Process Improvement Through the Analysis of Spatially Clustered Defects on Wafer Maps

Mark H. Hansen
Bell Laboratories
Murray Hill, New Jersey, 07974
cocteau@bell-labs.com

Vijayan N. Nair
University of Michigan
Ann Arbor, Michigan, 48109
vnn@umich.edu

David J. Friedman
Integral, Inc.
Cambridge, Massachusetts, 02138
dfriedman@ag-inc.com

Summary. This paper deals with statistical methods for process and yield improvement in integrated circuit (IC) fabrication. Our approach is based on an overall strategy that attempts to exploit important spatial information in wafer map data. The data in question are both high-dimensional and highly structured, requiring visualization tools and flexible methods of analysis. Our methodology goes beyond traditional statistical process control methods which primarily emphasize process monitoring for change-point detection. A considerable focus of our work is on using the spatial defect patterns to develop failure diagnostics and relating the spatial signatures to potential manufacturing problems in order to improve the manufacturing process. As demonstrated in this application, there is a lot of useful information that can be mined from the extensive and often complex manufacturing data that are being routinely collected. This is especially so in advanced manufacturing applications where there is a critical need to use such information for continuous process improvement.

Key Words: Clustering; Markov random field model; Semiconductor manufacturing; Spatial patterns; Spatial process monitoring; Statistical process control.

1. Introduction

Statistical methods have been used extensively in manufacturing industries for process monitoring and improvement, dating back to the pioneering work of Walter Shewhart in the 1920's. In recent years, however, the environment in high-technology industries has changed considerably, dictating a need to re-evaluate the traditional role and application of these methods.

On the one hand, pressures from the competitive marketplace are forcing manufacturers to constantly reduce product development cycle times. Product innovations can now move from design to full-scale production well before the technology and the fabrication processes are completely understood. In semiconductor manufacturing, for example, newly designed integrated circuits (IC's or chips) have features as small as $0.16\mu m$, and there are plans to reduce these sizes to $0.01\mu m$ in the next $5-10$ years. The IC fabrication process is very complex involving hundreds of steps and lasting up to several weeks. The separate steps are rarely "stable," and frequently interact in unexpected ways. Such complexity and instability are typical in advanced manufacturing applications. As a result, we need to go beyond the traditional emphasis of statistical methods on process monitoring. There is a critical need in these applications to exploit all available information from in-process and product quality data to do "real-time" process improvement. The information to be gleaned from these sources can greatly enhance process improvement efforts, especially during the introduction and "ramp up" phases for new products.

On the other hand, advanced manufacturing facilities now routinely collect extensive amounts of in-line and production data. These data are often high-dimensional with complex structure. In IC fabrication, the data are spatial measurements on a wafer map. Figure 1 shows some examples of inline and post production data. Due to the lack of appropriate statistical methods and software tools for analyzing such data, engineers typically summarize the data into simple measures such as overall yield and use these measures to track the process. These summary measures, however, ignore the structure in the data containing important information, often resulting in lost opportunities for process improvement.

Over the past several years, we have been involved in an effort to develop statistical methods and associated software tools for process and yield improvement in IC fabrication. Our approach is based on an overall strategy that attempts to exploit the spatial information in wafer map data. One of the components in our strategy is the development and use of methods for spatial monitoring of the wafer map data for defect clustering. Our methodology, however, goes beyond traditional process monitoring and change-point detection. A considerable focus of our work is the development of methods for identifying the spatial patterns of defects and using them to obtain spatial signatures at the lot or group level. These signatures can then be related to processing information to identify and correct process problems.

In Section 2, we provide a brief background of IC fabrication processes and describe the nature of the data as well as our strategy for exploiting the spatial information for process and yield improvement. The strategy

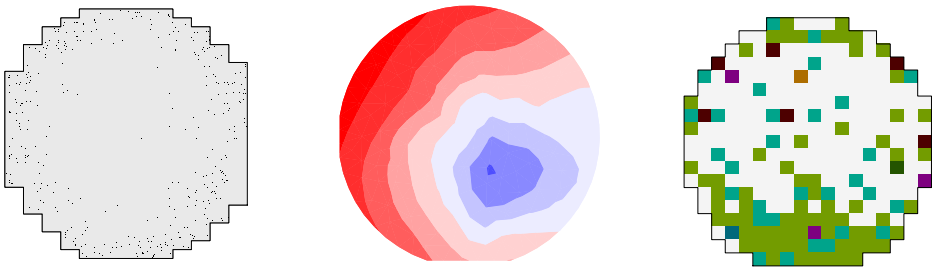


Figure 1: *Examples of inline and post-production data. The leftmost map was obtained by an optical inspection device, each black square representing a particle (we have magnified the defects by a factor of 1000 to make them visible). The color contour plot in the middle panel was obtained by smoothing inline thickness measurements taken from a single wafer. Finally, the wafer map at the right displays final probe test results. Each square represents a chip, and the colors denote different failure modes (white indicating the chip passed all of the tests and will be cleaved from the wafer and packaged).*

consists of several components which are taken up in detail in subsequent sections. Section 3 summarizes both formal and graphical methods for spatial monitoring of wafer map data. Section 4 deals with methods for identifying the spatial pattern of clustered defects at the wafer map level. In Section 5, we apply these methods to obtain signatures of the major patterns at a lot or group level. Section 6 describes methods for relating these signatures to process information in order to identify potential process problems.

A thread that will run through the paper is the role of technology transfer, making statistical methods available to process, product and yield enhancement engineers. To ensure that the complexity in the data does not give rise to equally complex (in terms of use or interpretation) statistical methods, emphasis is placed on bundling computations through a convenient software environment like S or S-PLUS. Graphical representations will also be extremely important for incorporating natural structures into traditional SPC displays. The paper concludes with a discussion of software tools and technology transfer.

2. Semiconductor Manufacturing

The fabrication of IC's is an extremely complex process, involving hundreds of separate steps and lasting many weeks. Several hundred IC's are fabricated simultaneously on a 6–8 inch disc of silicon called a *wafer*, and the wafers are themselves processed in groups called *lots*. A large manufacturing plant can start thousands of wafers a week. The process leading from bare silicon to computer chip is not unlike the development of black-and-white photographs. Transistors are built up in layers, with material being alternately deposited, exposed to light through a mask (the equivalent of a "negative" for a photograph) and

finally subjected to an etching process by which the unexposed material is removed. The features created in this way are currently as small as $0.16\mu m$, and as a result fabrication is done in a near sterile environment called a *clean room*. Samples of wafers from each lot are examined at various steps during the process to assess the impact of particulate defects, the thickness of different layers, the performance of test structures created in the areas between the chips, and so on (see Figure 1). In addition, at the end of the line, each chip on every wafer is subjected to a number of functionality tests. (To reduce probe time, testing of each chip is stopped after the first failed functionality test.) The results of these final probe tests can be displayed spatially in the form of a *wafer map* (see Figure 1).

It is well known that most of the manufacturing processes leading up to this testing phase are spatially inhomogeneous, and hence understanding the *patterning* in probe data can be very useful in diagnosing process problems. The rightmost wafer map in Figure 1 is a case in point. Defects clearly exhibit spatial clustering, an effect which is often attributable to specific process steps or equipment problems. Unfortunately, production volume precludes engineers from manually examining all the available inline and post-production data as part of their routine process improvement efforts. The quantity and complexity of data has forced the quality improvement efforts to focus primarily on one-dimensional summaries like the number of salable chips per lot. While this has the advantage of suggesting simple, non-ambiguous screening rules (such as tag a lot if less than than 90% of the chips are usable), it ignores important information in the data on the nature and possible causes of the defects, crucial information for process improvement.

In this paper, we discuss the use of flexible statistical techniques and visualization tools aimed at exploiting the information in the extensive amounts of manufacturing data. We will concentrate mainly on final probe data.

