Process and Productivity Improvement in Semiconductor Manufacturing

Mark H. Hansen
AT&T Bell Laboratories
Murray Hill, New Jersey
cocteau@research.att.com

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

Abstract: As part of an on-going research effort with engineers in AT&T Microelectronics, we have developed techniques for analyzing data collected during the integrated circuit (IC) fabrication process that systematically exploit the spatial character of these data. This approach includes simple, but powerful wafer visualization, the detection of defect signatures, and spatial analysis of variance to analyze data from final probe, end-of-the-line electrical tests, in-process particle scans and split-lot experiments. In this paper, we present an overview of these procedures as they have been implemented in three of AT&T's IC manufacturing sites.

1 Introduction

The fabrication of integrated circuits (IC's) is a complex process, involving hundreds of separate steps and lasting up to twelve weeks. Several hundred IC's, or chips, are fabricated simultaneously on a *wafer*, the exact number depending on the particular technology. The wafers themselves are processed together in groups called *lots*. Upon completion of the fabrication process, each chip on every wafer is subjected to a number of functionality tests and is declared to be either good or defective. Current quality control procedures are typically based on lot level summaries such as yield (the proportion of good chips in the lot). In a factory setting, these summaries are in turn used to design rules to identify problem lots:

- *Reject a lot if the overall yield is less than x%*
- *Reject a lot if the good-to-functional ratio is less than y%*

Here the term "functional" applies to a chip that is working but cannot be classified as good because it does not operate within specification limits. While such rules are easy to implement, they completely ignore the information about the spatial distribution of defective chips contained in this data.

For the most part, one-dimensional summaries like yield are motivated by the assumption that defects are distributed at random across the wafer. Unfortunately, very few processes in IC manufacturing are spatially homogeneous, and hence simple statistics can offer only an incomplete and often misleading view of the data. For example, in the left hand column of Figure 1, we present the final probe data for two wafers. In each case, the black squares represent defective chips, while the remaining white squares denote good chips. In the right column of the same figure, we display data from a particle scan of these wafers at two different steps in the process. These two maps share a common grey scale which assigns darker values to chips having a large number of particles. While particulate defects are a serious problem in IC manufacturing, they can impact the overall yield of a wafer differently, depending on their size and the stage at which they were introduced. In this case, each wafer has roughly the same overall yield and roughly the same total particle count. By inspecting the two sets of maps, however, we find that the particles represented in the grey level plot in the lower half Figure 1 have had a greater impact on the final yield. Therefore, from a process improvement standpoint, we should focus our attention on the step at which this particle scan was taken.

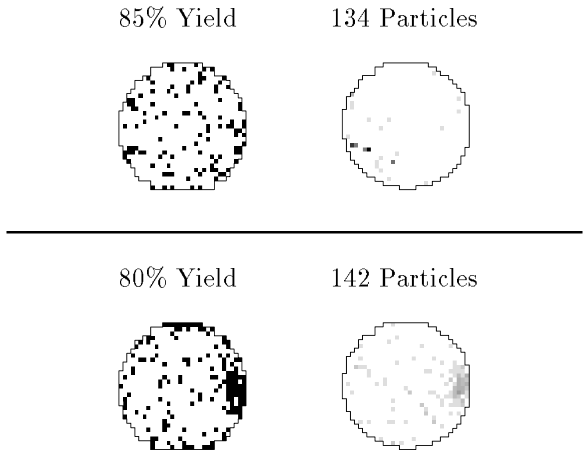

Figure 1: Final probe yield and particle scans of two wafers.

In designing tools for process and productivity improvement in this area, we have attempted to respect the spatial character of the data collected during the manufacturing process. In the pages that follow, we will briefly describe our current strategies for analyzing data from final probe, particle scans and split-lot experiments. Other aspects of this work can be found in a paper (also appearing in this volume) by Lorraine Denby and David James (Statistics Research, AT&T Bell Laboratories).

