dependence (the yield drops sharply at the outer edges of the wafer) and a moderate amount of spatial clustering of defective chips on the wafer maps from the development fab. The wafers from the production fab exhibit strong spatial clustering and very little radial dependence. In addition, a small amount of temporal dependence across wafers is present in both fabs, where certain chip locations have poor yield throughout the lot.

Although we have no plans to explicitly solve a mathematical problem, there are several important reasons to develop and validate a probabilistic model for chip yield. First, most semiconductor companies are very secretive about their yield figures, making it difficult for academic researchers to obtain yield data. Also, gathering yield data is a very computationally intensive undertaking: the 718,058 chips analyzed in this paper represent no more than one month's worth of output for a fab of typical size. Hence, if a model can be found that fits the data well, then it can be used by practitioners and researchers as a basis for both simulation and analytical studies. Finally, a decision variable in the problem posed here is to choose a start rate of wafers, and an explicit yield model is useful for *predictive performance analysis*: for example, given a screening policy and a yield model, estimate the start rate of wafers that causes the probe facility to work at exactly its effective capacity.

We employ a two-dimensional Markov random field to model the chip yield on a wafer. This model has been used extensively in statistical mechanics and allows the probability

of a chip being defective to depend upon the yield of the neighboring chips, where the neighborhood can be arbitrarily defined. Parameter estimation and goodness-of-fit tests show that the model appears to be a reasonably good fit for the wafer maps from the production fab and the homogeneous portion (that is, excluding the outer edges) of the wafer maps from the development fab; in particular, the null hypothesis that the chip yields on a wafer are iid Bernoulli random variables is rejected for nearly 90% of the wafers.

We identify a variety of chip screening strategies, including a class of strategies suggested by the Markov random field, that attempt to exploit the radial dependence, the spatial clustering, the temporal dependence, or some combination thereof. Their performance on the actual wafer maps is very impressive: our best policy in each fab performs nearly as well as a "clairvoyant" optimal policy. For example, 10% of the chips are discarded and they have an average yield below 3%, although the average incoming yield ranges from 50% to 80%.

Our numerical study assumes that an *adaptive* start rate is used; that is, the decision maker observes the average number of chips tested per wafer for a particular screening policy, and then chooses the optimal start rate of wafers. In practice, we would expect the decision maker to dynamically change the start rate, perhaps on a weekly basis, in response to the actual screening results. However, to aid in the determination of a wafer start rate, we attempt to predict the average number of chips tested per wafer for a variety of screening strategies. Exact results are derived for some simple policies, and for more complex policies, we resort to analytical bounds and tables generated by simulating Markov random fields.

The remainder of the paper is organized as follows. The problem is described in Section 1 and a preliminary analysis is performed in Section 2. Markov random field models are introduced in Section 3 and are used to model the industrial data in Section 4. The proposed chip screening policies are described in Section 5 and are tested on the industrial data in Section 6. The prediction of the optimal start rates also appears in Section 6, and concluding remarks about our study and the companion study in Ou and Wein can be found in Section 7.

1. Problem Description

This section contains a description of the problem. A precise mathematical formulation of the problem requires an underlying yield model, which is not given here, although a yield model will be proposed in Section 4. We retain much of the notation introduced in Ou and Wein, although some of the variables will be in different units (e.g., lots versus wafers). Readers are referred to Figure 1 for a description of the process flow. Wafers enter the fab at a start rate of λ wafers per week, where λ is a decision variable. The fab's effective capacity is μ_F wafers per week and a variable cost c_F is incurred for each wafer produced; it is assumed that any start rate above μ_F would lead to an unacceptably high level of work-in-process inventory. A fraction q of the wafers are scrapped during fabrication and each wafer contains M chips. Hence, chips enter the testing facility at rate $\lambda(1-q)M$ per week.

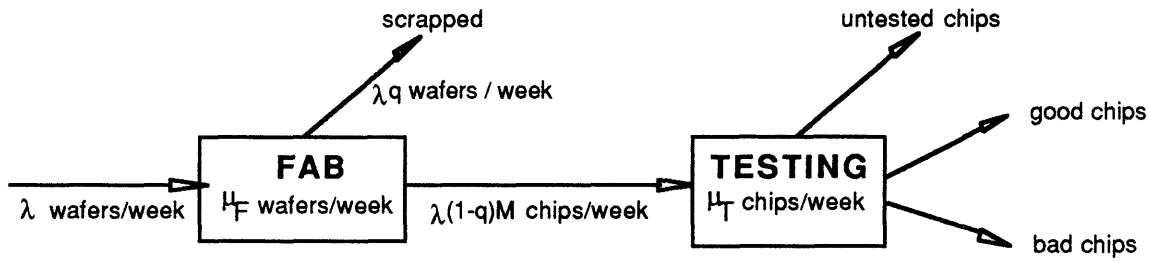


Figure 1. The semiconductor manufacturing facility.

The effective testing capacity is μ_T chips per week, and so we are assuming that the testing time per chip is a constant under any testing strategy. This assumption is not entirely accurate for three reasons, and tends to inflate the profit improvements that are reported in Section 6. The actual testing time per wafer consists of a setup time of perhaps 30 minutes plus the testing time of one to two minutes per chip, plus the travel time between chips. First, we are ignoring the setup time that is incurred before each wafer is tested. However, since a typical wafer contains hundreds or even thousands of chips (see Table II in Section 4), this assumption has minimal consequences. Second, we are assuming that the average testing time per chip is independent of the screening policy in use, and hence we are ignoring the travel time between chips. Even for our most complex policy, which makes three passes

over a wafer, the travel time is negligible compared to the actual testing time, and thus the omission of travel time is not serious. Third, we are assuming that the mean testing time of a chip does not depend on whether the chip is good or bad. However, the testing of a chip actually consists of a series of tests, and as soon as a chip is deemed to be defective, then the tester moves on to the next chip. Consequently, the mean testing time per bad chip may be somewhat less than the mean testing time per good chip. Based on our conversations with industry people, it appears that the mean testing time of a bad chip is at least half of the mean testing time of a good chip. Under the assumption that the mean testing time of a bad chip is exactly half that of a good chip, we reran some of our experiments, and found that the profit increase was reduced by a factor that was slightly less than one-half. For example, the 10.7% profit increase achieved by the mixed strategy in Table XIII was reduced to a 6.4% profit increase.

A screening strategy S determines which chips to probe and which chips to leave untested and discard. A third feasible option that is not considered here is to leave a chip untested and send it directly to packaging. Hence, in our model, one of three things can happen to a chip at testing: (i) the chip is tested at a cost c_T and found to be good, in which case it is sold and a revenue r is received; (ii) the chip is tested at a cost c_T and found to be defective, and is consequently discarded; or (iii) the chip is discarded and never tested. The screening policy can use any or all information about chips that have been probed thus far. The screening strategy S, along with the configuration of good and bad chips on the wafers exiting the fab, determines the expected fraction of chips tested, f_S, and the expected yield of the tested chips, Y_S.

Our goal is to choose the start rate λ and the screening strategy S to maximize expected profit subject to the effective capacity constraints. This optimization problem can be expressed as

$$\max_{\lambda,S} \quad rM(1-q)\lambda f_S Y_S - c_F\lambda - c_T M(1-q)\lambda f_S \tag{1}$$

$$\text{subject to} \quad 0 \le \lambda \le \mu_F \tag{2}$$

$$M(1-q)\lambda f_S \le \mu_T. \tag{3}$$

Parameter	μ_F	μ_T/M	q	$c_F/(rM)$	c_T/r
Value	0.9	0.7695	0.05	0.1	0.003

Table I. System parameters for our numerical study.

Two points bear repeating: (i) problem (1)-(3) is not precisely formulated because we have not defined a yield model that would essentially dictate the values of f_S and Y_S for any strategy S; and (ii) for any interesting yield model, problem (1)-(3) would be extremely difficult to solve, due to the huge number of possible screening strategies and the difficulty in determining f_S and Y_S for a given strategy. However, the optimality conditions for (1)-(3) will be analyzed in the next section to gain some initial insights.

2. Preliminary Analysis

Problem (1)-(3) is generic: by appropriate choice of system parameters, it could represent the situation at any semiconductor facility. For the sake of concreteness, and to develop insights for our numerical study, we hereafter assume that the system parameters take on the values given in Table I; these values are chosen to maintain consistency with Ou and Wein. By constraints (2)-(3), it follows that under the exhaustive testing policy ($f_S = 1$) that is commonly used in industry, the testing facility will be the bottleneck if $\mu_T < \mu_F M(1-q)$. In our example, this inequality simplifies to $0.81 < 0.9$, so that under exhaustive testing, the testing facility is the bottleneck and the fab cannot operate above $0.81/0.9 = 90\%$ of its effective capacity.

We first consider optimality conditions on λ. The objective function (1) can be factored as

$$\Pi(\lambda, S) = \lambda(rM(1-q)f_S Y_S - c_F - c_T M(1-q)f_S), \tag{4}$$

and thus the optimal value λ^* must be either the maximum possible value or zero depending upon whether the coefficient of λ in (4) is positive or negative. In particular, if we consider the exhaustive testing strategy, then $\lambda^* \geq 0$ if and only if the average yield is greater than

or equal to

$$\frac{c_F + c_T M(1-q)}{rM(1-q)} = 0.108.$$

Therefore, the fab must achieve an average yield of at least 10.8% in order to turn a profit on a marginal cost basis. We hereafter assume that this condition holds, and hence the optimal start rate is

$$\lambda^* = \min(\mu_F, \frac{\mu_T}{M(1-q)f_S}). \tag{5}$$

Note that λ^* will assume the second argument of the minimum in equation (5) provided that

$$f_S \geq \frac{\mu_T}{\mu_F M(1-q)} = 0.9. \tag{6}$$

Hence, probe will be at its effective capacity if it is optimal to discard no more than 10% of the chips.

We now derive the conditions under which a given screening policy A is preferable to another policy B. Without loss of generality, we assume that $f_A \geq f_B$; that is, policy B discards more chips on average than policy A does. Two cases will be considered. First, suppose $f_B < f_A < \mu_T/(\mu_F M(1-q)) = 0.9$. In this case, the optimal starting rate λ^* is μ_F for both A and B, and policy B is better than policy A if

$$Y_{A-B} \triangleq \frac{f_A Y_A - f_B Y_B}{f_A - f_B} < \frac{c_T}{r} = 0.003. \tag{7}$$

The left side of inequality (7) will be referred to as the *yield of the marginal untested chips.* If policies A and B are nested (that is, B discards all the chips that A discards), then Y_{A-B} is the fraction good of the chips that are tested under screening policy A but discarded under policy B. Thus, it pays to leave these marginal chips untested if their yield is so poor as not to cover the variable testing costs. Since 0.3% is a very low marginal yield, we expect condition (7) to be almost always violated in practice. Hence, very rarely will it be optimal to test less than 90% of the chips, and, as a consequence, have the testing facility work strictly below its effective capacity.

In the case where $f_A > f_B > \mu_T/(\mu_F M(1-q)) = 0.9$, the optimal start rate is $\mu_T/(M(1-$

$q)f_S$), and policy B is better than policy A if

$$Y_B - Y_{A-B} > \frac{K}{f_B}, \tag{8}$$

where $K = c_F/(rM(1-q)) = 0.1053$. Thus, it pays to leave the marginal chips untested if the marginal yield, Y_{A-B}, is *worse* than the average yield of the chips that have been tested so far, Y_B, by at least K/f_B (roughly 11%). We expect this condition to hold quite often (see Table XIV in Section 6 for marginal and average yields of various strategies), and consequently it will rarely be optimal to test more than 90% of the chips.