In Figure 2, colored squares denote chips that failed one or more final probe tests, while the white regions represent good chips that will eventually be cleaved from the wafer and sold. The colors represent different failure modes, labeled $f1$ through $f5$ in the figure. The three maps each have virtually the same yield, and even share similar distributions of failure types as seen from the accompanying barplots. Each, however, represents much different opportunities for process enhancement and yield improvement. In the first case, defects are (essentially) spatially random, and hence are likely the result of particulate defects. While manufacturing takes place in a near sterile environment, particles cannot be removed completely, and their levels tend to rise and fall with the overall cleanliness of the clean room. This sort of "background" effect can only be reduced through long term, gradual improvement or extensive equipment overhaul. Particulate defects in excess of this background level can, however, indicate that one or more machines in the line require cleaning or other routine maintenance.

In the second probe map of Figure 2, we observe a different defect signature. Particles are still an issue, but we also notice a large, clustered defect. Given the spatial character of IC fabrication, it should not

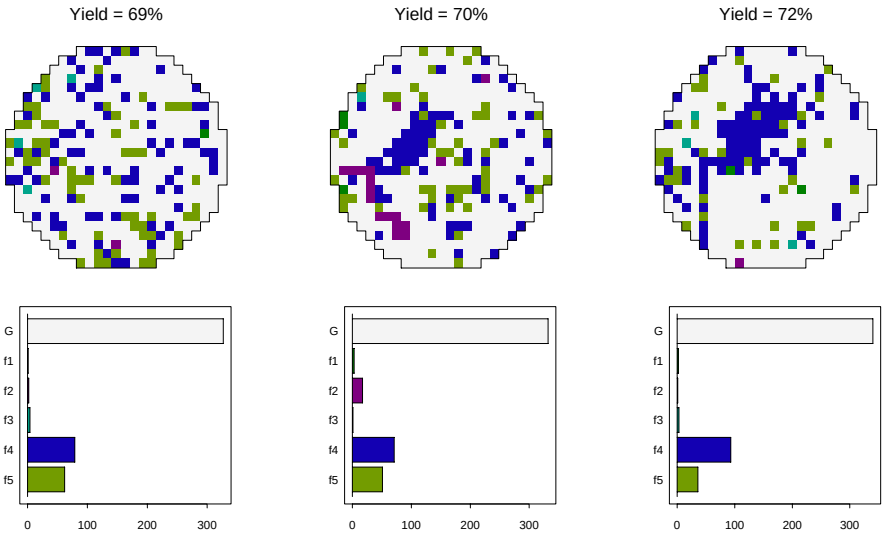


Figure 2: *The three wafers in this figure are of the same code and hence have the same number of chips per wafer (475). The colored squares denote defective chips. While the yields are essentially the same, each wafer represents a different challenge for yield improvement. Note also that the distribution of failure modes (f1, f2, f3, f4 and f5) is similar for each map, although the underlying failure mechanism is clearly much different.*

be surprising that process problems give rise to telltale *signatures* evident at final probe. Failure-mode-analysis (FMA) engineers use this fact to identify one or more processes (or even individual machines) that are responsible for the observed patterns in single lots. Clusters like that in the second map of Figure 2 represent a tremendous opportunity for yield enhancement. The final map in this figure, where many chips are lost to the large clustered defect, represents an even greater opportunity for yield improvement. Thus, the development and routine use of methods for analyzing probe data and for relating the spatial patterns to process problems can be very effective in enhancing current, mostly manual, FMA efforts.

In terms of classical SPC terminology, the particulate defects in the leftmost wafer of Figure 2 can be viewed as being due to *common* causes, while the clusters in the remaining two maps can be interpreted as being due to *special* or *assignable* causes. We do not mean to imply that there are never any assignable causes attributable to spatially random defects. As noted above, machines nearing their scheduled cleaning can produce heavier than normal random defects. Existing methods (which ignore spatial structure) can usually identify this situation early, helping to improve yields. Our focus here is on spatial patterning, which is extremely common in IC fabrication data and is also directly attributable to specific process problems.

For example, if the defects concentrate in the center of the wafer, there is likely a problem controlling the thickness of a chemical "resist" deposited on the surface of the wafer prior to lithography. If defective chips

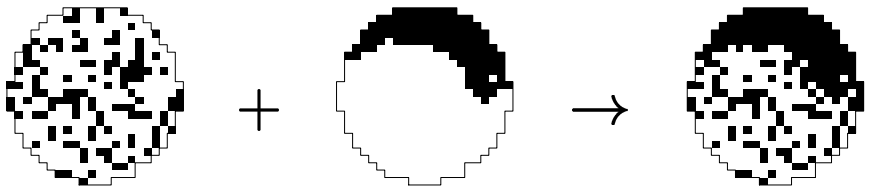


Figure 3: *A graphical model for overall wafer yield. The chips in the rightmost wafer occur essentially at random, while those in the middle are "clustered," reflecting a process problem. The final probe map is a superposition of these two processes.*

occur in a (distended) crescent shape along the bottom of the wafer, temperature control within a furnace is the likely cause. And observing defects arranged in an annulus typically signals problems with the chemical mechanical polishing process. This is not to imply that a fixed library of patterns is sufficient to categorize the results of final probe. On the contrary, as feature sizes drop, both processes and machines are pushed to their limits, so that new patterns emerge regularly. When there are gross processing problems that affect an entire lot, these patterns can be self-evident. With lesser problems, patterns are often difficult to discern, and the difficulties get magnified if an engineer is faced with thousands of such wafer maps a week.

The process improvement strategy in this paper is based on a simple and somewhat idealized model for *binary* wafer map data. The binary data are obtained by relabeling all failure modes as "failures" and creating a black-and-white map corresponding to defective and non-defective chips respectively (see Figure 3). In our idealized model, we view the pattern of defective chips on a given wafer map as being caused by a superposition of two destructive mechanisms. Acting independently, the first generates defects essentially at random, while the second produces "clusters." The dead chips on the leftmost wafer of this figure are primarily the result of particulate defects discussed previously. To model the yield lost on the middle wafer, we consider again the fact that the processing equipment does not affect all regions of the wafer equally. Here, the temperature variation in a furnace exposed chips near the upper edge to processing conditions that were beyond specification limits. The temperature gradient is in fact a smooth function across the wafer that gradually drifts out of control near top. The probability that a chip will fail because of this process problem tends to one in the region of dead chips. In general, one can envision variation in processing conditions across the wafer having a smooth, spatial effect on the probability that the chip will be defective. Thus, the defect probability will be close to one in regions where the processing conditions are far from specification limits.

Starting with this view of wafer yield, we have developed methods that treat mapped wafer data as an intrinsically spatial object and exploit the spatial information in the patterns of defects for process and yield

enhancement. Our overall strategy consists of the following steps.

1. Routinely monitor final probe data to *detect* the presence of clustered defects;
2. *Identify* spatial patterns when they exist, essentially reversing the arrow in Figure 3;
3. Create *spatial signatures* associated with these patterns, perhaps combining data across lots processed in a given time interval;
4. *Improve* the process by relating these signatures to processing information.

As we will see, each of these steps is important in its own right in addition to being part of an overall yield improvement strategy. We also note that, unlike traditional SPC efforts in which root cause analysis is treated as an engineering problem, much of the focus of our work is on developing procedures for relating the observed patterns in final probe maps to shop flow and various inline data. This leads to useful diagnostics that can help to identify the nature of the problem and thereby help improve the process.

The development of appropriate statistical methods for the various steps above raises interesting research questions. As technology transfer is essential in this context, the characteristics of the factory setting imposes hard constraints on our choice of procedures. In this paper, we discuss a particular set of methods for handling patterned probe map data, but certainly many others are possible and are the subject of continuing research. In the next four sections, we take up each step in our procedure above. At the end of each section, we describe how this work has evolved and some of our ongoing research to address open issues.

3. Process Monitoring

In most factories, routine monitoring procedures are typically based on overall lot- or wafer-level yield and related metrics. The spatial structure or pattern of defective chips is not considered explicitly until a lot has been flagged for FMA. For example, lots achieving less than 80% of planned yield might be tagged for FMA, at which point the yield-improvement engineers will manually examine the wafers and do root-cause analysis.

The view adopted in this paper is that spatially clustered defects indicate the presence of special or assignable causes of variation. Detection and identification of these causes provide important opportunities for process improvement, beyond those obtained from yield-based monitoring procedures. Thus, there is a need for statistical procedures to routinely monitor the wafer map data and detect the presence of defect clustering. In this section, we discuss both graphical and formal methods for spatial process monitoring.