2 Monitoring Probe Data

The defective chips on a wafer are either spatially random in nature or spatially clustered, and clustered defects are typically classified as being either particle or process related. Spatially random defects, on the other hand, are often the product of the cleanroom environment. This type of problem generally does not have an assignable cause and can only be reduced through long term gradual improvement or expensive equipment overhaul.

In general, we will assume that the arrangement of defective chips on a given wafer is in fact a superposition of two mechanisms. Acting in isolation, the first generates defective chips uniformly across the wafer, while the second produces clusters of defects. In Figure 2, we illustrate how these two mechanisms combine to yield an observed map. In statistical process control terminology, the spatially random defects are due to a *common cause*, while the clustered defects are the result of a *special* or *assignable cause*. Therefore, from a process improvement point of view, we want to identify clusters of defects from final probe data; in essence, reversing the arrow in Figure 2.

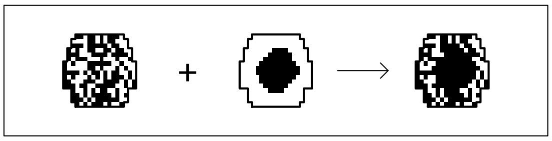

Figure 2: An idealized view of two destructive forces working independently to create the observed wafer map. This configuration has 250 chips per wafer. The cluster of defective chips in the middle of the wafer is often the result of a problem in the plasma area.

Toward this end, we have developed a two step procedure of smoothing and then thresholding a given wafer map. From each observed binary wafer map, we generate a smoothed map that is based on local averages (local defect probabilities). To these values we apply a variance-stabilizing transformation to account for the overall yield of the wafer followed by a probability-integral transform (after appropriate centering and scaling to account for possible edge effects). By construction, the resulting map takes on values in $[0, 1]$, and has the property that given the yield of the wafer, chips with too few defective neighbors are mapped to values near zero, while chips with too many defective neighbors are assigned values near one.

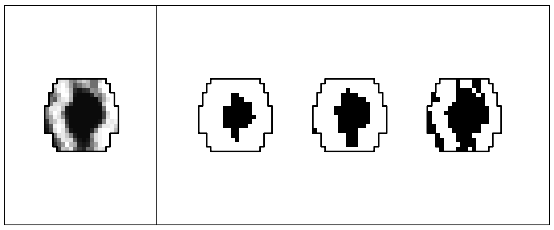

Figure 3: Thresholding the smooth map on the left at 0.99, 0.9 and 0.5.

The size and location of possible clustered defects can now be estimated by choosing a threshold value for our smoothed maps and coloring black all defective chips with values above the threshold. The remaining chips will be colored white. The two extreme threshold values of one and zero have the following interpretations. A threshold value of one means that all the chips will be turned white, implying that all of the defective chips in the observed wafer map were generated randomly. On the other hand, a threshold value of zero implies that all the defective chips in the wafer map are part of a cluster.

In the left-hand panel of Figure 3, we present the smoothed version of the observed wafer map appearing in Figure 2. The remaining three maps in this figure were obtained by thresholding this smoothed map with threshold values of 0.99, 0.9 and 0.5, respectively. By lowering the threshold value in this way, we create a sequence of nested clusters. As can be seen by comparing the results in Figure 3 with the original clustered defect in Figure 2, if the threshold is too large we underestimate the size of the cluster. On the other hand, a very low threshold leads to too many chips being included in the cluster and the resulting estimate becomes visually displeasing.

To choose a reasonable threshold, we have developed a procedure that iteratively applies a test for complete spatial randomness. Recall that under our framework as illustrated in Figure 2, defects are either spatially random or spatially clustered in nature. Therefore, if we have properly identified the clustered defects, the defective chips in the remaining part of the observed wafer map should not exhibit any spatial patterns and should pass any test for spatial randomness. Thus, we begin with a threshold value of one which corresponds to declaring that the entire wafer consists of random defects and apply a test of spatial randomness. If this test fails, we reduce the threshold value, identify the corresponding clusters and, conditional on these chips being defective, apply the test of spatial randomness to the remaining part of the wafer. The test proceeds in this way sequentially, until we reach a threshold value at which the hypothesis of complete spatial randomness is accepted.