In summary, it will often be optimal to test 90% of the chips on average, and discard the "worst" 10% of the chips. For any given screening strategy, the start rate λ will be set to its maximum feasible value. Thus, our goal is to find variations in yield that will allow us to simultaneously identify and discard chips that are likely to be bad.

3. Markov Random Fields

The chips on a typical wafer are laid out in a rectangular grid pattern, or *lattice*. A wafer is represented by a matrix of binary random variables $X = \{X_{i,j}\}$, where $X_{i,j} = 1$ if the chip at lattice point (i,j) is good, and $X_{i,j} = 0$ if the chip at (i,j) is bad. Although it is clear from the semiconductor yield literature (see the Introduction of Ou and Wein) that defects, and hence defective chips, tend to cluster together, Flack (1985) has proposed the only model that explicitly takes into account the spatial dependence of chips on a wafer. She lets each chip site (i,j) have an associated binomial random variable, and then defines the number of point defects on the chip in site (i,j) to be the sum of its associated binomial random variable and the binomial random variables associated with 20 neighboring chip sites. Since a chip is good if it contains no point defects, this model captures spatial correlations in the yield of neighboring chips. In contrast to Flack, we model chip yield directly (that is, without modeling point defects) and propose a binary Markov random field for representing the number and location of good chips on a wafer. After the model is described, we discuss parameter estimation and the simulation of wafers.

3.1. Model Description

Markov random fields (MRFs) - or equivalently, Gibbs distributions - have been used extensively in statistical mechanics, dating back to the celebrated Ising model of ferromagnetism (Ising 1925). More recently, they have been successfully applied to image processing (see, for example, Geman and Geman 1984, Besag 1986 and Derin and Elliott 1987). Readers are referred to Besag (1974) for a lucid description of the model used in this paper.

A Markov random field is a probability model for describing uncertain quantities on a lattice X. In particular, we consider here the binary MRF model, where each lattice point (i,j) has a random variable $X_{i,j}$ that takes on the values 0 or 1. For each site (i,j), the model specifies a *neighborhood* of nearby sites, and each of these sites are referred to as a nearest neighbor of site (i,j). The MRF possesses the Markov property that X_{ij} depends on the rest of the lattice only through its nearest neighbors. We will begin by focusing on a four nearest neighbor model with two parameters β_0 and β_1, where

$$P(x_{i,j} \mid \text{ rest of lattice}) = \frac{\exp[x_{i,j}(\beta_0 + \beta_1(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1}))]}{1 + \exp[\beta_0 + \beta_1(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1})]}. \tag{9}$$

If a boundary of zeroes is assumed to surround the lattice, the joint distribution for the entire lattice is (see Besag 1972)

$$P(x) = \frac{1}{Z(\beta_0, \beta_1)} \exp\left(\sum_{i,j} (\beta_0 x_{i,j} + \beta_1 x_{i,j}(x_{i-1,j} + x_{i,j-1}))\right), \tag{10}$$

where

$$Z(\beta_0, \beta_1) = \sum_{\text{all possible } x} \exp\left(\sum_{i,j} (\beta_0 x_{i,j} + \beta_1 x_{i,j}(x_{i-1,j} + x_{i,j-1}))\right).$$

$Z(\beta_0, \beta_1)$ is a normalizing constant referred to as the *partition function*. If the rectangular lattice has m rows and n columns, then Z is the summation of 2^{mn} terms, one for each of the possible realizations of the random matrix X. Hence, for any lattice of size greater than 6×6, the function Z is essentially uncomputable.

The parameter β_1 measures the clustering effect. If $\beta_1 = 0$, then the conditional dis-

tribution (9) implies that each $X_{i,j}$ is an iid Bernoulli random variable. If $\beta_1 > 0$, then configurations where the ones are clustered together are more likely. Similarly, if $\beta_1 < 0$, configurations where ones are dispersed from each other are more likely. If β_1 exceeds a certain critical value, then the MRF exhibits a form of long range dependence known as a phase transition (see Pickard 1987). In this case, the MRF is no longer ergodic, and parameter estimation becomes problematic; see Gidas (1991) for a detailed discussion. In our study, 17 of the 316 wafers in our study have parameter estimates that exceed this critical value.

Two extensions of (9) will be considered so that more complicated types of spatial dependence can be captured. The *non-isotropic* MRF allows the spatial interactions along the horizontal and vertical axes to differ in strength, and is characterized by

$$P(x_{i,j} \mid \text{rest of lattice}) = \frac{\exp[x_{i,j}(\beta_0 + \beta_1(x_{i-1,j} + x_{i+1,j}) + \beta_2(x_{i,j-1} + x_{i,j+1}))]}{1 + \exp[\beta_0 + \beta_1(x_{i-1,j} + x_{i+1,j}) + \beta_2(x_{i,j-1} + x_{i,j+1})]}. \qquad (11)$$

A second generalization results by also considering interactions with diagonal terms. The *eight nearest neighbor* MRF conditional distribution is identical to (11), with the addition of the term $\beta_3(x_{i+1,j+1} + x_{i-1,j+1} + x_{i-1,j-1} + x_{i+1,j-1})$ to the β_0, β_1, and β_2 terms.

3.2. Fitting the Model

To model semiconductor wafers with MRFs, we need a reliable method to estimate parameters. Unfortunately, the inability to compute the partition function $Z(\beta)$ precludes maximum likelihood estimation as a practical approach. However, several methods have been developed to attempt to overcome this estimation problem, and readers are referred to Comets and Gidas (1992) for references. We employ the *maximum pseudo-likelihood* method developed by Besag (1975), where the likelihood function (10) is replaced by the pseudo-likelihood likelihood function

$$f(X) = \prod_{i,j} \pi_{i,j}^{x_{i,j}} (1 - \pi_{i,j})^{(1-x_{i,j})},$$

where

$$\pi_{i,j} = \frac{\exp[\beta_0 + \beta_1(X_{i-1,j} + X_{i+1,j} + X_{i,j-1} + X_{i,j+1})]}{1 + \exp[\beta_0 + \beta_1(X_{i-1,j} + X_{i+1,j} + X_{i,j-1} + X_{i,j+1})]}.$$

Although this simple method assumes that all the lattice sites are independent of one another, it provides consistent estimates (see Geman and Graffigne 1987) and simulation studies (see Strauss 1991 for references) suggest that it loses little efficiency compared to the maximum likelihood estimates, provided that the MRF is below the phase transition point.

After the MRF parameters are estimated, the goodness-of-fit must be assessed. Since the maximum pseudo-likelihood technique is essentially a logistic regression, the goodness-of-fit tests for logistic regression can be applied. In particular, we may wish to test whether the MRF model provides a significantly better fit than a Bernoulli model, where the chip yields on the wafer are iid Bernoulli random variables. This test is equivalent to testing whether the inclusion of the β_1 term in model (9) significantly improves the fit. More generally, we can test whether the addition of one additional covariate significantly improves the fit of the logistic regression. If the simpler model is true, then the *deviance* (twice the increase in the pseudo-likelihood function) is distributed approximately as a chi-squared random variable with one degree of freedom; see McCullagh and Nelder (1983) for details. Hence, a simple χ^2 test can be performed.

Since the maximum pseudo-likelihood method provides only an approximate goodness-of-fit test, we also employ the *generalized Monte Carlo significance test* developed by Besag and Clifford (1989), which allows exact tests to be carried out. This technique is best illustrated by an example. Suppose we wish to test the hypothesis that the simple MRF (9) describes the data x versus the null hypothesis that the x are Bernoulli. As can be seen from equation (10), $\sum_{i,j} x_{i,j}$ and $u = \sum_{i,j} x_{i,j}(x_{i-1,j} + x_{i,j-1})$ are sufficient statistics for the simple MRF model, whereas $\sum_{i,j} x_{i,j}$ is a sufficient statistic for the Bernoulli model. Starting with the realization x, we visit each site (i,j) in turn and permute the value of $x_{i,j}$ with the value of a random site $x_{i',j'}$. Each permutation leaves the value of $\sum_{i,j} x_{i,j}$ unchanged but may change the value of u. In this way, we generate many different realizations $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(R)}\}$ of the matrix x that are equally likely under the null hypothesis. To test the null hypothesis, the actual value of u is compared with the empirical distribution $\{u^{(1)}, u^{(2)}, u^{(3)}, \ldots, u^{(R)}\}$

under the null hypothesis. The null hypothesis is rejected at the significance level M/R, where M is the number of realizations i such that $u \geq u^{(i)}$ (in a one-tailed test).

More generally, this Monte Carlo technique can be used to test any two nested spatial models. This involves constructing permutations that leave the sufficient statistics of the null hypothesis unchanged, and allow the additional sufficient statistic u of the alternative hypothesis to vary. Performing the hypothesis test entails comparing the actual value of u with the empirical distribution of u generated under the null hypothesis.

3.3. Simulating Markov Random Fields

There are several instances in Section 6 where we need to generate simulated wafers, which can be achieved by drawing random samples from a given MRF. Due to the mutual dependence inherent in the MRF model, this cannot be done directly. We employ the *Gibbs sampler* algorithm for simulating a MRF, which was developed by Geman and Geman (1984), who drew upon a general method developed by Metropolis et al. (1953).

The Gibbs Sampler algorithm works in the following manner. Given a configuration $x^{(t)}$, visit each site (i,j) in turn and update its value $x_{i,j}$ by drawing a sample from the conditional distribution of $x_{i,j}$ (equation (9) for a simple MRF). After each site is updated, we have a new realization $x^{(t+1)}$. In this manner the algorithm generates a sequence of matrices $\{x^{(0)}, x^{(1)}, x^{(2)}, \ldots\}$ that forms a Markov chain. It is possible to show that, regardless of the initial configuration $x^{(0)}$, the steady state distribution of the Markov chain is the same as the distribution of the MRF (10).

Because successive matrices are not independent, generating many independent random samples from the same distribution requires some care; Gelman and Rubin (1991) outline several of the pitfalls. Typically, a large number of iterations is required for the Markov chain to approach the steady state distribution. Since it is difficult to determine the number of iterations required to reach steady state from a single Markov chain, Gelman and Rubin suggest using multiple Markov chains, each with a different initial configuration. Only when each of these chains "converges" to the same distribution has steady state been reached. We found it necessary to first skip a large number of iterations (on the order of 10,000) to reach

steady state, and then only take every N^{th} instance of the Markov chain, where N is on the order of 1000; this allowed us to produce approximately independent random samples from any given Markov Random Field.

4. Analysis of the Wafer Maps

The data analyzed in this paper come from two very different semiconductor fabs. The first fab, which will often be referred to as the D fab, is a relatively low volume *development* facility producing a diverse set of leading edge components with forefront technologies. As a result, the D fab typically experiences low yields by industry standards, as it continually redesigns and reintroduces the latest high performance devices. The second fab, referred to as the P fab, is a higher volume *production* facility producing commodity chips. It applies better understood technology to a relatively stable product mix, and consequently achieves higher yields than the D fab.