We begin with simple graphical methods that have proved successful in a factory setting. In situations where there is little wafer-to-wafer variability, a display of the *composite* or average wafer map data often

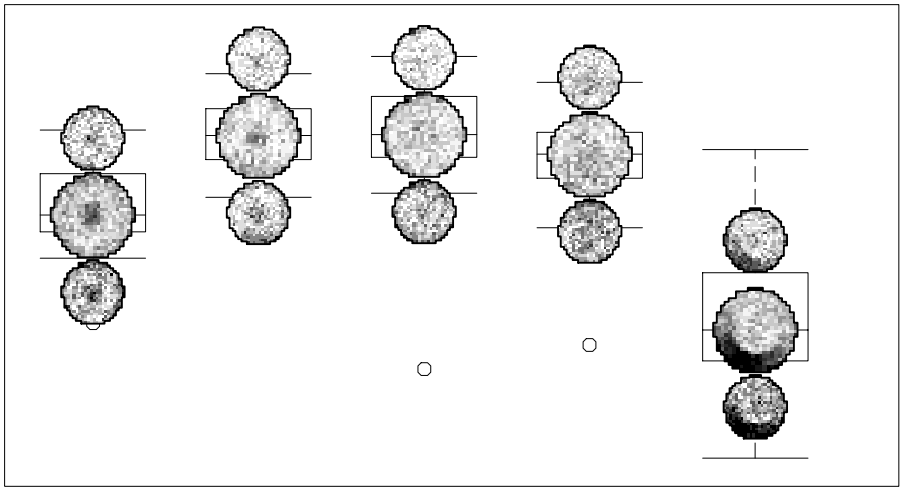


Figure 4: *Extending a simple boxplot to incorporate spatial information collected across a number of lots of the same code. Yield summaries for 5 lots are presented in the order of final probe. (The axes and legend for the greyscale composites are intentionally missing.)*

conveys most of the relevant spatial information. Here, composite wafer map data refers to sitewise (or chipwise) averages of yield from a group of wafers (typically taken from the same lot). Figure 4 shows a particular implementation of this idea that factory engineers at Lucent have found useful. This figure represents static yield reports with spatial clustering information displayed through boxplots of composite maps. In this figure, a lot of 25 wafers was divided into three groups based on their yield: those with yields below the 25th percentile (for the lot), those yielding between the 25th and 75th percentiles, and those with yields above the 75th percentile. The composite wafer maps corresponding to these three groups are added to the simple boxplot to aid engineers in identifying lots in which a large fraction of the wafers experienced a similar processing problem. In this case, the 5 lots in Figure 4 were probed in the order they appear in the figure. We selected lots separated by a week to give a sense of the trend in spatial patterning that is possible. Here we clearly see a gradual improvement as the yield in the center of the wafers improves, only to have another defect mechanism appear. Typically, trends of this sort occur "across the fab" in the sense that new problems affect several similar codes at once.

Plots such as these can be generated daily to summarize the lots probed during the previous 24 hours. Within Lucent, these daily reports are generated routinely and made available through the fab's intranet. Users can also generate custom reports summarizing a single code over a specified period of time.

When there is a lot of wafer-to-wafer variation or if the production volume is large, individual displays such as these may not be very useful or can be time consuming to examine manually. With this in mind,

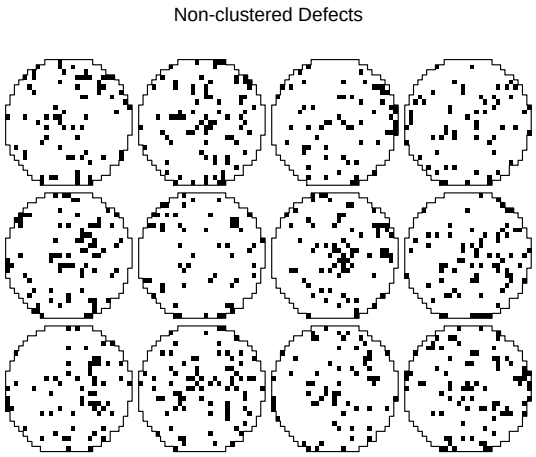


Figure 5: *Samples from a lot that is free of spatial patterning. While engineers classify these patterns as "uninteresting", only two of the wafers pass a strict test for spatial randomness.*

we now discuss more formal monitoring procedures that we have developed to routinely identify lots with significant spatial patterning. Additional details can be found in Hansen, Nair and Friedman (1997), hereafter referred to as HNF.

In order to develop effective test procedures, one needs to specify the appropriate null hypothesis for spatial "randomness." Under the usual notion of complete spatial randomness or CSR (see Cressie, 1993; and Cliff and Ord, 1981), the binary wafer map data (where the chips are classified as being either defective or non-defective) are independent and identically distributed as Bernoulli random variables. Unfortunately, this notion of CSR is an idealization that does not adequately capture the null situations of practical interest in our problem. Figure 5 shows a sample from a lot that a typical product engineer would not want to classify as being spatially clustered. But common tests for CSR will reject the null hypothesis of CSR for most of these wafers. Therefore, to be useful, a reasonable null distribution should allow for some form of mild spatial clustering.

HNF proposed a suitably parameterized Markov random field (MRF) model with small-scale dependence to characterize the null hypothesis of mild clustering. We need some notation to describe this model and the resulting monitoring procedures. Throughout this discussion, we use a script letter when referring to a set and let the corresponding upper case letter denote its cardinality. For example, $\mathcal{W}$ represents the collection of chip locations on a given wafer, while W denotes the total number of chips on the wafer. Let $(X(1), ..., X(W))$ denote the binary wafer map of final probe data with $X(i) = 0$ if chip $i \in \mathcal{W}$ is defective and $X(i) = 1$ otherwise. We let D be the total number of defective chips so that $1 - D/N$ is the observed

wafer yield. Finally, we let $p_i = EX(i)$ denote the probability that the chip at location i is defective, and set $q_i = 1 - p_i$, the probability that the same chip is not defective.

Let $\{w_i(\cdot)\}$ be a set of weights associated with chip i, $i = 1, \ldots, W$. The monitoring procedure in HNF was based on the statistic

$$T \;=\; \sum_{i \in \mathcal{W}} \sum_{j \in \mathcal{W}} w_i(j) X(i) X(j) \;+\; \sum_{i \in \mathcal{W}} \sum_{j \in \mathcal{W}} w_i(j)[1 - X(i)][1 - X(j)]\,,$$

a weighted join-count statistic of neighboring chips that are either both defective or both non-defective (Moran, 1948). Taam and Hamada (1993) also apply this statistic to IC probe data. For identifying clusters that are "blobs" similar to that in Figure 3, we take $w_i(j)$ to be constant in a 3×3 neighborhood around chip i. Near the edge, the weight function $w_i(j)$ is adjusted to leave out chips referenced beyond the wafer boundary. Other neighborhoods and weight functions would be preferable when searching for problems that do not manifest themselves as a blob. For example, a common lithography defect leaves behind a checkerboard pattern of defective chips. In such cases, other choices of $w_i(j)$ will be more meaningful. We will say more about this class of defects in the next section.

Under some mild conditions on our choice of weights $\{w_i(\cdot)\}$, the probability measure that assigns to the vector $\mathbf{X} = (X(1), \ldots, X(W))$ the probability

$$P(\mathbf{X}) \;=\; C \exp\left(\beta_0 \frac{N_1}{W} + \beta_1 T \right) \tag{1}$$

can be shown to be an MRF (Strauss 1977; Cressie 1993). Here, the constant C is a normalizing constant. The parameter β_1 in the model captures the amount of (mild) clustering that one is willing to accept; $\beta_1 = 0$ reduces to the idealized case of CSR. We note that, while the MRF framework has historically been used in the context of hypothesis testing as a convenient class of alternatives to complete spatial randomness, our interest here is in using them to generate more realistic null distributions.

HNF studied the distribution of the monitoring procedure T under the model (1). In particular, let μ and σ^2 be the mean and standard deviation of T under CSR, conditional on the observed wafer yield $1 - D/N$ given in HNF (see also Cliff and Ord, 1981). The mean μ depends only on the yield while the variance σ^2 depends on the geometry of the wafer map and has to be computed numerically. Under a MRF with mild clustering (small values of β_1 in the sense of HNF), T is approximately normally distributed with mean $\mu + \beta_1 \sigma$ and variance σ^2. One can now use this fact to easily develop a large-sample procedure for testing the null hypothesis of mild spatial clustering in (1).