In Figure 4 we illustrate how our strategy for choosing the threshold value works in practice. In this graph, we plot threshold level versus the Z-score of the associated randomness test. The chips in each wafer map are shaded so that good chips are white, chips belonging to the estimated cluster are black, and the remaining defective chips are grey. We refer this plot as a *threshold profile*.

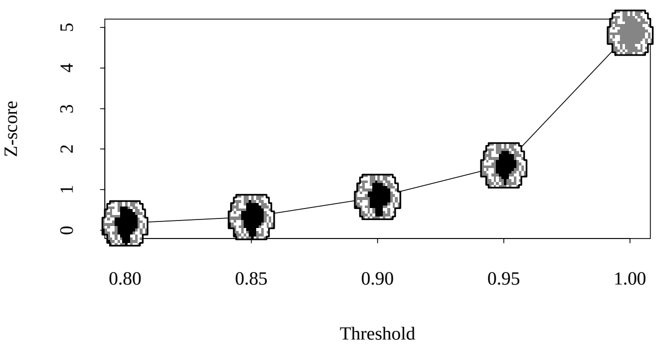

Figure 4: Threshold profile for threshold levels between 0.8 and 1.0

There are many different tests for spatial randomness available in the literature, see [Cressie, 1991], and several of these are based on *join-count* statistics, see [Moran, 1948]. We have investigated the properties of the join-count statistics and have found a particular linear combination of them to perform well in our context. In addition, we have extended the usual test procedure to the conditional situation, where we need to test for spatial randomness in part of the wafer map after conditioning on the observed pattern in the remaining part. The details of this procedure as well as the use of join-count statistics for process monitoring are given in [Hansen et al., 1995], see also [Taam and Hamada, 1992]. Not surprisingly, we have also auditioned many different methods for smoothing binary wafer maps other than the simple technique described here. Spaces of bivariate splines, for example, have been used in a logistic regression model to achieve a similar goal. (In the next section we discuss applying this spline-based methodology to particle scans.)

It has been our experience that routinely identifying significant patterns in wafer map data and relating them to possible causes and critical process variables is a cost-effective approach to yield and process improvement. The automatic procedure for identifying clusters of defects described above is the starting point for this type of analysis. The interested reader is referred to [Denby and James, 1995] (in this volume) for more complete a discussion of how this strategy is implemented in practice. Finally, as a byproduct of this analysis, we can accurately estimate both the extent of spatial clustering and the intensity of the random defects, two quantities that have been examined in the literature on semiconductor manufacturing in the context of yield modeling, see [Friedman et al., 1995].

3 Monitoring Particle Maps

As mentioned in the introduction, particle scans of product wafers are taken at various stages in the process. Unfortunately, very little is known about how particles introduced at different levels impact the final probe yield. Mimicking the development of the previous section, we wanted to smooth this data in some way, and isolate common patterns of particle deposition.

If we assume that particles are deposited on a wafer according to a homogeneous Poisson process, then a simple smoothing procedure can be derived that is analogous to the smooth for binary wafer maps (based on transformations of local averages). In the top half of Figure 5 we present (the square root of) chipwise particle counts for four wafers. In the lower portion of this figure, we present the smoothed maps which again take on values in $[0, 1]$ and have the property that the darker regions indicate areas where there are too many particles given the average count per chip across the wafer.

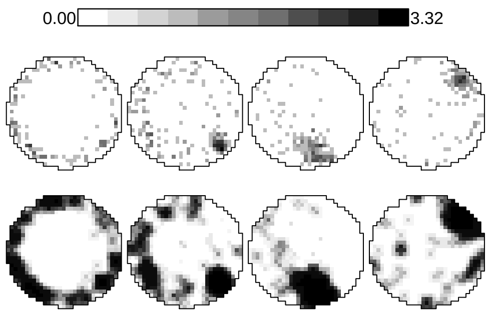

Figure 5: Smoothing particle maps.