We analyze 275 wafer maps (consisting of six lots denoted by $P1, \ldots, P6$) from the P fab and 41 wafer maps from the D fab. These 41 wafers are of the same product (i.e., chip type), and although they were not produced as a lot, they will sometimes be collectively referred to as the D lot. Also, with the exception of lots $P3$ and $P4$, the six lots at the P fab correspond to different products. A *wafer map* is a graphical depiction of the good and bad chips on a wafer; see the Appendix for six sample wafer maps. Due to restrictions on space, these six wafers have been selected as a representative sample of the 316 wafers in this study. When presenting results, we will give exact results for these six wafers and a summary of the results for all 316 wafers. Blank locations in the interior of the wafer correspond to either fiducial areas, which are reference marks used to align the wafers during the various processing steps, or test chips, which contain test patterns used for process control. Neither fiducials nor test chips are included in the yield calculations. Since the D fab uses two inch wafers and the P fab uses four inch wafers, the P wafers have many more chips on them. Detailed information about the wafers is summarized in Table II. In the remainder of this section, we conduct an exploratory data analysis and then use MRFs to model these wafers.

Lot	No. Wafers	Wafer Diameter (mm)	Chip Size (mm)	Average Chips/Wafer	Average Yield
D	41	50	0.8×3.0	627	50.8%
$P1$	47	100	2.6×1.6	1800	89.7%
$P2$	44	100	2.4×1.6	1975	87.5%
$P3$	45	100	1.6×1.6	2786	82.9%
$P4$	47	100	1.6×1.6	2807	45.6%
$P5$	46	100	1.5×1.3	3720	87.5%
$P6$	46	100	1.9×1.8	2009	92.9%

Table II. Wafers in this study.

4.1. Exploratory Data Analysis

The goal of our exploratory data analysis is *to find variations in yield.*

Yield variation by lot. Table II shows that lot $P4$ has much lower yield than the other five P lots. Most of this lot-to-lot variability cannot be accounted for by the different product mix, since lots P3 and P4 both consist of the same product.

Yield variation by wafer. Under a Bernoulli model, where the number of good chips on each wafer in the lot is distributed as a binomial random variable with parameters (p, M), the variance in the yield of a wafer is $p(1-p)/M$. As can be seen in Table III, all of the lots show a yield variation by wafer much greater than that predicted by a Bernoulli model. This high degree of wafer-to-wafer variability is also seen in Example 1 of Albin and Friedman (1989) and in Bohn (1991).

Yield variation by radial distance. The yield of a chip location for a given lot is the number of wafers in the lot that have a good chip at the location divided by the number of wafers in the lot that have a chip at the location. For the D lots, Figure 2 plots the yield of each chip location as a function of the distance between the chip location and the center of the wafer. This figure reveals a drastic reduction in yield beyond 17 mm from the wafer center. In contrast, yield is a relatively constant function of radial distance for five of the six P lots, except for the outer 5mm of the wafer; see the figures in Appendix B of Longtin (1992). The lone exception is the low yielding lot P4, which exhibits a nearly linear

Lot	p	M	$(p(1-p)/M)^{1/2}$	Actual σ
D	50.8	627	2.00	10.19
$P1$	89.7	1800	0.67	4.53
$P2$	87.5	1975	0.66	4.44
$P3$	82.9	2786	0.68	9.60
$P4$	45.6	2807	0.94	24.31
$P5$	87.5	3720	0.51	10.33
$P6$	92.9	2009	0.47	1.95

(All results above are in percentages.)

Table III. Yield variation by wafer.

relationship between yield and radial distance; for this lot, the yield at the wafer edge is about half as high as the yield at the wafer center.

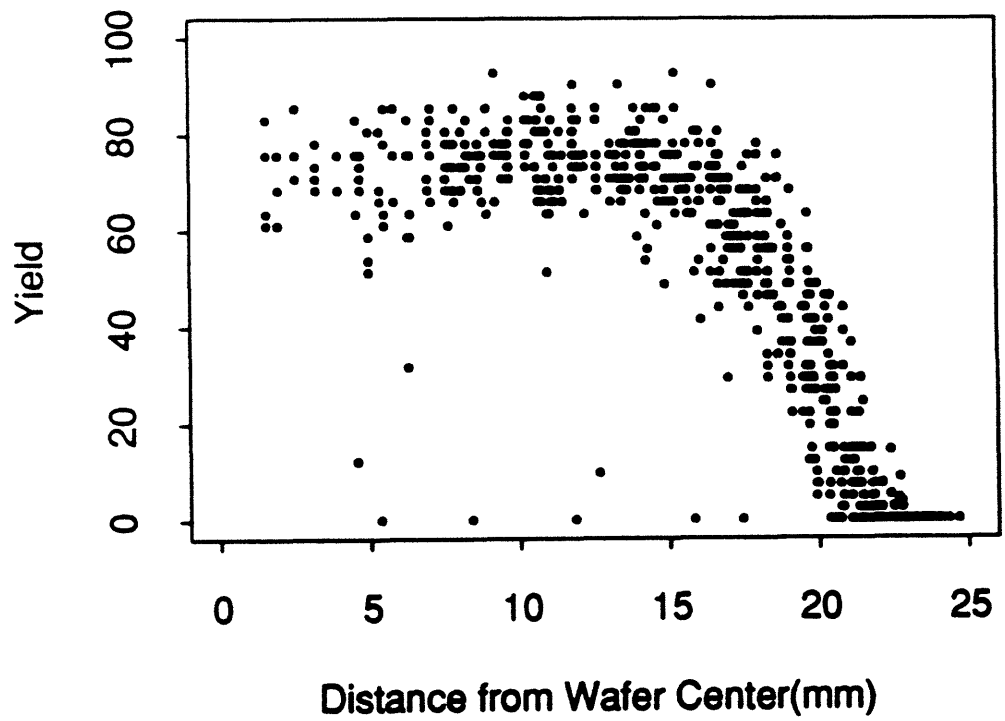


Figure 2. D Yield vs. radial distance by location.

Yield variation by location. Figure 2 also uncovers several randomly placed "outlier" locations with yield close to zero. To investigate this phenomenon further, we define a *hole*

as a zero yield location that is at least 5mm from the edge of the wafer. The percentage of holes on each of the seven lots is shown in Table IV. Under the Bernoulli model, the probability that no holes will occur on any of the seven lots is over 0.999999, and hence the existence of these holes represents a serious departure from the Bernoulli hypothesis. Further investigation shows that between 40%–50% of the holes on lot $P3$ also appear at the same locations on lot $P4$. Therefore, some of the holes may be due to causes that affect all lots of a given product (such as mask defects), and other holes may be due to causes that vary from lot to lot.

Lot	D	P1	P2	P3	P4	P5	P6
Holes/Chips	0.8%	3.9%	0.9%	1.7%	1.4%	2.7%	0.9%

Table IV. Percentage of holes by lot.

Yield variation by nearest neighbor yield. As can be seen from the wafer maps in the Appendix, bad chips tend to cluster together. The rows labelled "Actual" in Table V display the yield of all the chips that have n out of four nearest neighbors good, where $n \in \{0, 1, 2, 3, 4\}$. Only the inner 16mm of the D wafers are included to avoid confusing nearest neighbor dependence with radial dependence. This table shows that the spatial dependence of any chip on the four adjacent chips is very strong: a chip surrounded by four bad chips is often bad, and a chip surrounded by four good chips is usually good. However, some of this dependence may be due to the fact that wafer yield varies considerably within a lot, and hence good chips would be more likely to be surrounded by other good chips because good chips are more likely to appear together on the same wafers. To distinguish between wafer-to-wafer variability and spatial clustering, we suppose the yield of chips on each wafer i is distributed according to a Bernoulli model with parameter p_i, where the parameter for each wafer equals the wafer's actual yield. Under this hypothesis, the yield of chips in a lot with n nearest neighbor chips good is approximately

$$\overline{Y}_n = \frac{\sum_i p_i^{n+1}(1-p_i)^{n-4}}{\sum_i p_i^{n}(1-p_i)^{n-4}},$$

Lot	Method	No. of Neighbors Good				
		0	1	2	3	4
D	Bernoulli	45.4	57.2	68.4	76.0	80.7
	Actual	20.3	31.8	62.6	81.1	86.7
$P1$	Bernoulli	66.0	74.9	86.2	90.7	91.8
	Actual	32.0	56.8	77.2	89.9	93.4
$P2$	Bernoulli	78.6	82.1	86.2	89.3	91.2
	Actual	18.8	38.3	67.0	88.0	95.1
$P3$	Bernoulli	52.6	63.7	76.8	84.5	87.8
	Actual	4.5	37.7	70.0	86.0	94.1
$P4$	Bernoulli	14.5	38.7	53.4	64.0	71.4
	Actual	1.2	26.7	57.1	83.1	95.6
$P5$	Bernoulli	37.4	51.6	77.4	88.8	91.5
	Actual	5.9	49.2	72.3	89.5	93.6
$P6$	Bernoulli	93.1	93.7	94.3	94.9	95.4
	Actual	26.7	45.8	77.0	93.0	96.6

Table V. Yield by nearest neighbor yield: Bernoulli model vs. actual.

where the summations are over the wafers in the lot; we have approximated the expected value of the ratio by the ratio of the expected values, which is quite accurate since each lot contains nearly 50 wafers. The large discrepancy in Table V between the actual yield and the approximate Bernoulli yield shows that the probability a chip is good is highly dependent upon the yield of its nearest neighbors.

We also consider average spatial first-order autocorrelations, which allow us to look for correlation between adjacent chip sites in the $\hat{\imath}$, $\hat{\jmath}$, or $\hat{k}$ directions, where $\hat{\imath}$ and $\hat{\jmath}$ correlations represent dependencies along the two axes of a wafer, and $\hat{k}$ correlations represent dependencies *across adjacent wafers.* Average spatial first-order autocorrelations for each lot are calculated in the following manner. Consider each of the K wafer maps $x = \{x_{i,j}\}_{i=1,\ldots,I;j=1,\ldots,J}$ as being stacked one on top of another so as to form an array of dimension $I \times J \times K$. The average first-order autocorrelation along the $\hat{\imath}$ direction is given by

$$\overline{\rho_{\hat{\imath}}} = \mathrm{avg}_{j,k}\left(\frac{\sum_i (x_{i,j,k} - \overline{x_{j,k}})(x_{i-1,j,k} - \overline{x_{j,k}})}{\sum_i (x_{i,j,k} - \overline{x_{j,k}})^2}\right),$$

where $\overline{x_{j,k}} = \mathrm{avg}_i(x_{i,j,k})$. The corresponding quantities for the $\hat{\jmath}$ and $\hat{k}$ directions are calcu-

Lot	$\overline{\rho_{\hat{\imath}}}$	$\overline{\rho_{\hat{\jmath}}}$	$\overline{\rho_{\hat{k}}}$
D	0.095	-0.068	-0.006
$P1$	0.042	0.051	-0.015
$P2$	0.215	0.150	0.003
$P3$	0.308	0.170	-0.029
$P4$	0.503	0.527	0.103
$P5$	0.214	0.144	-0.005
$P6$	0.388	0.170	-0.002

Table VI. Average spatial autocorrelations.

lated in an analogous manner. These results, which are presented in Table VI, show that spatial dependencies in the $\hat{\imath}$ and $\hat{\jmath}$ are quite strong for lots $P2$–$P6$. Correlations across wafers in a lot $\overline{\rho_{\hat{k}}}$ appear to be negligible (which is expected for the D wafers, since these wafers were not processed together), with the exception of lot $P4$.

4.2. Fitting the Markov Random Field Models

Given the strong evidence of the spatial clustering of defective chips on a wafer, the Markov Random Field seems like a promising model for chip yield. It is important to note, however, that the MRF will not capture all the yield nonuniformities that were found in the previous section. Since the MRF is stationary (translation-invariant), it will capture neither yield variation by radial distance nor yield variation by location (we will return to these yield variations later). Nonetheless, for most of the wafers, particularly the P fab wafers, the spatial clustering appears to dominate the other yield dependencies.