To implement the test procedure, we have to either fix the value of β_1 at an acceptable level of mild clustering or estimate it from historical data. Cressie (1993) discusses several methods for estimating this parameter, the simplest involving a logistic regression. Typically, this cluster parameter will vary from technology to technology and possibly from code to code. Recall that particles are considered as being

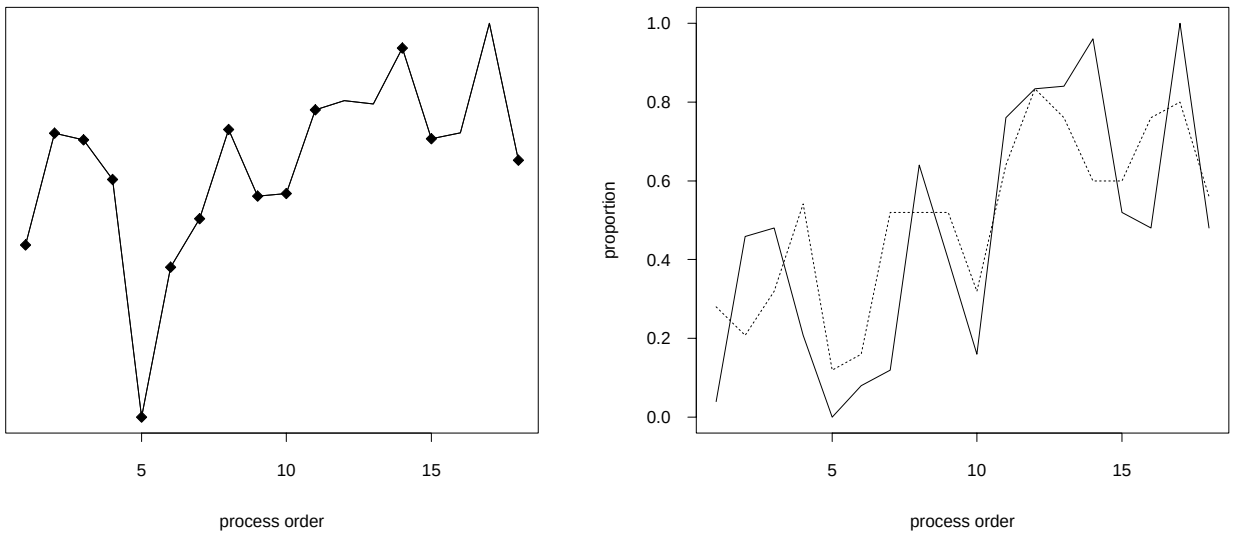


Figure 6: *Monitoring wafer maps data at the lot level. (Left) Information from wafer-level tests for spatial randomness have been added to a chart for monitoring lot-level yield. Black diamonds indicate lots for which more than* 25% *of the wafers exhibited significant clustering. (Right) Simple screening rules to flag lots for further failure mode analysis. The proportion of wafers yielding above plan is given by the solid line, while the proportion of wafers not exhibiting spatial patterning is given by the dashed line.*

responsible for the non-clustered defects. It is well known that the effects of particles will depend strongly on the design of the chip (the number of levels of metal, the line width, etc.) so that the dependence of β_1 on technology or code is to be expected.

Once the values of μ, σ and β_1 have been determined or estimated from historical data, it is straightforward to develop a Shewhart-like monitoring procedure for defect clustering at the individual wafer map level. HNF describe a particular graphical implementation of such a procedure, with the spatial clustering information as a simple addition to the usual $p-$chart for yield monitoring. They also discuss some operational procedures for extending the monitoring procedure to a lot or group level. Of course, in situations with little wafer-to-wafer variability, it is possible to adapt the techniques in HNF to composite wafer maps and monitor the lot-level data directly. Below, we describe another graphical technique for incorporating and displaying spatial monitoring information at the lot level.

Figure 6 is based on information from the five lots in Figure 4 together with 13 additional lots (also processed roughly a week apart). The solid line in the leftmost plot tracks the average yield in each lot with the horizontal line representing the planned yield for the code. (The pattern of improved yield continues over the months following this period, eventually resulting in revised planned yield.) We then tested each wafer in each lot for the presence of spatial clustering using an estimated value of $\beta_1 = 30$. Those lots for

which more than 25% of the wafers exhibit some kind of pattern are marked with a black diamond. We see that even lots with relatively high yields have clustered defects and are labeled as "out of control" with respect to spatial patterning. For example, the wafers in Lot 14 have a center problem that leaves a small, tight cluster of dead chips in the center of the wafer. Thus, even though this lot is high-yielding, there is still an opportunity for yield improvement.

In the rightmost plot of Figure 6, we have drawn the proportion of wafers in each lot that yield above plan (the solid line) as well as the fraction of wafers that do not exhibit any spatial clustering. For the most part, these curves track each other reasonably well. In cases like Lot 4, however, we see a big difference. Here, the overall yield is quite low, but many of the wafers are not patterned (compare the composites for this lot in Figure 4). It is likely that this lot was processed just before a tool was taken off line for a cleaning cycle. As mentioned at the beginning of this section, lots have traditionally been screened for FMA on the basis of simple rules like achieving "80% of planned yield." A plot like the one on the right suggests spatial analogs of these rules. In the next section, we add more information to this kind of display by introducing spatial signatures summarizing the clusters seen in the lot.

4. Identifying spatial patterns

Once the lots or wafers have been identified as having significant clustering, the traditional approach in SPC would be to view root-cause analysis as an engineering problem. However, as we shall see in the sequel, there is considerable information in the "out-of-control" data that can be used for process improvement. In this section, we consider methods for identifying the nature of spatial patterns at the wafer map level and discuss some direct applications. In the next section, we will consider methods for grouping these patterns to create spatial signatures at the lot or group level. These signatures of process problems can then be compared to processing information to identify likely facilities that are possibly out of control.

Given the relationship in Figure 3, we now consider reversing the arrow and segmenting each wafer into large, clustered defects and essentially random background clutter. As mentioned earlier, clusters of defects are typically the result some smooth, spatially varying effect drifting out of specification. In Section 2, we described a furnace problem of this type. We discuss another simple example here. Recall from the Introduction that electronic connections are built up in layers on the wafer through a process that mimics the development of ordinary black-and-white photographs. A thin layer of "resist" is applied to the wafer, exposed to light through a mask, and then those portions not exposed to light are removed. The thickness of the resist can determine whether or not a chip will function. Because this layer is applied while the wafer is spinning, there is a considerable center-to-edge effect in thickness measurements, which changes the probability that a chip will function as you move to the boundary. Clustering occurs when this probability

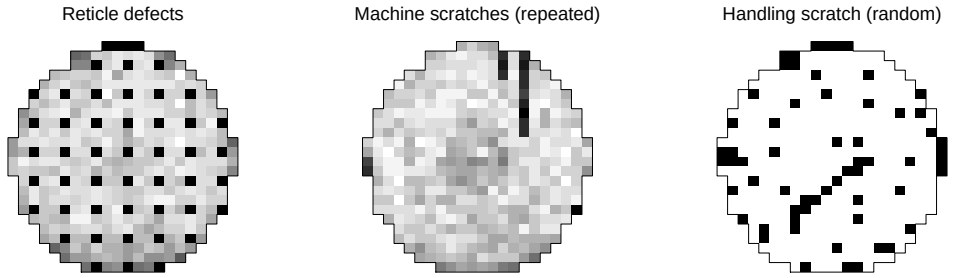


Figure 7: *Examples of "high-frequency" defect mechanisms. In the rightmost wafer map we see an example of a handling scratch on a single wafer. The middle and leftmost maps are lot-level composites. The regular scratches visible in the middle composite were caused by the same machine while wafers were being loaded into its chamber. The regular checkerboard of defects in the leftmost lot occurred because of a small hole in the lower-right hand corner of the* 3×3 *mask used at the photolithography step.*

drops near zero, and a ring of dead chips at the edge is the result. Traditionally, yield losses of this type are labeled as "parametric," because some critical parameter (i.e. resist thickness) has drifted beyond specification limits. In general, these parametric effects are driven by smooth surfaces which control the overall probability that a chip will function.