Starting instead with the assumption that particles are deposited across the wafer according to a nonhomogeneous Poisson process, we can attempt to model the intensity function directly. In this case, we will assume that the logarithm of the intensity (approximately) follows a bivariate spline. In Figure 6, we display an example of a (subdivided) triangular mesh used to define a continuously differentiable space of piecewise quadratic polynomials, see [Chui, 1988]. We have entertained both penalized likelihood techniques (see [O'Sullivan et al., 1986]) as well as ordinary maximum (penalized) likelihood over adaptively constructed triangulations, see [Stone et al., 1995]. We are currently exploring the use of principal components to characterize the variation in these intensity estimates.

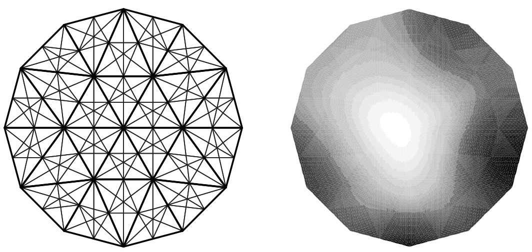

Figure 6: Bivariate splines and smoothing particle maps.

Data visualization is just the first step in understanding the nature of the particle deposition process. Our immediate goals for this investigation include:

1. The development of a system that routinely provides engineers with a snapshot of the particle deposition process level by level;

2. The development of models for understanding the spatial character of particle deposition and for assessing the impact of particles on wafer yield level by level; and

3. Combining particle information with processing information to determine the root causes of yield loss.

Since the Interface conference, we have made considerable progress toward achieving items 1 and 2 on this list, and we are in the process of transferring the resulting tools to AT&T's IC manufacturing sites in Madrid and Orlando. For example, we are now able to monitor particle maps for the presence of excessive clustering level-by-level and send a warning signal when the predicted impact on final probe yield is beyond a tolerable limit.

4 Analysis of Data from Designed Experiments

Experiments are frequently being conducted within AT&T Microelectronics. Unfortunately, the analysis of these data depends on an initial reduction of each wafer map to a one-dimensional summary like yield. A simple alternative to this approach involves modeling both yield and the join-count statistics discussed in Section 2, see [Taam and Hamada, 1992]. Another alternative involves modeling the various effects sitewise, and displaying the results as wafer maps. This approach was developed by Daryl Pregibon (Statistics Research, AT&T Bell Laboratories).

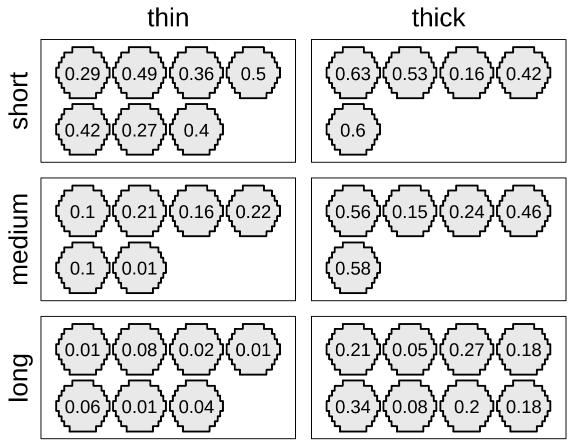

Figure 7: The results of a designed experiment.

Suppose, for example, that we have a (roughly balanced) experiment with two factors: the thickness of a given coating (thick and thin) and the duration of a particular process (short, medium and long). A plot of the design and the observed wafer yields is given in Figure 7. Coding each factor using linear and quadratic polynomials, a standard regression analysis on the yields of the wafers in this experiment produces the following table:

Coefficient	Value	Std.Err.	T-val
(Intercept)	0.269	0.020	13.76
coating	0.116	0.028	4.23
duration.L	-0.225	0.032	-6.87
duration.Q	0.003	0.034	0.09
coating duration.L	0.040	0.046	0.86
coating duration.Q	-0.087	0.050	-1.76

Observe that the longer processing times and a thinner coating tend to lower the overall wafer yield.