We now fit the simple Markov Random Field (9) to a *window* of locations on each wafer map using the maximum pseudo-likelihood technique outlined in Section 3. For the P fab, the window consists of all locations (i,j) such that (i,j) and its eight nearest neighbors are on the wafer for all wafers in the lot. Due to the strong radial dependence of yield in the D lot, we further restrict the window to the inner 16mm of these 41 wafers.

Table VII shows the maximum pseudo-likelihood parameter estimates β_0 and β_1 and the significance levels α, for the selected wafers in the Appendix. The parameter β_1 is positive for all six of the wafers, indicating positive spatial dependence. What is perhaps

Wafer	β_0	β_1	Δd.f.	χ^2	α
$D.26$	-0.7247	0.5110	1	9.3	0.0022
$D.34$	-1.3715	0.4219	1	6.2	0.0131
$P1.47$	-0.5095	0.4232	1	85.8	< 0.0001
$P2.44$	-0.8554	0.9277	1	80.8	< 0.0001
$P3.2$	-2.0807	1.2193	1	948.9	< 0.0001
$P4.28$	-3.4702	1.5947	1	1787.0	< 0.0001

Table VII. MRF parameter estimates for selected wafers.

	Mean Estimates		No. reject H_0
Lot	β_0	β_1	at $\alpha < 0.05$
D	-0.425	0.589	25/47
$P1$	0.347	0.583	46/47
$P2$	-1.144	1.033	44/44
$P3$	-1.430	1.073	45/45
$P4$	-3.658	1.397	44/47
$P5$	-0.062	0.702	46/46
$P6$	-0.142	0.887	37/46

Table VIII. Average MRF Parameter Estimates for All Wafers.

most striking about these results is not that the Bernoulli model is overwhelmingly rejected for wafer $P4.28$ (P fab, 4^{th} lot, 28^{th} wafer), where the spatial dependence is obvious to the naked eye, but that the Bernoulli model is overwhelmingly rejected for wafers $D.26$ or $P1.47$, where the spatial dependence is not so obvious. Table VIII gives the average MRF parameter estimates over all the wafers in each lot, and shows the number of wafers for which we can reject the null hypothesis (Bernoulli) with 95% confidence or greater. Nearly all of the P wafers and over half of the D wafers exhibit significant spatial dependence, and the magnitude of this dependence is larger in the P fab. Only four of 275 P wafers and three of 41 D wafers have a negative β_1 coefficient, and none of these seven are significantly negative at the 90% confidence level.

Since the maximum pseudo-likelihood method does not produce unbiased parameter estimates, the Monte Carlo significance test (using 1000 samples) described in Section 3.2 is also applied to the wafer maps. The results of these tests for the six selected wafers are

Wafer	MPL α	Monte Carlo α
$D.26$	0.0022	4/1000—11/1000
$D.34$	0.0131	15/1000—19/1000
$P1.47$	< 0.0001	$< 1/1000$
$P2.44$	< 0.0001	$< 1/1000$
$P3.2$	< 0.0001	$< 1/1000$
$P4.28$	< 0.0001	$< 1/1000$

Table IX. Monte Carlo goodness-of-fit test.

shown in Table IX. Although the maximum pseudo-likelihood technique tends to somewhat overstate the confidence level (i.e., underestimate α), the Bernoulli model can still be rejected with very high confidence.

We now consider several extensions to the simple MRF model to see whether a better fit can be found. Since the maximum pseudo-likelihood technique will be used for parameter estimation, our comparisons of the different versions of the MRF will be done using the chi-squared test discussed in Section 3.2. Since the chips on a wafer are not always square (see Table II), we might expect the strength of the spatial dependence to differ in the two directions. The results of fitting the non-isotropic MRF (11) to the six selected wafers can be found in Table C.1 of Longtin, and a summary of the non-isotropic MRF estimates for all wafers is given in Table X. Note that for lot D, where the chips are nearly four times as long in the $\hat{\jmath}$ direction as the $\hat{\imath}$ direction, the spatial interactions are much stronger along the $\hat{\imath}$ (β_1)

	Mean Estimates			No. reject H_0
Lot	β_0	β_1	β_2	at $\alpha < 0.05$
D	-0.326	0.694	0.431	11/41
$P1$	0.359	0.582	0.578	2/47
$P2$	-1.113	1.080	0.973	13/44
$P3$	-1.425	1.139	1.005	9/45
$P4$	-3.659	1.539	1.260	16/47
$P5$	-0.045	0.720	0.675	7/46
$P6$	0.020	1.032	0.662	9/46

Table X. Non-isotropic MRF parameter estimates.

direction, as one might expect. In contrast, no systematic relationship is apparent between the estimated paramater values and the chip geometry for the P lots. The last column of Table X gives the fraction of wafers for which the null hypothesis $H_0 : \beta_1 = \beta_2$ (i.e., the isotropic MRF) can be rejected. Overall, the non-isotropic MRF provides a significantly better fit than the isotropic MRF for about 20% of the wafers.

The next model we consider is the eight nearest neighbor MRF, which allows interactions with the four diagonal nearest neighbors as well as the $\hat{\imath}$ and $\hat{\jmath}$ nearest neighbors. Table C.2 of Longtin gives the parameter estimates for the six selected wafers and Table XI presents the mean MRF parameter estimates for all wafers, along with the fraction of wafers for which the null hypothesis $H_0 : \beta_3 = 0$ (i.e., the four nearest neighbor non-isotropic MRF) can be rejected. The null hypothesis is rejected for the overwhelming majority of the P wafers. In contrast, the β_3 interaction for the D wafers appears to be negligible, and the four nearest neighbor model is rejected with 95% confidence on only one of the 41 wafers, less than what one would expect from chance alone.

	Mean Estimates				No. reject H_0
Lot	β_0	β_1	β_2	β_3	at $\alpha < 0.05$
D	-0.272	0.664	0.400	0.020	1/41
$P1$	-1.015	0.420	0.420	0.557	46/47
$P2$	-1.819	0.870	0.765	0.413	35/44
$P3$	-2.155	0.837	0.667	0.536	44/45
$P4$	-4.108	1.018	0.609	0.863	45/47
$P5$	-1.060	0.550	0.492	0.465	46/46
$P6$	-0.731	0.862	0.458	0.393	26/46

Table XI. 8 nearest neighbor MRF parameter estimates.

The Markov random field variants we have considered so far are defined on a two-dimensional lattice. It is also possible to define a MRF on a three-dimensional lattice, by adding an interaction term for cross-wafer dependence to the two dimensional MRF model (9). This model could then be used to model an entire lot of wafers, viewed as a three dimensional lattice. However, the three-dimensional MRF does not appear to be an appropriate model for representing chip yield. It requires every wafer in a lot to have the

same expected yield and the same amount of spatial dependence, whereas our exploratory analysis in Section 4.1 found significant yield variability by wafer and Table VII displays significant differences in spatial dependence. Although we do not report the results here, the three-dimensional MRF provides a worse fit for all of the lots in this study than the simple two-dimensional isotropic MRF model defined in (9).

We conclude this section by commenting on the relative complexity of the various models. If a lot consists of K wafers, then K parameters are needed to characterize the Bernoulli model. In contrast, the original MRF model (9) requires $2K$ parameters, the non-isotropic model requires $3K$ parameters and the eight nearest neighbor model requires $4K$ parameters. Among the models considered here, the non-isotropic model provides the best fit for the wafers from the D fab, and the eight nearest neighbor model provides the best fit for the wafers from the P fab. Hence, these two extensions to the original MRF model provide a significantly better fit (particularly at the P fab) at the expense of some additional complexity. In this light, the relatively poor performance of the three-dimensional MRF is hardly surprising, since it contains only four parameters. Finally, Section 4.3.2 of Longtin considers the simple isotropic MRF (9), except that the spatial dependence parameter β_1 is assumed to be identical for every wafer in the lot. This allows a lot of K wafers to be characterized by only $K+1$ parameters. Table 4.11 of Longtin shows that the simple Bernoulli model is rejected (at $\alpha = 0.05$) in favor of the constant$-\beta_1$ MRF model for 94.5% of the P wafers, and the constant$-\beta_1$ model is rejected in favor of model (9) for 40.4% of the P wafers. Hence, the constant$-\beta_1$ MRF offers a significant improvement in fit over the Bernoulli model at the expense of only one additional parameter.

5. Screening Strategies

In this section, we develop screening strategies that attempt to exploit the yield nonuniformities found in Section 4. The eight strategies described below will be evaluated on the actual wafers in Section 6. Note that the decision of whether to test or discard a particular chip under some of our testing strategies can depend upon neighboring chips that are off the edge of the wafer or part of a non-chip area. For purposes of implementation, any missing

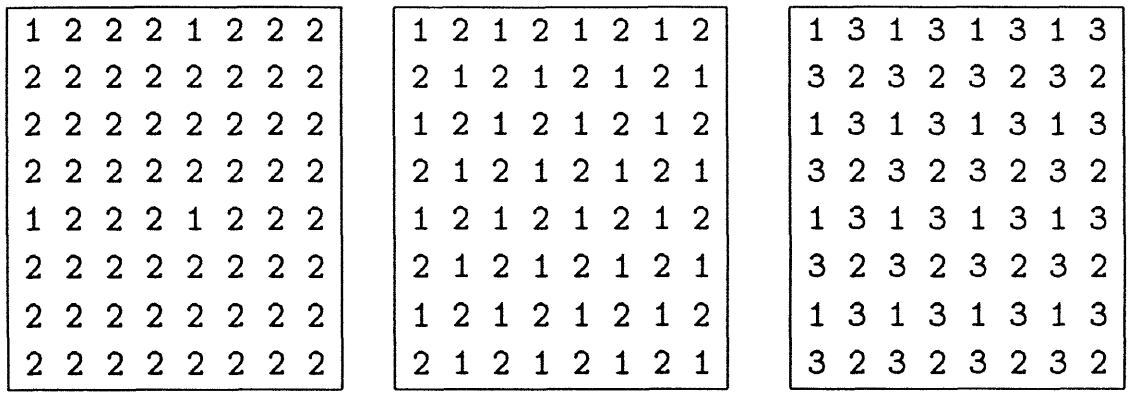


Figure 3. Chip testing patterns.

nearest neighbor chip is assumed to be bad.

1. E: *Exhaustive testing.* Test every chip. This testing strategy is the one commonly used in industry and will serve as a benchmark to compare with other strategies.

2. V: *Clairvoyant.* Test only those chips that are good, and leave all bad chips untested. This strategy does not represent an actual achievable screening strategy. Rather, as an ideal testing strategy, its purpose is to provide an upper bound on the performance of all other strategies.

3. $R(r)$: *Radial.* Test those chips within r mm of the center of the wafer, and discard the remaining chips. This strategy attempts to exploit radial nonuniformity in yield.

4. $W(y)$: *Wafer screening.* First test all the chips labeled "1" in Figure 3; these chips represent 1/16 of the chips on the wafer. If the yield of the "1" chips is above y, then test all of the remaining chips (the "2" chips) on the wafer. If the yield of these sample chips is at or below y, then leave the "2" chips untested. This simple strategy attempts to exploit wafer-to-wafer yield variability.