Compare this type of failure to say particulate defects or scratches. Here, we do not get large clusters, but instead "high-frequency" patterns (a typical scratch is rarely more than 1-chip wide). The rightmost map in Figure 7 is a good example. Completely random scratches such as this are the result of improper handling and should be disappearing in newer fabs. (New robotic equipment makes it increasingly unnecessary for an operator to have direct contact with the wafers. Most fabs now employ a so-called Smif technology so that wafers are moved through the line in an air-tight box.) Machine-related scratches, however, are much more common, but tend to repeat across a large fraction of the wafers in the lot. This makes it easy to identify the problem simple from a lot-level composite (see the middle map in Figure 7). Another process problem that does not fall into the "parametric" category comes from the lithography step. The mask mentioned above (the analog of a negative in photography) typically prints a 2×2 or 3×3 grid of chips at once time. If there is a small hole or tear on this mask, it will produce a repeating defect that occurs in every third or fourth column and row for virtually every wafer in the lot (see the rightmost map in Figure 7). Again, these high-frequency failures tend to be easily identifiable via simple thresholding techniques applied to lot-level composites.

In this section, we will focus on parametric failures, and hence we will attempt to identify clusters with smooth boundaries. There are a myriad of approaches that could be applied to address this problem, from direct modeling through Markov random fields and the associated Bayesian computational techniques (Besag

and Green, 1993), to greedy segmentation algorithms that would "grow" clustered regions (Zhu and Yuille, 1996). Some of these methods are the subject of ongoing research and will be mentioned at the end of the section.

We have implemented a straightforward smoothing and thresholding approach to identify spatial patterns. We choose an ordinary kernel smooth, with a kernel that adapts to the edges of the wafer. For each $i \in \mathcal{W}$, we let $\mathcal{N}_i$ denote the collection of chip locations that we take to be the neighbors of chip i and let N_i denote their number. To each chip in $j \in \mathcal{N}_i$, we assign a weight $w_i(j)$ and we require that the sum of the $w_i(\cdot)$ over the neighbors of chip i be equal to one. Typically, we choose neighborhoods $\mathcal{N}_i$ that deform as chip i approaches the boundary of the wafer, although we could instead simply readjust the weights near the boundary as we have done for the test statistics from the previous section. We have adapted our kernels in this way because during fabrication, chips near the boundary are exposed to similar processing conditions, conditions that can vary quickly as you move from the center. Therefore, when smoothing near the boundary, we do not want to reach into the center of the wafer to borrow strength.

Next, for each $i \in \mathcal{W}$, we define the kernel estimator

$$\hat{p}_i \;=\; \sum_{j \in \mathcal{N}_i} w_i(j) X(j). \tag{2}$$

Under the null MRF model in the last section, the $X(i)$'s are binary random variables with constant defect probability $p_i = p$, $i \in \mathcal{W}$ (this is achieved by restricting the support of the kernel as well as its values in the MRF neighborhood). By conditioning on the yield of the wafer, we can set $p = D/W$. The presence of clusters of defective chips will result in areas in which $\hat{p}_i$'s are consistently larger than D/W. It is possible to use the wafer map with these $\hat{p}_i$'s in order to visually characterize spatial clustering, should it exist. However, it is difficult to assess this information when comparing wafers with different overall yields. Ideally, we would like to separate the clustering information from the wafer yield. One way to do this to obtain sitewise $p-$values that also account for the statistical variability in the estimates. The sitewise $p-$value is the probability that $\hat{p}_i$ is greater than p for each chip $i \in \mathcal{W}$. We used a normal approximation to obtain the $p-$values in our computations. Examples are given in Figure 8. As a visual tool, this procedure is similar to the probability maps of Diggle (1984).

Once a smooth map of $p-$values has been created, we can produce a segmentation into random (small-scale) and parametric (large-area) defects by thresholding. An illustration of the technique is given in Figure 9. A binary probe map (the lower left hand map in each panel of Figure 8) is smoothed via (2) and a map of sitewise $p-$values is obtained. A series of thresholds is then applied to this map. For threshold t, each defective site i corresponding to a $p-$value that is strictly larger than t is classified as being part of the parametric defect. The distributional results in HNF are used to test the null hypothesis that the pattern of defective chips (conditional on the identified cluster) in the remaining region appears "random" as modeled

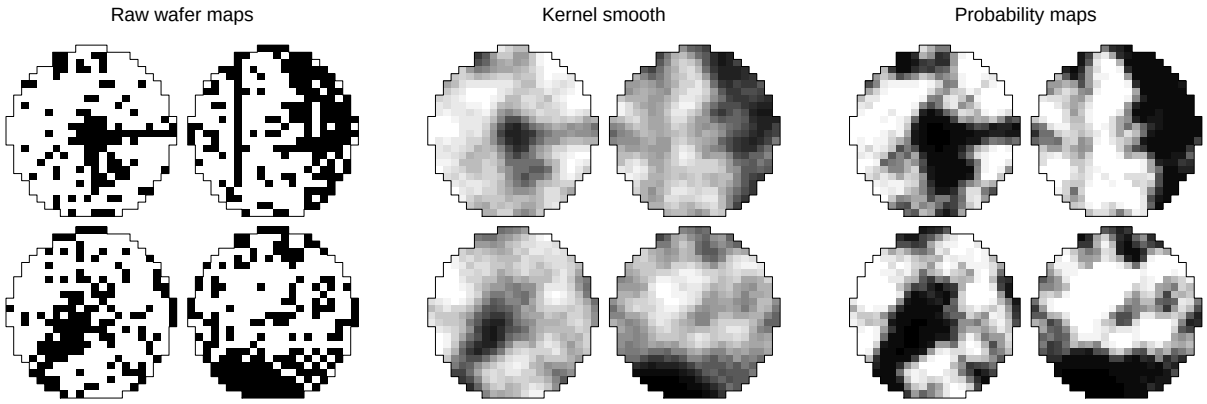


Figure 8: *Smoothing binary probe data to highlight spatial effects. A sample of wafer maps (left) are smoothed with* 5×5 *kernel (middle) and a probability transformation is applied (right).*

by the null MRF model described previously. If the null hypothesis is rejected, we decrease the threshold, leading to a larger region of clustered defects, and we repeat the test. The sequential testing stops as soon as we accept the null hypothesis of randomness for defects not belonging to the large-scale cluster. At each step in the process, we examine the join-count statistic T described in the previous section. Once standardized, this statistics has (approximately) a standard normal distribution under the null hypothesis. In Figure 9, we illustrate this procedure, with a critical value of 0.95 for our sequential tests. (We have studied more elaborate methods to correct for the sequential nature of the procedure with repeated testing but these yielded very similar results.) In order to make this scheme computationally tractable, we interpolate the Z-scores calculated at a modest grid of threshold values.

In Figure 9, the wafer maps illustrate how a cluster (displayed in red) is grown as the threshold drops. The chips classified as being part of the background are colored in blue. The optimal threshold in this case is 0.98. (The values chosen in this way are typically greater than 0.9.) In the right hand panel of Figure 9 we present the identified clusters from the samples in Figure 8. Comparing these two figures, we see that the sequential thresholding procedure is performing well except for the wafer map in the top right hand corner. This wafer has a long scratch that is smoothed over by our procedure. As mentioned above, we can remedy this problem by choosing a different type of neighborhood structure in our smoothing procedure. Our focus here was on large-area clusters, and as a result we chose a kernel with large support. As a final note, it is possible to relate this technique to a scheme for approximately maximizing a certain Bayesian posterior distribution. For this, we assume that the prior of the parametric defect is a Markov random field having neighborhood structure that agrees with the support of our kernel.