If we repeat this analysis sitewise (which in this case means 248 separate regressions, one for each chip on the wafer), then we can construct the following somewhat more informative table:

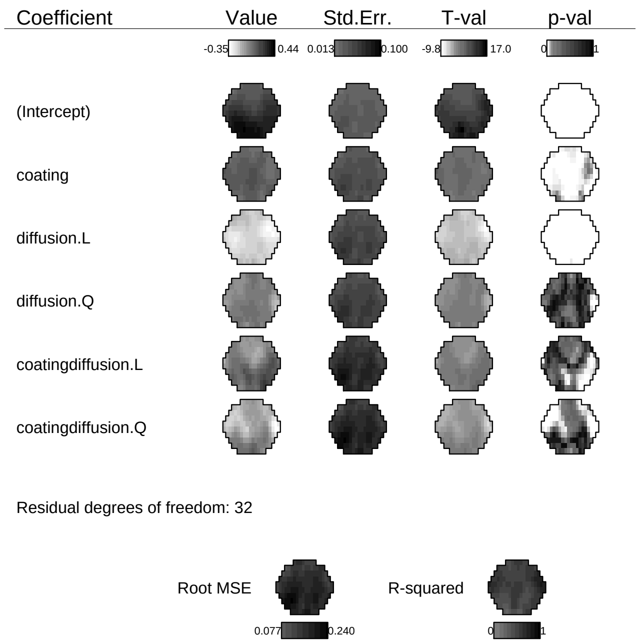

Figure 8: The results of a designed experiment.

While it is difficult to discern from this black and white plot, we again find that the longer processing times and thinner coating act to reduce wafer yield. In addition, however, we find that the thinner coating has a radial effect on yield. That is, changing the coating from thick to thin results in a greater yield loss on the edge of the wafer than in the middle. Insights such as these are extremely important for engineers contemplating changes in the manufacturing process. Keeping with the spirit of the table in Figure 8, we have also designed spatial analogs of design and interaction plots, each aimed at understanding the spatial effect of the factors.

If there is good reason to expect that certain processes will produce a given spatial signature, then these effects should be modeled directly. The use of appropriately defined bivariate spline spaces is an efficient way to incorporate this type of prior knowledge and is the subject of our current investigations. At present, the simple analysis described in this section has been implemented in the graphical user interface discussed in [Denby and James, 1995].

5 Acknowledgements

This work began roughly five years ago when David Friedman (Engineering Research Center, AT&T Bell Laboratories) introduced us to IC manufacturing and the traditional approaches to process and productivity improvement in the field. Since that point, the project has benefited from the involvement of Daryl Pregibon, David James and Lorraine Denby (Statistics Research, AT&T Bell Laboratories) and James Hoyer and Steve Whitlock (AT&T Microelectronics).

References

[Chui, 1988] Chui, C. (1988). *Multivariate Splines.* SIAM, Philadelphia, PA.

[Cressie, 1991] Cressie, N. (1991). *Statistics for spatial data.* Wiley-Interscience.

[Denby and James, 1995] Denby, L. and James, D. A. (1995). A graphical user interface for spatial data analysis in integrated circuit manufacturing. In Meyer, M. and Rosenberg, J., editors, *Computing Science and Statistics*, volume 27. Interface Foundation of North America.

[Friedman et al., 1995] Friedman, D., Hansen, M., Nair, V., and James, D. (1995). Model-free estimation of d_0 and y_0 from wafer map data. *submitted to IEEE Trans. on Semiconductor Manufacturing.*

[Hansen et al., 1995] Hansen, M., Nair, V., and Friedman, D. (1995). Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects. *in preparation.*

[Moran, 1948] Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, 10:243–251.

[O'Sullivan et al., 1986] O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81:96–103.

[Stone et al., 1995] Stone, C., Hansen, M., Kooperber, C., and Truong, Y. (1995). The use of polynomial splines and their tensor products in extended linear modeling. *submitted to The Annals of Statistics.*

[Taam and Hamada, 1992] Taam, W. and Hamada, M. (1992). Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics*, 35:149–160.