5. $C(n)$: *Checkerboard.* Make two passes over the wafer. On the first pass, test every other chip in a predetermined checkerboard pattern; that is, test all the chips labeled "1" in Figure 3. On the second pass over the wafer, consider the remaining ("2")

chips; notice that each of these "2" chips has four nearest neighbors that were tested during the first pass. If more than n ($n = 0, 1, 2, 3$) of these neighboring chips are good, then test the "2" chip. If the number of good nearest neighbor "1" chips is less than or equal to n, then leave the "2" chip untested. Thus, from 50% to 100% of the chips on the the wafer will be tested. This strategy is motivated by the MRF and attempts to exploit the spatial clustering of defective chips. Besag (1974) employs this checkerboard pattern in his coding method for MRF parameter estimation.

6. $M(n_1, n_2)$: *Metacheckerboard.* Make three passes over the wafer. On the first pass, test every fourth chip in a predetermined grid pattern; that is, test every chip labelled "1" in Figure 3. On the second pass, visit all the "2" chips, which are diagonally adjacent to the "1" chips. If the number of good diagonal nearest neighbor "1" chips is more than n_1 (out of four), then test the "2" chip. If the number of good diagonal nearest neighbor "1" chips is less than or equal to n_1, then leave the "2" chip untested. At this point, one half of the chips (in a checkerboard pattern) have been either tested or discarded. On the third pass, visit all the remaining ("3") chips. If the number of good *and tested* nearest neighbor "1" and "2" chips is more than n_2 (out of four), then test the "3" chip. If the number of good and tested nearest neighbor "1" and "2" chips is less than or equal to n_2, then discard the "3" chip. Like the checkerboard strategy, this strategy also attempts to exploit spatial clustering. The checkerboard strategy never discards more than 50% of the chips in a given region, whereas the metacheckerboard strategy has the advantage of discarding up to 75% of the presumably bad chips. Thus, it can be expected to perform well when the spatial clustering of defective chips is particularly strong.

7. $S(a, b)$: *Sequential.* For each wafer, visit each location (i, j) on the wafer in turn. For each location (i, j), keep a running total of the number $N_{i,j}$ of chips that were tested at this location on previous wafers in the lot, and the number $B_{i,j}$ of these $N_{i,j}$ chips that were bad. For a given location, if

$$B_{i,j} < ab + (1 - b)N_{i,j}, \tag{12}$$

where $a > 0$ and $0 < b < 1$, then the chip at this location is tested on the current wafer. If, on the other hand, inequality (12) does not hold, then location (i, j) is left untested on the current wafer *and all subsequent wafers* in the lot. This strategy attempts to exploit yield variation by location, and is inspired by the classical sequential sampling technique developed by Wald (1947). The parameter a represents the minimum number of chips at a given location that will be tested before rejection, and as the number of chips tested gets very large, a new chip will be rejected only if the average yield for this location is less than b.

8. $SM(a, b; n_1, n_2)$: *Mixed.* The last strategy is a combination of the metacheckerboard and sequential strategies. For each wafer, visit each location (i, j), and check the sequential strategy acceptance condition (12). If this condition is violated for (i, j), discard the chip at (i, j) on this wafer and all subsequent wafers. Test every "1" chip on the metacheckerboard pattern (see Figure 3), except those that have been discarded by condition (12). Test only those "2" chips that have not been discarded by condition (12) and are surrounded by at least n_1 good tested diagonally adjacent chips. Discard the remaining "2" chips. Test only those "3" chips that have not been discarded by condition (12) and are surrounded by at least n_2 good tested adjacent chips. Discard the remaining "3" chips. Finally, for each location (i, j), increment $N(i, j)$ by one if the chip at (i, j) was tested and increment $B(i, j)$ by one if the chip at (i, j) was also found to be bad.

6. Numerical Results

In this section, the screening strategies described in the last section are applied to the actual wafers from the two fabs. The strategies are assessed in the context of the optimization problem (1)–(3), which maximizes expected profit over possible wafer start rates λ and screening strategies S, subject to congestion constraints. The performance of a given strategy S can be summarized by two variables: the expected fraction of chips tested, f_S, and the expected yield of the tested chips, Y_S. By (5), the value of f_S is required to determine the optimal wafer start rate λ^* for a given strategy S. However, for five of the eight strategies in

Section 5, the true value of f_S is very difficult to estimate in advance. In Section 6.1, we allow an *adaptive* start rate, where each strategy S is used on the actual wafers and the average fraction of chips tested, f_S, is observed; the start rate λ^* is then chosen using (5) and the resulting profit Π is calculated using (1). We believe that the adaptive case is a reasonable representation of industrial practice in a production fab, since the fraction of chips tested can be frequently observed, and hence an accurate estimate for f_S should be readily available. In Section 6.2, the *nonadaptive* case is considered, where the decision maker must choose the start rate before any testing is performed. Here, we address the problem of estimating f_S, and hence estimating λ^* via (5), for several particular strategies. Readers who are interested in only a summary of the numerical results may proceed to Section 6.3.

6.1. Adaptive Start Rates

In this subsection, calculated profits are reported separately for the D fab wafers and the P fab wafers; no attempt is made to find a separate optimal strategy for each of the six different lots at the P fab. We discuss each strategy's performance in detail below.

Base cases. We begin by considering the two base cases: the exhaustive strategy (denoted by E) commonly used in industry and the clairvoyant strategy V that represents an upper bound on achievable performance. For the D fab, Table XII records four quantities for each strategy S that are expressed in percentage terms: the fraction of tested chips, the yield of the tested chips, the yield of the untested chips, and the profit improvement over exhaustive testing. The table also records the optimal start rate λ_S^* for each policy S. Recalling that the effective capacity of the fab is 0.9, we can interpret $\lambda_E^* = 0.81$ to mean that *the fab is operating at 90% of its effective capacity* under the exhaustive testing policy. The clairvoyant strategy V increases the start rate to 100% of the fab's effective capacity, resulting in an 11.5% increase in profit over strategy E.

Table XIII gives the P fab results for each strategy, for all six lots taken as a whole. The clairvoyant strategy achieves an 11.2% profit increase, and the discrepancy between 11.5% and 11.2% is due to the larger amount of testing in the P fab. Whereas Table XIII includes only the best performance achieved by a given strategy, Table XIV considers

Strategy	Yield of Chips Tested	Yield of Chips Discarded	Fraction Tested	Start Rate	Increase in Profit
Exhaustive Testing E	50.8	—	100.0	0.810	0.0%
Wafer Screening $W(0.35)$	53.2	27.6	90.6	0.894	3.3%
Metacheckerboard $M(0,0)$	61.6	12.0	78.1	0.900	4.0%
Checkerboard $C(0)$	57.4	5.8	87.1	0.900	9.2%
Sequential $S(12,.05)$	56.4	5.5	88.9	0.900	9.5%
Radial $R(23.5)$	56.5	0.8	89.8	0.900	11.0%
Clairvoyant V	100.0	0.0	50.8	0.900	11.5%

Table XII. D fab profitability results.

Strategy	Yield of Chips Tested	Yield of Chips Discarded	Fraction Tested	Start Rate	Increase in Profit
Exhaustive Testing E	79.7	—	100.0	0.810	0.0%
Radial $R(48)$	81.5	53.5	93.5	0.866	1.6%
Checkerboard $C(1)$	84.8	10.4	93.1	0.870	6.4%
Wafer Screening $W(0.48)$	85.2	25.3	90.8	0.892	6.5%
Sequential $S(2,0.6)$	85.7	15.8	91.4	0.886	7.3%
Metacheckerboard $M(0,1)$	87.4	8.7	90.2	0.898	9.6%
Mixed $SM(10,0.05,0,0)$	89.3	2.8	88.9	0.900	10.7%
Clairvoyant V	100.0	0.0	79.7	0.900	11.2%

Table XIII. P fab profitability results.

strategies under several parameter values, to shed some insight into the nature of these policies. Table XIV provides the fraction of tested chips (f), the yield of tested chips (Y), and the yield of discarded chips (Y_{E-S}) for all six lots together (column "all P") and each lot taken separately (columns 1–6).

Radial strategy. Table XII shows that the radial strategy is extremely effective on the D wafers, discarding chips that are over 99% bad. This strategy is able to achieve an 11% profit increase, which is nearly equal to the theoretical upper bound of 11.5%. The performance of the radial strategy on the P wafers is much less impressive. Although the strategy is more profitable than exhaustive testing, Table XIII shows that more than 50% of the chips it discards are good.

Wafer screening strategy. The wafer screening strategy increases the yield of tested

chips in the D fab by only 2.4% over the exhaustive testing strategy, and achieves a modest 3.3% profit increase. On the P wafers, the strategy rejects 22 wafers from lot $P4$, and one each from lots $P3$ and $P5$, resulting in a 6.5% increase in profit. In both fabs, the yield of discarded chips is over 25%. Longtin shows that the wafer screening strategy performs nearly as well as a clairvoyant wafer screening strategy that tests all of the chips on wafers with yield above a certain cutoff value y, and none of the chips on wafers with yield below y.

Checkerboard strategy. Table XII shows that the checkerboard strategy $C(0)$ works well on the D wafers, resulting in a 9.2% profit increase. Although this strategy is exploiting nearest neighbor dependence, to some extent it is also exploiting radial dependence. Because the $C(0)$ strategy already discards too many ($12.9\% = 1 - f$) chips, we do not consider the other checkerboard strategies. More generally, a strategy that deterministically or randomly mixes between two checkerboard strategies $C(n)$ and $C(n+1)$ (where $C(-1)$ represents the exhaustive testing strategy) can be found that discards exactly 10% of the chips; although such a policy will usually outperform a pure checkerboard strategy, this avenue is not pursued here. Similar comments apply to the metacheckerboard strategy.

The best checkerboard strategy for the P fab is $C(1)$, which achieves a 6.4% profit increase in Table XIII. Like the wafer screening strategy, the checkerboard strategy discards mostly chips from lot $P4$. In Table XIV, we also report on the $C(0)$ and $C(2)$ strategies. The $C(0)$ strategy only discards chips with four bad nearest neighbors, which have a yield of only 2.4%. However, these chips only constitute 5.4% of all chips from the P fab, and hence only a 5.6% profit increase is achieved. The $C(1)$ checkerboard strategy discards 6.9% of the chips, but the yield of the incremental chips that $C(1)$ discards but $C(0)$ does not discard, denoted by $Y_{C(0)-C(1)}$ in Table XIV, is a rather high 39%. The $C(2)$ strategy also discards chips that have two bad nearest neighbors; these chips have a yield of 67.6% and hence a small incremental drop in profit is experienced.