Engineers can now use these identified patterns of spatial clustering as a first step in their root-cause analysis of process problems. We discuss a statistical approach in the next section that uses these identified

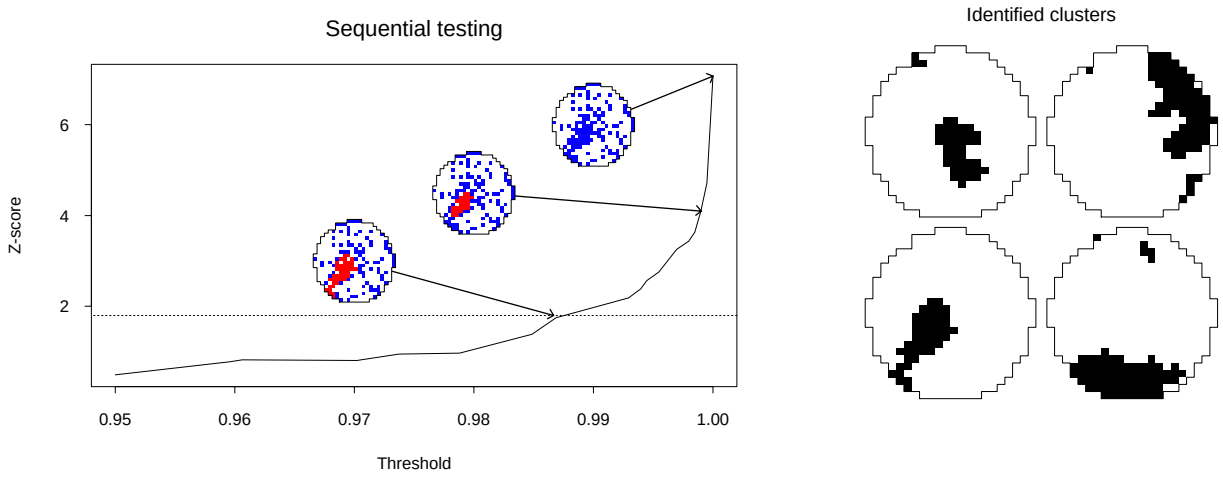


Figure 9: *Thresholding to create a segmentation of defects into large- and small-scale effects. A sequential testing procedure is proposed that highlights clusters.*

patterns to develop lot or group-level signatures of patterns. In the rest of this section, we discuss another important application of this segmentation scheme.

There has been considerable interest in the semiconductor manufacturing literature over the last thirty years on developing yield models (see Friedman, Hansen, Nair and James, 1997 and references therein). Much of this interest has focused on decomposing the yield into metrics associated with parametric (large, clustered) and random (small-scale clustered) defects and analyzing the random defect density as a function of technology, code, and chip size. Early work was based on a simple Poisson model for the number of particles falling in each chip under the assumption that such defects are distributed uniformly across the wafer. This results in the expression $Y = \exp(-D_0 A)$ for the probability that a chip of size A is defect free (will "yield," hence the variable label Y). The quantity D_0 is the intensity of the homogeneous Poisson process assumed to be scattering particles in the fabrication line. However, this yield model was seen to be inadequate in modeling actual yield data which exhibited large-area defect clustering. This spurred researchers in the area to consider various alternatives including mixtures of Poissons. One alternative that has become popular in the yield modeling area is the Murphy model (Murphy, 1964) given by $Y = Y_0 \exp(-D_0 A)$ where Y_0 is an "area usage factor" meant to describe the impact of clustered or parametric defects. Typically Y_0 and D_0 are estimated from lot-level yield data (i.e., the simple yield summaries), completely ignoring the spatial arrangement of defective chips. (See Cunningham, 1990, for an extensive overview of yield modeling.)

Friedman et al. (1997) discuss the use of the above segmentation technique to separately estimate the yield metrics Y_0 and D_0. The former is simply the "size" of the clustered defect, while the latter is a transform of the background yield, each parameter being estimated on a wafer-by-wafer basis. Friedman et

al. (1997) compare this method with the time-honored "windowing" technique popular in the semiconductor manufacturing literature and show that their approach is decidedly superior. They also propose the use of control charts for separately monitoring the Y_0 and D_0 yield components from final probe data.

Finally, we briefly describe our ongoing research on other approaches for cluster identification. Currently, we are considering parametric models, employing Zernike polynomials (see Born and Wolf, 1964), to describe the spatially smooth component of yield loss (Ionides, 1999). The Zernike polynomials form a complete set in L_2 (defined over a disc), but low-order terms appear to be sufficient to capture a significant portion of the patterns observed in practice. One can use quasi-likelihood (Cressie, 1993) rather easily to identify the smoothly varying spatial component of yield loss that can be traced to processing conditions. Formally, maximizing the quasi-likelihood involves a logistic regression, and because of the flexibility required of the bases in this context, we have experienced discouragingly frequent convergence problems. Regularization of some sort is necessary before scaling this procedure to meet production volume. We have found, however, that even with these amendments, the simple procedures we have described above yield virtually equivalent results. The number of wafers to be processed each day dictates that candidate procedures must perform reliably and with relatively short computing times.

5. Creating spatial signatures

In general, spatial defect patterns can vary both between and within lots. In situations where there is little wafer-to-wafer variability, the spatial patterns can be summarized easily by lot-level composite maps. In other cases, however, much more subtle process problems can arise as a result of complex interactions between machines on the factory floor. It is common to see only a few wafers in a lot exhibiting significant spatial clustering or to have several different types of patterns caused by different machine or process problems. In such cases, it is more effective to combine processing information from wafers in different lots exhibiting similar patterns of defects and use this to identify the machines or facility.

In this section, we will describe one approach to identifying common patterns at the lot or group level based on hierarchical clustering. Our strategy has been to design easily-understood, interactive displays which allow engineers to make decisions about practically important differences between patterns based on their knowledge of the process. The approach described here has met with considerable success in our factory settings. However, there are clearly other methods that can also be used for this purpose. We are also currently studying the use of alternative methods based on the parametric model discussed in the last section.

In Figure 10, we present probe maps for a 25-wafer lot. Recall that the different colored squares represent different failure modes, and that only the non-defective chips are white. A cursory inspection reveals that

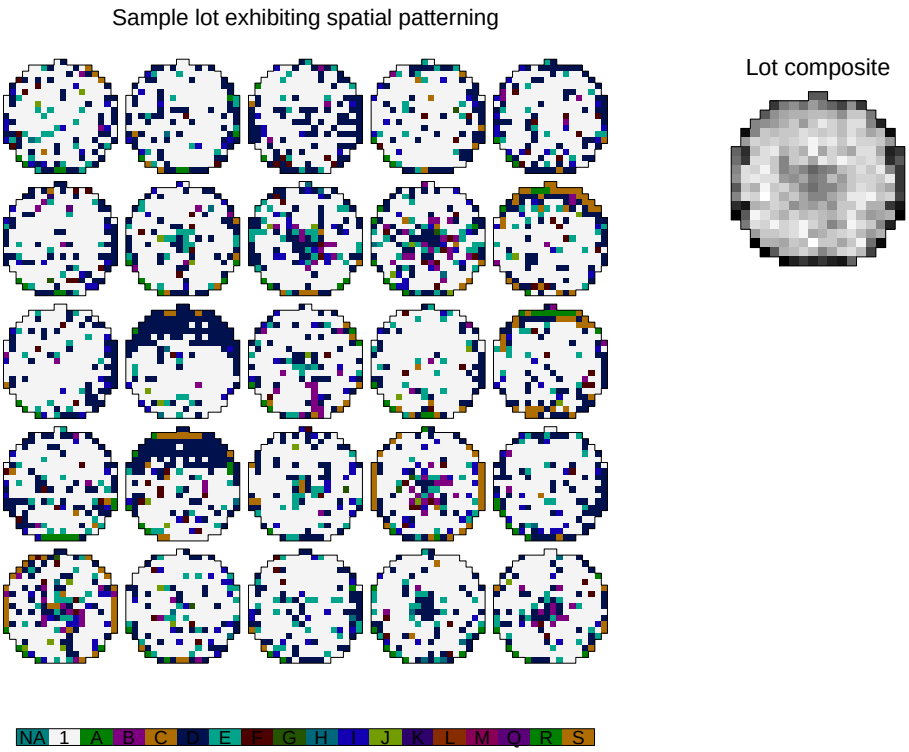


Figure 10: *Final probe data for a sample lot. Failed chips are represented by colored squares with white regions corresponding to non-defective chips. The legend gives dummy names to the failure modes. The map on the right is a simple sitewise composite.*

there are at least two types of patterning present in this lot. About 10 of the maps have clusters of defects in the center of the wafer, while four exhibit a problem along the upper edge (similar to that in the middle map of Figure 3). The composite for this lot is given in the right hand panel of Figure 12. We find that 90% of the wafers also have a pronounced edge effect. Historically, the boundary of the wafer was extremely difficult to manage, and most of the yielding sites were in the center (recall the "area usage factor" described in Section 4). In modern fabs, the problem is much less pronounced, killing consistently only one or two chips on the border. Here, the loss of chips is likely due to some combination of resist thickness and plasma etch.

To add spatial information beyond what could be seen in either the composite or the box-plot display for this lot, we first produce a wafer-by-wafer segmentation into large-scale (clustered) and small-scale (spatially random) defects. The automated thresholding procedure described in the previous section was applied with significance level of 0.95. Then, we use hierarchical clustering to group maps of the identified large-scale defects. While we have tried several options for this procedure, the best involves clustering on the thresholded

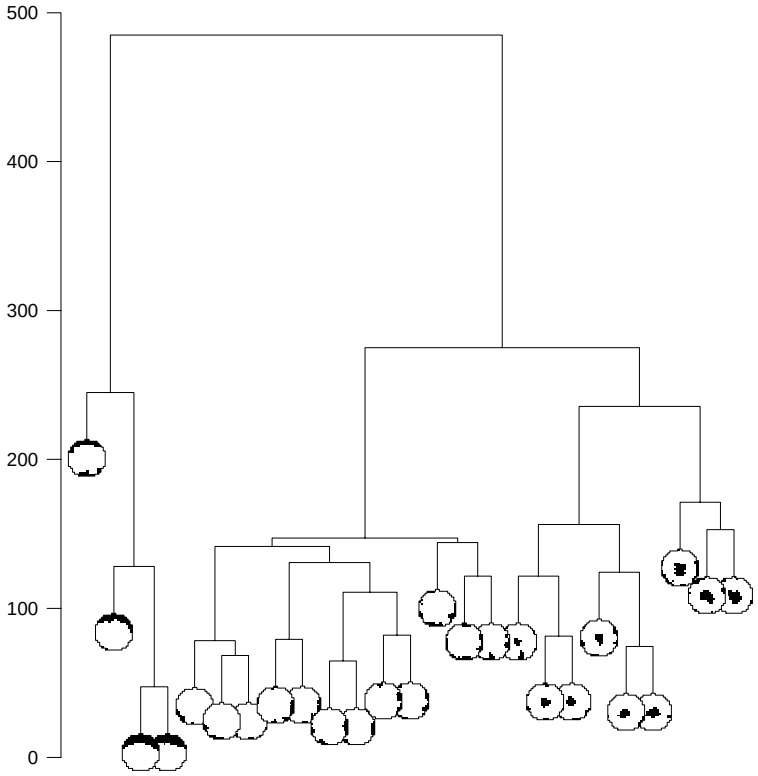


Figure 11: *Hierarchical clustering display based on the thresholded wafer maps from Figure 10.*

maps. The raw, binary probe maps are simply too noisy, and they are unduly influenced by the differences in yield across wafers. We also found that using the smoothed maps in Figure 8 do not separate the clusters as well as the thresholded maps. In addition, we have experimented with various metrics for measuring distance between wafer maps, but found that the usual L_1 distance performs adequately in most cases.

Wafers that pass our initial test for randomness (those with a threshold value of 1) are not included in the cluster analysis. In the above case, each map showed some degree of patterning so the analysis involved all 25 wafers. The final cluster tree is shown in Figure 11. At each leaf node, we have superimposed the thresholded maps to give a sense of the patterns. Interactive versions of this display allow the user to cut the tree at any node, replacing the subtree with a composite of the selected wafers. Historical data from several weeks can be analyzed quickly in this way. This device is routinely used by product engineers who are interested in tracking yield loss for one or more selected codes.

In Figure 12 we use the cluster tree to identify three patterns or spatial signatures. The leftmost panel in this figure is a boxplot, augmented with composites as described in Section 3. The edge problem is clear in each composite group, as is the heavier yield loss in the center. In the right hand panel, we present the identified spatial signatures. These are simply composites of the thresholded maps (the information from the

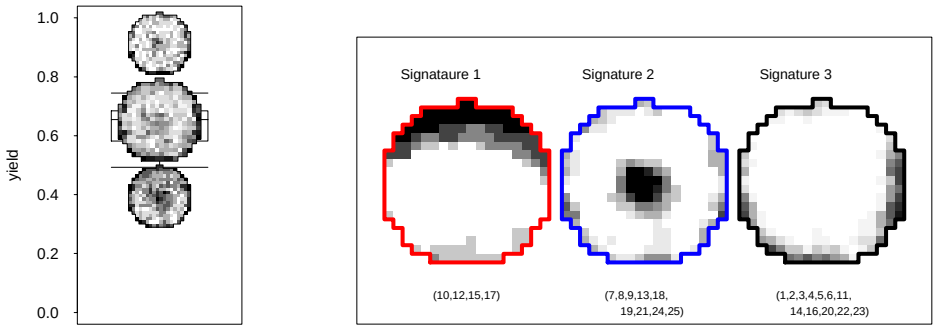


Figure 12: *Spatial data analysis. The left hand panel shows a simple boxplot described in Section 3. The edge and center yield loss problems are clearly evident. The lowest yielding 25% of the wafers seem to have an additional problem near the top edge. In the right hand panel, we present a set of defect "signatures." These were produced by hierarchical clustering as described in the text and shown in Figure 11. Cluster membership is given at the bottom of the figure, where the indices refer to maps in Figure 10 (where 1 is the upper left hand wafer, and we count from left to right and top to bottom).*

colored borders will be used in the next section). From this display, we see that in addition to the pronounced center and edge problems, four of the wafers have a crescent-shaped defect along the upper edge. This type of lot-summary is routinely produced for wafers processed in the fabs at Lucent. It is often helpful to add the percentage of chips lost to each signature, assessing its overall impact. In the next section, we relate these maps to actual process information and identify the processes responsible for these patterns.

There are clearly many options for identifying clusters of this type. We have chosen hierarchical clustering because the resulting interactive displays are quite natural for the product and process engineers. We acknowledge that this approach may not adequately capture the natural variations in particular types of patterns. Because we have chosen to develop interactive tools, this shortcoming is mitigated somewhat in that an engineer can group "equivalent" patterns easily. This idea of incorporating the spatial information directly into the output from a standard clustering routine is fairly powerful. The S-software that enables this type of display is described in Hansen and James (1997). As noted earlier, we are studying the use of alternative parametric methods for this problem. Nevertheless, the simplicity of the approach outlined in this section has proved quite useful in practical applications.

6. Improving the process

The final and critical component of the strategy consists of relating the spatial signatures to available process information in order to identify potential sources of problems. The methods that can be useful here will depend on the type of process information that are collected, how much engineering or historical knowledge