Strategy		Lot 1	2	3	4	5	6	all P
Exhaustive:	f	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Y	89.7	87.5	82.9	45.6	87.5	92.9	79.7
Clairvoyant:	f	89.7	87.5	82.9	45.6	87.5	92.9	79.7
	Y	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Y_{E-V}	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Radial(48):	f	92.0	95.3	93.3	93.2	93.5	94.2	93.5
	Y	91.1	89.4	85.0	47.8	88.7	95.0	81.5
	$Y_{E-R(48)}$	74.1	50.0	54.2	15.8	70.1	58.9	53.5
Wafer Screen(0.48):	f	100.0	100.0	97.9	56.3	97.9	100.0	90.8
	Y	89.7	87.5	83.8	62.5	88.6	92.9	85.2
	$Y_{E-W(0.48)}$	—	—	42.4	23.9	34.7	—	25.3
Checkerboard(0):	f	99.7	99.2	97.0	77.1	98.6	99.7	94.6
	Y	89.9	88.1	85.3	58.8	88.7	93.1	84.1
	$Y_{E-C(0)}$	27.6	14.0	5.6	1.3	5.5	12.9	2.4
Checkerboard(1):	f	98.7	97.5	95.1	74.8	97.5	99.0	93.1
	Y	90.3	89.0	86.1	59.8	89.1	93.5	84.8
	$Y_{E-C(1)}$	49.2	28.2	19.9	3.6	25.9	33.8	10.4
	$Y_{C(0)-C(1)}$	55.0	35.1	42.5	26.2	50.1	41.6	39.0
Checkerboard(2):	f	94.2	93.2	90.8	71.9	93.9	96.1	89.4
	Y	90.9	90.3	86.9	60.0	89.8	94.3	85.6
	$Y_{E-C(2)}$	70.9	50.2	43.3	8.8	52.4	57.5	30.3
	$Y_{C(1)-C(2)}$	77.1	62.7	69.6	55.1	71.2	66.1	67.6
Metacheckerboard(0,0):	f	99.6	98.8	95.3	65.3	97.9	99.6	91.7
	Y	90.0	88.4	86.6	69.1	89.2	93.2	86.6
	$Y_{E-M(0,0)}$	28.2	18.0	7.5	1.6	7.5	14.3	3.1
	$Y_{C(0)-M(0,0)}$	29.5	25.8	11.0	2.2	11.0	16.8	4.4
Metacheckerboard(0,1):	f	98.4	96.9	93.4	63.1	96.6	98.5	90.2
	Y	90.5	89.4	87.5	70.5	89.8	93.9	87.4
	$Y_{E-M(0,1)}$	42.4	26.4	17.8	3.2	22.0	26.5	8.7
	$Y_{M(0,0)-M(0,1)}$	47.4	32.0	42.7	29.0	47.0	31.8	37.9
Sequential(10,0.05):	f	95.5	96.5	97.4	94.0	96.3	98.0	96.2
	Y	94.0	90.7	85.1	48.2	90.8	94.8	82.7
	$Y_{E-S(10,0.05)}$	0.5	1.6	0.9	5.1	1.3	0.6	2.3
Sequential(5,0.6):	f	94.0	95.3	96.0	74.8	94.5	97.2	91.4
	Y	95.3	91.6	86.0	53.1	92.1	95.4	85.7
	$Y_{E-S(10,0.2)}$	2.9	5.6	7.2	23.4	7.8	3.6	15.8
	$Y_{S(10,0.05)-S(5,0.6)}$	10.2	17.9	19.1	29.0	21.5	11.3	26.4
Mixed(10,0.05;0,0):	f	95.1	95.7	93.1	63.2	94.4	97.7	88.9
	Y	94.2	91.2	88.7	71.2	92.4	95.0	89.3
	$Y_{E-SM(10,0.05;0,0)}$	2.9	6.3	5.5	1.8	3.8	3.2	2.8
	$Y_{M(0,0)-SM(10,0.05;0,0)}$	0.5	1.7	1.2	5.8	1.5	0.6	1.8
	$Y_{S(10,0.05)-SM(10,0.05;0,0)}$	34.4	27.5	8.2	1.2	8.7	19.3	3.1

Table XIV. P fab results by lot.

Notice that we can use equations (9) and (11) to directly estimate the marginal yield $Y_{C(n)-C(n+1)}$ or the yield of discarded chips $Y_{E-C(n)}$. For example, a chip that is discarded under the $C(0)$ strategy is surrounded by four bad chips, and has probability $e^{\beta_0}/(1+e^{\beta_0})$ of being good under model (9). Referring back to Table VII, we can see that wafer $P.3.2$ has $\beta_0 = -2.0807$ and wafer $P4.28$ has $\beta_0 = -3.4702$. Hence, our yield estimate for the discarded chips is $\hat{Y}_{E-C(0)} = 0.111$ and 0.030, respectively, for these two wafers. The actual yield of discarded chips $Y_{E-C(0)}$ is 0.083 and 0.012, respectively.

Metacheckerboard strategy. On the D wafers, the metacheckerboard strategy $M(0,0)$ discards too many (21.9%) chips, which explains why it performs worse than the $C(0)$ policy. In contrast, the $M(0,0)$ strategy results in a 8.6% profit increase on the P wafers in Table XIV, and discards 8.3% of the chips, mostly from lot $P4$. The marginal yield $Y_{C(0)-M(0,0)}$ is only 3.1%, which is nearly as low as the 2.4% discarded chip yield under strategy $C(0)$. The $M(0,1)$ strategy achieves a 9.6% profit increase, although the yield of the marginally tested chips $Y_{M(0,0)-M(0,1)}$ is a rather high 37.9%.

Sequential strategy. The sequential strategy $S(12,0.05)$ works very well on the D wafers, resulting in a 9.5% profit increase. It is essentially exploiting radial effects here: the vast majority of the discarded chips are on the outer 3mm of the wafer. The sequential strategy also performs well on the P wafers. The S(10,0.05) strategy discards 3.8% of the chips in Table XIV, and these have a yield of only 2.3%. This strategy is also fairly consistent across lots, in that the fraction of chips discarded varies from 2.0% to 6.0% and the yield of the discarded chips varies from 0.5% to 5.1%. $S(5,0.6)$ is the best sequential strategy that was found, achieving a 7.3% profit increase.

Mixed strategy. The combined strategy $SM(10,0.05;0,0)$ achieves a 10.7% profit increase, which is close to the theoretical upper bound of 11.2%. Although the combined strategy discards 11.1% of the chips instead of the optimal 10%, few negative consequences are incurred since the marginal yield $Y_{M(0,0)-SM(10,0.05;0,0)}$ is only 1.8%.

Actual vs. simulated wafers. We conclude this subsection by considering the homogeneous (inner 18 mm.) portion of the D fab wafers under adaptive start rates, and testing a subset of the strategies on the actual wafers and on simulated wafers. More specifically,

for each of the 41 wafers, the MRF parameters for model (9) were estimated from the homogeneous portion, and 25 realizations of each MRF was randomly generated using the Gibbs sampler. The comparison between the actual wafers and the 1025 simulated wafers allows us to assess the validity of using simulated MRFs as an alternative to actual wafer maps (that have no strong radial effects) in future studies. The results are reported in Table XV, and the strong similarity between the actual and simulated wafers is readily apparent.

Strategy	Yield of Chips Tested	Yield of Chips Discarded	Fraction Tested	Start Rate	Increase in Profit
Actual Exhuastive Testing E	73.4	—	100.0	0.810	0.0%
Simulated Exhuastive Testing E	73.4	—	100.0	0.810	0.0%
Actual Checkerboard $C(0)$	74.0	19.2	98.9	0.819	0.8%
Simulated Checkerboard $C(0)$	74.5	16.3	98.6	0.821	1.0%
Actual Checkerboard $C(1)$	75.7	28.7	95.1	0.852	2.8%
Simulated Checkerboard $C(1)$	76.1	30.6	94.6	0.857	3.1%

Table XV. Actual vs. simulated results on the inner D wafers.

6.2. Predicting the Start Rate

The dramatic results in Subsection 6.1 are achieved under the assumption of adaptive start rates; that is, the expected fraction of tested chips, f, can be readily observed, and the optimal wafer start rate λ^* can be chosen via (5). Now we consider non-adaptive control, where the wafer start rate must be determined in advance by estimating f. A reasonable estimate of f for every strategy S under consideration is vital for preventing overutilization or underutilization of the facility.

The expected fraction of chips, f_S, that a given strategy S will test depends upon the exact nature of the wafers encountered by the strategy. In many situations, however, detailed yield data, such as the wafer maps analyzed here, may be difficult to obtain. Even if such maps are available, backtesting the strategies on the actual wafers to determine f_S would be a fairly time-consuming task. Hence, it would be useful to be able to predict f_S from more accessible data. In particular, we focus here on predicting f_S when only the average yield $(p_1, p_2, \ldots, p_K)$ of each wafer from a representative past sample is known.

Since the radial strategy is a static, deterministic strategy, determining the fraction of chips tested is a simple matter of counting the number of locations within the cutoff radius. That is, for a radial strategy with cutoff radius r,

$$\hat{f}_{R(r)} = \frac{\sum_{i,j} 1_{\{r_{i,j}<r\}}}{N},$$

where N is the number of chips on a wafer and $r_{i,j}$ is the distance from location (i,j) to the center of the wafer. Likewise, estimating the fraction tested under a wafer screening strategy is straightforward if we have access to yield data by wafer:

$$\hat{f}_{W(y)} = 1 - \frac{1-\phi}{K} \sum_{k=1}^{K} 1_{\{p_k \leq y\}},$$

where y is the cutoff yield level and ϕ is the fraction of chips tested in the first pass of the wafer screening strategy (1/16 in our study).

Using the data $(p_1, p_2, \ldots, p_K)$ to predict the fraction of chips tested for the sequential strategy is problematic, since, as we saw in Section 4.1, a Bernoulli model vastly underpredicts the number of "holes" that occur on actual wafers. We believe that a reliable estimate of f_S for the sequential strategy can only be found by analyzing wafer maps for holes and/or by developing a chip yield model that explicitly captures inter-wafer effects and radial effects.

Predicting f under the checkerboard strategy is also difficult to do without an explicit spatial model, such as the MRF. However, two extreme cases can be easily evaluated using only the mean wafer yield. The first case assumes that the yield of each chip on a wafer is an iid Bernoulli random variable with parameter p. In this case, the expected fraction of chips tested by the checkerboard strategy $C(n)$ is

$$\hat{f}_{C(n)} = 1 - \frac{1}{2} \sum_{i=0}^{n} \binom{4}{i} p^i (1-p)^{4-i} \quad \text{for} \quad n = 0, 1, 2, 3. \tag{13}$$

The second limiting case assumes perfect dependence, where every chip on the wafer is good

with probability p and every chip on the wafer is bad with probability $1-p$. In this case,

$$\hat{f}_{C(n)} = \frac{1}{2} + \frac{p}{2} \quad \text{for} \quad n = 0, 1, 2, 3, \tag{14}$$

which is independent of n. Although neither of these cases is very realistic, they should offer upper and lower bounds on f for most wafer maps, since clustering of defective chips is very common on wafers.

Similarly, predicting f under the metacheckerboard strategy $M(n_1, n_2)$ can be done under the Bernoulli model, although the resulting equations are more complicated. For example,

$$\hat{f}_{M(0,0)} = 1 - \frac{1}{4}(1-p)^4 - \frac{1}{2}(1-p)^2[(1-p)^2 + (1-(1-p)^2)(1-p)]^2. \tag{15}$$

Under perfect dependence,

$$\hat{f}_{M(n_1,n_2)} = \frac{1}{4} + \frac{3p}{4} \quad \text{for} \quad n_1, n_2 = 0, 1, 2, 3. \tag{16}$$

Since the checkerboard policy exploits spatial clustering, it seems natural to employ the MRF model to estimate f for these strategies. Although an exact analytical calculation of f is not computationally feasible, we can apply the checkerboard/metacheckerboard strategies to simulated wafers (i.e., realizations of MRFs generated by the Gibbs sampler) to determine f. The approach requires historical wafer maps to estimate the MRF parameters, and hence, at first glance, it appears to be very roundabout; backtesting the wafers on historical wafer maps would produce a more direct and reliable estimate of f. However, this approach can predict f as a function of the yield and the MRF parameter β_1, and, as explained below, fab managers can get a crude estimate of β_1 by visually inspecting their wafer maps.

Simulation results for the checkerboard strategy $C(0)$ are shown in Table XVI. This table gives the estimated fraction tested for the $C(0)$ strategy as a function of wafer yield and the spatial dependence parameter β_1. The constant-β_1 MRF described at the end of Section 4 was used to generate the table. Even though the more complicated MRF models provide a better fit for the actual wafers, the constant-β_1 MRF is the most reasonable choice, given the

	β_1					
Yield	0.0	0.5	1.0	1.5	1.75	"∞"
0	50	50	50	50	50	50
10	67	67	66	65	64	55
20	80	79	76	74	71	60
30	88	87	85	81	75	65
40	94	93	89	86	80	70
50	97	96	94	90	84	75
60	99	98	97	94	89	80
70	100	99	98	96	93	85
80	100	100	100	99	96	90
90	100	100	100	100	100	95
100	100	100	100	100	100	100

Table XVI. Estimated percentage of chips tested for checkerboard(0).

minimal amount of data accessible to the typical fab manager. These results were generated by using a Gibbs Sampler on a 16×60 torus, with 1000 iterations between samples. The column $\beta_1 = 0$ corresponds to the fraction tested under the Bernoulli assumption, as in equation (13), and the column $\beta_1 =$ "∞" corresponds to the fraction testing under perfect dependence, as in equation (14). Note that although the Bernoulli case corresponds to an actual Markov Random Field (with $\beta_1 = 0$), the $\beta_1 =$ "∞" case *does not* correspond to an actual Markov Random Field. Table XVI shows that the influence of β_1 on the fraction tested is small for wafers with very high or very low yields, and is large for wafers with yields close to 50%. Similar tables for the $C(1)$ and $M(0,0)$ strategies can be found in Tables 5.17 and 5.18 in Longtin.

Although generating these tables required extensive computer simulation, once generated, the tables can be used in any fab. To predict the fraction of discarded chips using one of the tables, a fab manager would need the yield of each of the wafers under consideration, as well as an estimate of β_1. Although β_1 is difficult to calculate, the human eye can often detect spatial dependence. If the fab manager has access to wafer map pictures like those in the Appendix, then a visual comparison of the actual wafer maps with the simulated wafer maps will provide a crude estimate of β_1. Towards this end, Appendix E of Longtin contains ten simulated wafers with different levels of spatial dependence. Since spatial dependence is

Strategy	Lot 1	2	3	4	5	6	all P
$C(0)$ Bernoulli	100.0	100.0	99.8	89.5	99.7	100.0	98.0
$C(0)$ MRF	100.0	99.8	99.2	82.5	99.6	100.0	96.1
$C(0)$ Actual	99.7	99.2	97.0	77.1	98.6	99.7	94.6
$C(0)$ Dependent	94.9	93.8	91.4	72.8	93.8	96.4	89.9
$C(1)$ Bernoulli	99.7	99.7	98.6	80.3	98.9	100.0	95.7
$C(1)$ MRF	99.4	98.1	96.3	75.4	97.9	99.2	93.7
$C(1)$ Actual	98.7	97.5	95.1	74.8	97.5	99.0	93.1
$C(1)$ Dependent	94.9	93.8	91.4	72.8	93.8	96.4	89.9
$M(0,0)$ Bernoulli	99.9	100.0	99.5	82.0	99.5	100.0	96.4
$M(0,0)$ MRF	99.9	99.6	98.5	72.7	99.2	100.0	94.3
$M(0,0)$ Actual	99.6	98.8	95.3	65.3	97.9	99.6	91.7
$M(0,0)$ Dependent	92.3	90.6	87.2	59.2	90.6	94.7	84.8

Table XVII. Fraction of chips tested in the P fab: actual vs. predicted.

most obvious to the eye and has the biggest impact on predicted f for wafers with yields near 50%, these simulated wafers all have expected yield of 50%. Several representative wafers from the fab with yields near 50% can then be compared to these simulated wafers to choose β_1. Alternatively, if no such wafer maps are available, a crude estimate of $\beta_1 \approx 1.0$ may be reasonable, based upon the results of Table 4.11 in Longtin. Using Tables 5.17 and 5.18 in Longtin, Table XVI and the β_1 estimates from Section 4.2, we can predict f for the actual wafers via simple linear interpolation. For the P fab wafers, Table XVII gives the actual fraction tested under three checkerboard-type strategies, and gives the corresponding quantities estimated by the Bernoulli model, the perfect dependence assumption, and the MRF. As expected, the Bernoulli model overestimates the fraction tested, and the perfect dependence assumption underestimates the fraction tested. Although the MRF model exhibits a systematic upward bias, it predicts f to within 1–3% on the P lots as a whole.

As mentioned earlier, if the true value of f is known, then the optimal wafer start rate λ^* is determined by (5). Using this result, we also showed that for any feasible strategy A, the optimal fraction of chips tested would usually be given by

$$f_A = \frac{\mu_T}{\mu_F M(1-q)}, \tag{17}$$

which in our case corresponds to testing 90% of the chips. This value of f ensures that both the fab and the testing facility operate at their effective capacity. However, if the exact value of f_A is unknown and we overestimate the true value of f_A (that is, our estimate is $\hat{f}_A > f_A$) and set $\lambda = \lambda^*$, then a feasible, but suboptimal, solution is obtained; notice that this would occur if the MRF estimates in Table XVII are used. However, if we underestimate the true value of f_A (that is, $\hat{f}_A < f_A$) and set $\lambda = \lambda^*$, then the testing capacity constraint will be violated, leading to excessive work-in-process inventory and long lead times. Two remedies are available to guard against this latter possibility: reduce the wafer start rate below λ^* or reduce the fraction of chips tested below our estimate, $\hat{f}_A$. We conclude this subsection by attempting to determine which of these two approaches results in higher profits.

Let us suppose, for the sake of concreteness, that the utilization of the testing facility must be reduced by a factor ω to limit the probability of violating the testing constraint to an acceptable level. Denote the reduced start rate approach by the start rate $\omega\lambda^*$ and the testing strategy A, where f_A is given in (17), and denote the reduced testing approach by the start rate λ^* and the testing strategy B, where $f_B = \omega f_A$. By (4), (5) and (17), it can be shown that the reduced testing policy is more profitable than the reduced start rate policy whenever

$$Y_A - Y_{A-B} > \frac{c_F \mu_F}{r \mu_T} = 0.117. \tag{18}$$

Therefore, the reduced testing policy is preferable (i.e., discard more chips) if the marginal yield of those discarded chips is less than the average yield of the tested chips by 11.7%. This inequality is easily satisfied in Table XIV by the four examples $A = C(0), B = C(1); A = C(1), B = C(2); A = M(0,0), B = M(0,1)$; and $A = S(10, 0.05), B = S(5, 0.6)$. In these cases, it is preferable to select a strategy that discards slightly more than 10% of the chips rather than to reduce the start rate; testing slightly less than 90% makes excess congestion unlikely, while giving up little in profits.

6.3. Summary

In this section, we tested our screening strategies on the actual wafers under the assumption that f_S, the fraction of chips tested under a given strategy S, could be observed by the

decision maker, who then chose the optimal start rate using (5). Since it is important to have a reasonably accurate estimate of f_S in advance, this quantity was estimated for a variety of testing strategies, under limited information. Also, if a fab manager does not want to risk overutilizing the testing facility, and his or her estimate of f_S is uncertain, then conditions are derived under which reducing the fraction tested below its optimal value is preferable to decreasing the wafer start rate below its optimal value.

For both fabs, we found sequential screening strategies that were extremely effective at discriminating between good and bad chips. In particular, for each fab's actual wafers, we calculated the maximum possible increase in profit that a clairvoyant sequential strategy could achieve over the exhaustive strategy, given the capacity-constrained hypothetical fab parameters in Table I. A sequential screening strategy for each fab was found that achieved 95% of the maximum possible profit increase.

Moreover, since the wafer maps exhibited a wide variety of yield nonuniformities, there was more than one way to achieve a large profit increase. Not surprisingly, the radial policy was the best among the proposed strategies on the D fab wafers, and was the worst on the P fab wafers. The wafer screening policy, which exploits yield variation by wafer, performed better on the P fab wafers than the D fab wafers, because of the large number of low yielding wafers in lot $P4$. The sequential policy exploits holes and radial effects, and performed well in both fabs, particularly in the D fab, which was dominated by radial effects. The best of the checkerboard-type policies performed very well in both fabs. The checkerboard policy was successful in the D fab by exploiting the radial effects and the moderate clustering of bad chips, and the metacheckerboard policy achieved a high profit increase in the P fab by capitalizing on the strong spatial clustering of bad chips. Furthermore, on the P wafers, the sequential and metacheckerboard strategies complemented each other and their combination led to the best performance: the sequential strategy effectively discarded 3-5% of the worst chips from every wafer, and the metacheckerboard strategy discarded up to 45% of the worst chips on bad wafers, while leaving good wafers relatively untouched.

The desirability of a screening policy depends not only on its profitability performance, but on other factors as well. A screening strategy is more likely to be implemented if (i)

the parameters for the strategy can be easily chosen, (ii) the fraction of chips tested can be accurately estimated, (iii) it can exploit different types of yield nonuniformities, which should lead to more robustness, and (iv) it is easy to explain to fab managers. The concepts behind all of our proposed policies are easy to explain, and hence condition (iv) is satisfied.

The radial policy satisfies all but condition (iii). Although its lack of robustness prevents it from being broadly applicable, it should be the policy of choice for a fab that is dominated by the radial effects apparent in Figure 2. In fact, in this case, it would be more economical to go one step further and *cease fabrication* on the outer edges of the wafer. However, other factors related to organizational learning may still make it worthwhile to fabricate chips on the outer edges of the wafers.

The wafer screening strategy can only exploit wafer-to-wafer yield variability, and its limited robustness prevents it from performing well in the D fab. The sequential policy is robust, but does not satisfy conditions (i) and (ii). In particular, among the basic policies tested, its parameters are hardest to choose and an accurate estimate of the fraction of chips tested is most difficult to obtain.

The checkerboard-type policies, and the $C(n)$ policy in particular, appears to satisfy all four conditions. More specifically, the parameters are easy to choose (a production fab would rarely choose a value of n other than zero or one) and equation (9) or (11) can be used to obtain a rough estimate of the fraction of chips tested and the marginal yields (such as $Y_{E-C(0)}$ or $Y_{C(0)-C(1)}$), both of which are helpful for determining the start rate and the proper parameter values.

The mixed strategy that performed so well in the P fab is the most robust, since it can exploit radial effects, yield holes, and spatial clustering. Unfortunately, it suffers from the same drawbacks as the pure sequential strategy. However, if the effort is taken to fine tune the parameters and estimate f, then this policy should be very effective in a wide variety of facilities.

7. Concluding Remarks

This paper and its companion, Ou and Wein, have examined a particular quality man-

agement issue in semiconductor manufacturing. Many semiconductor facilities are capacity constrained, and we have focused on those facilities where electrical probing, or testing, is the bottleneck in the process. By finding and exploiting nonuniformities in chip yield, we showed how to increase the throughput, and hence the profitability, of the facility. Since a 1% increase in revenue minus variable costs corresponds to about 10 million dollars in a typical production fab, the procedures developed here can potentially result in huge savings. Moreover, the analysis can also be used to assess the value of additional testing capacity.

Three features of our study are worth emphasizing. A key goal of the study was to analyze industrial data to identify the types and magnitudes of various yield nonuniformities. Our results suggest that screening at the wafer level (that is, discarding the remaining wafers in a lot) is not nearly as effective as screening at the chip level. This is due to the fact that the lot-to-lot variability, although significant in magnitude, is dominated by the within lot variability. We are not in a position to make any sweeping generalizations about the relative magnitude of lot-to-lot versus within lot variability or about the relative magnitudes of the various types of within lot variability, such as radial effects, spatial clustering and temporal effects (i.e., yield holes). However, we have developed a procedure, consisting of exploratory data analysis, model building and analysis, and proposed screening policies, to determine the best sequential screening approach for any given facility.