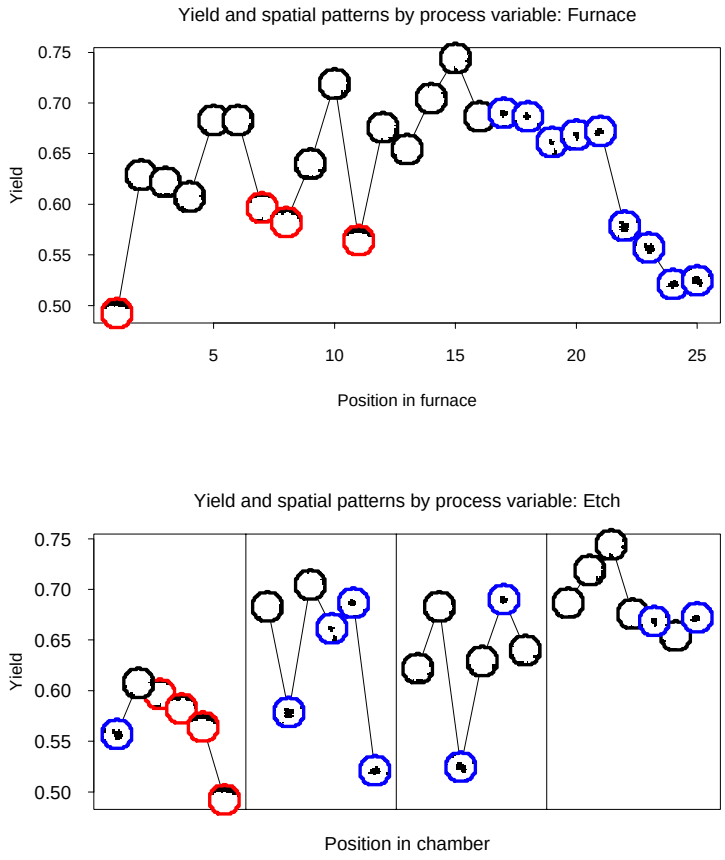


Figure 13: *Relating patterning to process information.*

is available about the technology and the manufacturing process, and so on. Often, engineering judgment is relied upon to make the connection between patterns of defects and their possible causes. For example, if a half moon pattern of shorts or opens is exhibited by ten wafers in the lot, the engineer might quickly conclude that the problem is a bad implanter. While this type of diagnosis will be effective for known patterns, new technologies and new designs place heavy demands on manufacturing, frequently producing new interactions and new patterns.

In most IC fabrication facilities, some type of database is used to record SPC and reliability data on the various machines. This data base tracks information on service times, down time, and the cause of any service interruptions for all processed lots. If a lot is found to have what is believed to be process related problems, the lot can be traced through this system to see if any processing problems were encountered, and in turn to see if any other lots processed at the same time exhibited similar patterns. In addition to machine

maintenance schedules, most facilities also compile a shop flow data base that records the paths followed by lots as they wind through the manufacturing line. Often, this routing information can be augmented with the order in which wafers in the lot were processed at each step. Since wafers are routinely shuffled in the line, process order can often uniquely define a process step or suite of likely steps. If we assume that an observed pattern of failures is process-related, then it is logical to assume that if a particular step could be found where all wafers displaying that pattern were processed contiguously, that step would be a likely cause of the problem.

We discuss an application of this idea in the the two plots of Figure 13. The upper panel is a plot of wafer yield against processing order. To this plot, we have added thresholded maps (calculated in Section 4) indicating the pattern exhibited by the corresponding wafer. Colored borders help to relate spatial clustering to process order. Here, we have used the data from Figure 10, and present the plot for just one step: wafer position in a furnace. In this case, the center problem worsens as we move along the furnace. The remaining signatures have little correlation with this step. It is known that temperature variations in the furnace can (in concert with other process steps functioning near the boundary of specification limits) produce these patterns. The interesting feature of this example is that examining yield as a function of process order would not identify this step as the likely culprit. In this way, spatial analysis is an invaluable component in post-production failure mode analysis.

In the lower panel of Figure 13, we present a similar plot, this time for process order through the four chambers of a plasma etcher (denoted by vertical lines in the plot). In the first chamber, we see a slow degradation in yield. In this case, wafers are loaded into the chambers one at a time. As the lot was processed, the conditions in this chamber drifted out of specification limits. In this case, examining average yield per chamber might identify this particular machine. By augmenting spatial information, we guarantee it.

So far, we have considered interactive methods based on simple plots that incorporate spatial information. Many rules such as those discussed in the previous two paragraphs can be built into an automated system that routinely relates the patterning identified in our lot summaries with relevant process information. This approach has been used within the fabs in Lucent to aid FMA efforts.

7. Technology Transfer and Fab Impact

We have made reference at several points in the paper to software tools for displaying mapped wafer data and Web-based reporting schemes that disseminate information about the spatial clustering evident in final probe data. Over the course of our involvement in this research project, a library of S-PLUS functions has been created to organize the type of computations demanded by our analysis. This collection of tools, called

S-wafers (Hansen and James, 1997), provides an environment for spatial data analysis tailored to questions relating to IC fabrication. Through the object-oriented facilities in S-Plus, it is possible to transparently (from the viewpoint of the user) extend routine analyses suggested by traditional SPC, incorporating the special structures inherent in complex, manufacturing data. We have seen this, for example, in the case of the waferized boxplots of Figure 4. We take this opportunity to tout the benefits of well-designed statistical software as a tool for technology transfer.

The S-wafers platform began as a research tool to help us understand the patterning evident in final probe data. At that point, most product, process and FMA engineers recognized the importance of retaining the spatial character of these data. Unfortunately, very little was done in terms of routine analysis because the existing software was too restrictive. Over time, the S-wafers tools migrated to the fabs and provided a natural bridge for technology transfer. New views of the data gave way to interactive graphical displays that, with input from the engineers, led to standard reports and enhanced process monitoring. From a software perspective, several ingredients have made this approach possible:

- A powerful graphical and analytical engine provided by the underlying S language,
- The encapsulation of spatial information into objects in the language, and
- The definition of methods for manipulating these new classes of objects.

Although the range of "specialized" analyses made possible with S-wafers is fairly extensive, an object-oriented approach allows users to invoke familiar S commands that extend automatically to use the spatial information. Finally, the benefits of such a system go beyond a simple accounting of what can be computed. We have found that the availability of software for visualizing and manipulating complex data directly influences how users frame process improvement problems. From yield models to metrics that track the health of the fab, the development of important concepts relating to manufacturing data can be limited by the available software.

Automated reports resulting from our collaboration are now available through the company's intranet to engineers in each of the major fabs. (Because these sites are geographically disparate – two in the United States and one in Spain – the use of the Web has also helped to fuel cooperation in tackling common processing problems.) An example is given in Figure 14. Recently, we have also developed Java applets to further enhance the interactive capabilities of S-wafers, offering greater analysis opportunities. A sample display is presented in Figure 15.



Figure 14: *The S-wafers home page, offering a variety of daily (spatial) yield reports as well as interactive displays of probe data.*

8. Acknowledgments

Our research has benefitted greatly from discussions with and input from engineers in Lucent Microelectronics. David James and Daryl Pregibon also provided valuable input at various phases of this project. V. N. Nair's research was supported in part by NSF-DMS grant 9803281 and AFOSR/ARPA MURI grant F49620-95-1-05.



Figure 15: *An interactive Java applet for analyzing selected lots of final probe data.*

References

Besag, J., and Green, P. J. (1993) "Spatial statistics and Bayesian computation," *JRSS B*, **55**, 25-37.

Born, M., and Wolf, E. (1997) *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, Cambridge, UK: Cambridge University Press.

Cliff, A. D., and Ord, J. K. (1981) *Spatial Processes, Models and Applications*, London: Pion Limited.

Cressie, N. (1993) *Statistics for Spatial Data*, New York, NY: John Wiley & Sons.

Cunningham, J.A. (1990) "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Trans. Semicond. Manufact.*, **3**, No. 2, 60-71.

Friedman, D. J., Hansen, M. H., Nair, V. N. and James, D. A. (1997) "Model-free estimation of some yield metrics in integrated circuit fabrication " *IEEE Transactions on Semiconductor Manufacturing*, **10**, No. 3, 344-359.

Hansen, M. H. and James, D. A. (1997) "A computing environment for spatial data analysis in the microelectronics industry," *Bell Labs Technical Journal*, **2**, No. 1, 114-129.

Hansen, M.H., V.N. Nair, D.J. Friedman (1996) "Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects," *Technometrics*, **39**, No. 3, 241-253.

Ionides, E. (1999) "Spatial models for wafer yield: Formulation and applications," unpublished manuscript.

Moran, P. A. P. (1948) "The interpretation of statistical maps," *JRSSB*, **10**, 243-251.

Murphy, B.T. (1964) "Cost-Size optima of monolithic integrated circuits," *Proc. IEEE*, **52**, 1537-1545.

Strauss, D. (1977) "Clustering on coloured lattices," *J. Appl. Prob.*, 14, 135-143.

Taam, W. and Hamada, M. (1993) "Detecting spatial effects from factorial experiments: An application in integrated circuit manufacturing," *Technometrics*, **35**, 149-160.

Zhu, S. C., and Yuille, A. (1996) "Region competition – Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE PAMI*, **18**, 884-900.