The second aspect of our study concerns the modeling of yield. We proposed two models, the Bayesian gamma-gamma model and the Markov random field, that appear to be new to the quality control literature. Although alternative models may exist that fit the yield data just as well, the model-fitting results in both papers make it very clear that traditional quality control models, which are based on simplifying independence assumptions and employ the Poisson, Bernoulli or binomial random variables, are woefully inadequate at representing the data.

The final feature of the study is the systems approach that is taken. We believe that quality control problems in manufacturing cannot be looked at in isolation. The economics of the process and the quantity control aspects, such as process flows and bottlenecks, also need to be considered. In particular, the premise upon which this analysis is based, *profitability*

can be increased by using sequential screening to prevent bottlenecks from working on bad parts, can be employed in virtually every manufacturing facility, not just in semiconductor manufacturing.

Acknowledgment

We thank Michael Harrison and Michael Pich for helpful comments during a presentation of this work. This research is supported by a grant from the Leaders for Manufacturing Program at MIT and National Science Foundation grant DDM-9057297.

References

Albin, S. and D. J. Friedman. 1989. The Impact of Clustered Defect Distribution in IC Fabrication. *Management Science* **35**, 1066-1078.

Besag, J. 1972. Nearest-neighbor Systems and the Auto-logistic Model for Binary Data. *Journal of the Royal Statistical Society* **34**, 75-83.

Besag, J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society* **36**, 192-236.

Besag, J. 1975. Statistical Analysis of Non-lattice Data. *The Statistician* **24**, 179-195.

Besag, J. 1986. On the Statistical Analysis of Dirty Pictures (with discussion). *Journal of the Royal Statistical Society Series B* **48**, 259-302.

Besag, J. and P. Clifford. 1989. Generalized Monte Carlo Significance Tests. *Biometrika* **76**, 633-642.

Bohn, R. E. 1991. Noise and Learning in Semiconductor Manufacturing. Center for Technology Policy and Industrial Development, MIT, Cambridge, MA.

Comets, F. and B. Gidas. 1991. Parameter Estimation for Gibbs Distributions from Partially Observed Data. *Annals of Applied Probability* **2**, 142-170.

Derin, H. and H. Elliott. 1987. Modeling and Segmentation of Noisy and textured Images using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**, 39-55.

Flack, V. F. 1985. Introducing Dependency into IC Yield Models. *Solid-State Electronics* **28**, 555-559.

Gelman, A. and D. B. Rubin. 1991. A Single Series from the Gibbs Sampler Provides a False Sense of Security. Technical Report 305, Department of Statistics, University of California, Berkeley.

Geman, S. and D. Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine In-*

telligence **6**, 721-741.

Geman, S. and C. Graffigne. 1980. Markov Random Field Image Models and their Applications to Computer Vision. In *Proceedings International Congress Mathematics*, A. M. Gleason, ed., American Mathematical Society, Providence, R.I.

Gidas, B. 1991. Parameter Estimation for Gibbs Distributions, I: Fully Observed Data. In *Markov Random Fields: Theory and Applications*, R. Chellapa and R. Jain, eds., Academic, New York.

Ising, E. 1925. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik* **31**, 253-258.

Longtin, M. D. 1992. Sequential Screening in Semiconductor Manufacturing: Exploiting Spatial Dependence. M.S. thesis, Sloan School of Management, MIT, Cambridge, MA.

McCullagh, P. and J. A. Nelder. 1983. *Generalized Linear Models.* Chapman and Hall, Ltd., London.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087-1092.

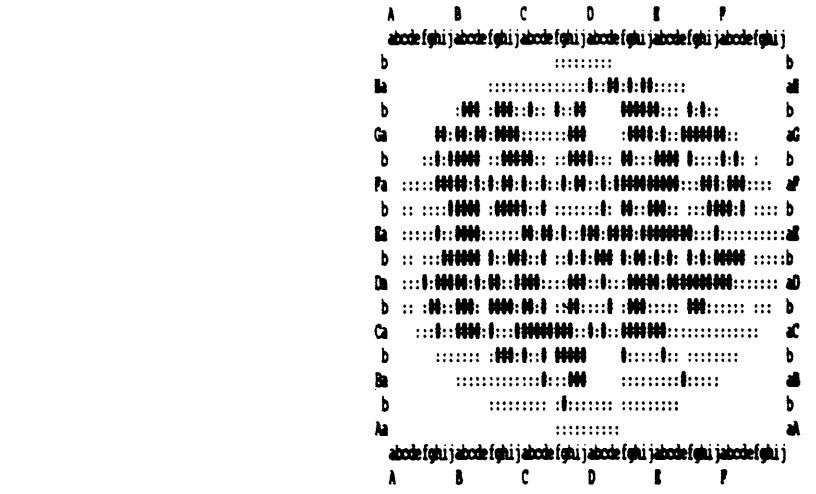
Ou, J. and L. M. Wein. 1992. Sequential Screening in Semiconductor Manufacturing, I: Exploiting Lot-to-Lot Variability. Sloan School of Management, MIT, Cambridge, MA.

Pickard, D. K. 1987. Inference for Discrete Markov Fields: The Simplest Nontrivial Case. *Journal of the American Statistical Association* **82**, 90-96.

Strauss, D. 1991. The Many Faces of Logistic Regression. Department of Statistics, University of California, Riverside.

Wald, A. 1947. *Sequential Analysis.* Wiley, New York.

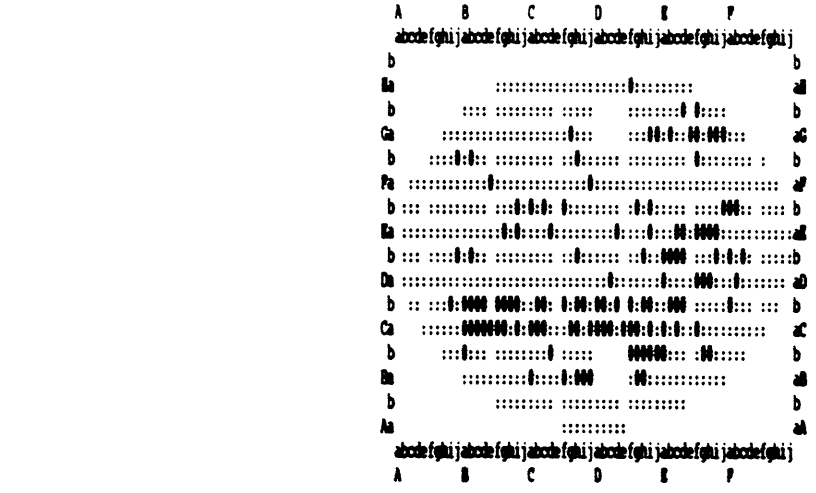
Appendix. Sample wafer maps



"#" = good; 275 chips total

":" = bad; 369 chips total

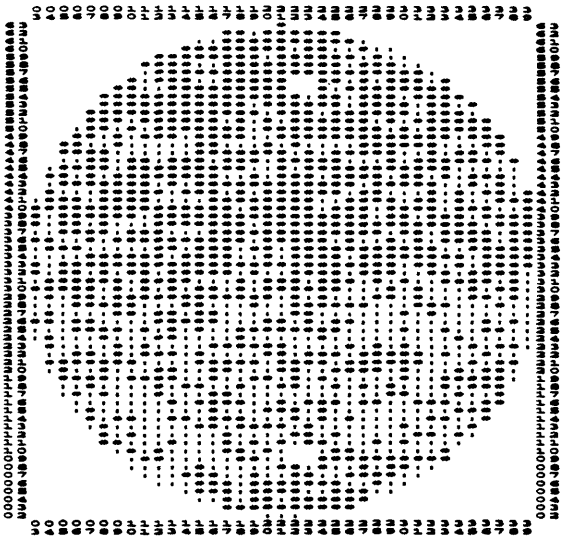
Wafer $D.26$, the 26^{th} wafer from the D fab.



"#" = good; 120 chips total

":" = bad; 519 chips total

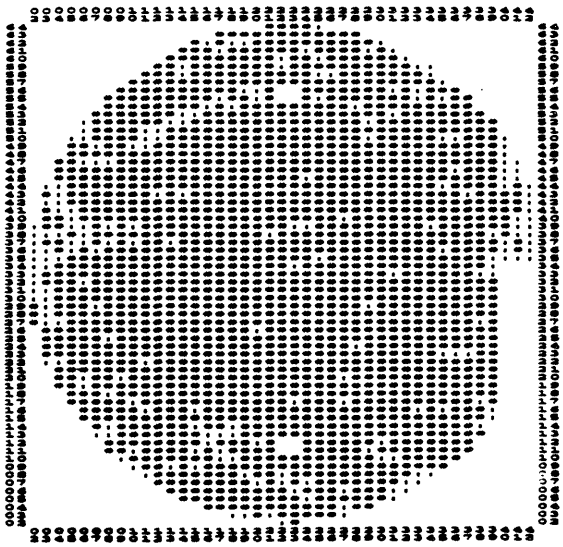
Wafer $D.34$



"#" = good; 1635 chips total

":" = bad; 163 chips total

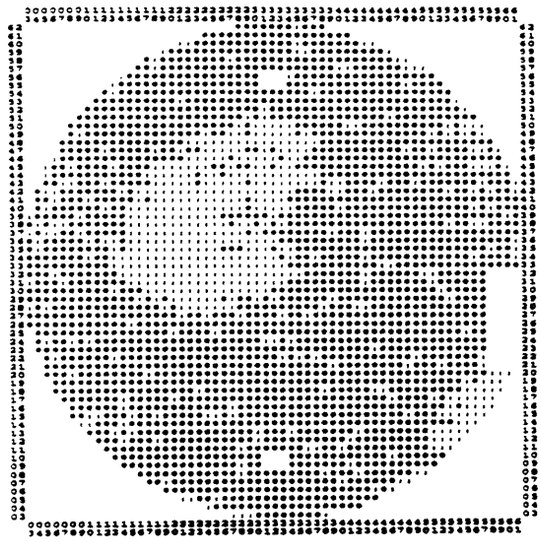
Wafer $P1.47$, the 47^{th} wafer from the 1^{st} lot from the P fab.



"#" = good; 1755 chips total

":" = bad; 219 chips total

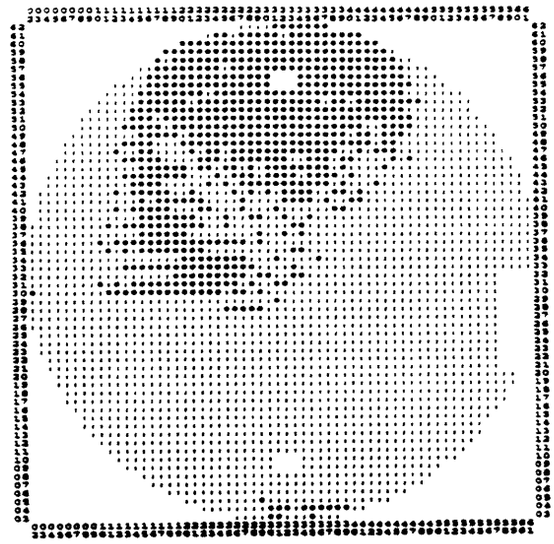
Wafer $P2.44$



"#" = good; 2170 chips total

":" = bad; 617 chips total

Wafer *P*3.2



"#" = good; 716 chips total

":" = bad; 2101 chips total

Wafer *P*4.28