Proceedings *Process Control, Diagnostics, and Modeling in Semiconductor Manufacturing II*, Electrochemical Society Meeting, Montreal, May, 1997.

SPATIAL VARIATION IN SEMICONDUCTOR PROCESSES: MODELING FOR CONTROL

Duane Boning, James Chung, Dennis Ouma, and Rajesh Divecha
Massachusetts Institute of Technology, EECS
Microsystems Technology Laboratories, Cambridge MA 02139
Room 39-567, Phone: (617) 253-0931, Email: boning@mtl.mit.edu

ABSTRACT

Variation in critical device or structural parameters occurs at multiple scales in semiconductor processes; in most control scenarios, only the largest of these scales is typically addressed. New “statistical metrology” methods are emerging for the modeling of parameter variation not only across the wafer, but also within each die. Systematic dependencies to be captured include layout pattern or topography effects in addition to process relationships. Appropriate sampling and modeling of wafer-level variation in the face of within-die variation is important for process control applications. Modeling of within-die dependencies, particularly those related to pattern density effects, can benefit from signal processing analogies to identify appropriate parameters for both spatial and process modeling. In several semiconductor processes, models of both wafer-level and die-level spatial dependencies will become increasingly important for effective multi-objective process control that encompasses uniformity, throughput, and environmental goals.

1. INTRODUCTION

Practical applications of advanced process control (APC) to semiconductor manufacturing are beginning to emerge. In its most common form, APC is employed to maintain uniform process behavior and product characteristics over extended periods of time in the face of equipment, consumable, or material drifts and disturbances. In more aggressive cases, the goal is to maintain not only some target for the mean of one or more parameters from one wafer or lot to the next, but also to achieve a desired uniformity metric, usually a wafer-scale measure. Less attention has been paid to modeling and control when within-die (or intradie) variation is also important. Such variation is increasingly important in advanced circuits where yield and performance depend on consistent parameters within the die.

Statistical metrology is the body of methods for understanding *variation* (particularly *spatial variation*) in microfabricated structures, devices, and circuits. The goals of this paper are (1) to raise awareness in this area of research, (2) to examine some of the tools and methods that are emerging for modeling spatial variation, and (3) to consider the

implications and applications of spatial variation modeling for advanced process control. In Section 2, key ideas and methods emerging from statistical metrology are briefly reviewed. Many of these methods have been developed for the purposes of identifying variation sources and for understanding the possible impact of such variation on circuit performance. In this paper, the implications for process control are the focus. In particular, this paper considers issues in wafer-level modeling when die-level variation is also present, and die-level modeling to capture pattern (layout) process dependencies in a fashion appropriate for use in process control. The modeling of wafer-level variation is considered in more detail in Section 3, where we see that special care must be taken in modeling wafer-level uniformity when variation is also occurring at the die-level. The modeling of die-level variation is the focus of Section 4. While previous work in statistical metrology has sought to understand and characterize layout pattern dependencies, for process control we must also develop methods to characterize the *process* dependencies of within-die variation. Finally, in Section 5 we offer conclusions and directions for future research.

2. STATISTICAL METROLOGY

Statistical metrology is an emerging methodology for variation modeling and characterization in semiconductor manufacturing [1]. Historically, process control has paid a great deal of attention to temporal issues (e.g., variation in some thin film structure or device parameter from run to run or lot to lot), and to "large scale" spatial issues (e.g., uniformity in a parameter of concern across the wafer or across a batch of wafers in multi-wafer processing tools). Increasingly, however, variation at the die or circuit level is becoming a key concern, for not only must device or interconnect parameters be near target, they must also be well "matched" across the die (e.g. to avoid clock or signal skew, uncertain delays, etc.). When wafer and die variation is poorly understood, it is conventionally treated as a large "random" variation, and worst-case design approaches employed. Thus the first goal of statistical modeling is to identify the *systematic* components of variation, and to develop models and relationships for wafer- and die-level determinants of deviation from target.

A key tool emerging from statistical metrology is *variation decomposition* — the separation of variation elements into components most closely related with different spatial scales and corresponding different physical causes. In such an approach, an additive model is assumed in which wafer-level, die-level, and wafer-die interaction components are summed to express the systematic components of variation in measurements across many die on a wafer. In the additive linear model [2] we have

$$f_{\mathrm{RAW}}(x, y) = f_{\mathrm{WLV}}(x, y) + f_{\mathrm{DLV}}(x, y) + f_{\mathrm{INTERACTION}}(x, y) + \varepsilon \qquad (1)$$

where $f_{\mathrm{RAW}}(x, y)$ is the total response, $f_{\mathrm{WLV}}(x, y)$ is the wafer-level variation, $f_{\mathrm{DLV}}(x, y)$ is the "die pattern" identical for every die, and $f_{\mathrm{INTERACTION}}(x, y)$ captures the interaction between the wafer-level and die-level variation. Wafer-level variation is

generally considered to be relatively smooth or slowly varying across the wafer, and can be modeled using simply spatial regression, moving average (or down-sampled moving average) approaches, or spline methods [2]. Die-level variation is that component considered to be identical for every die on the wafer — that is, the component directly related to the repeated layout pattern present on every die. The spatial die "signature" can be extracted by taking advantage of the known periodicity in the stepping of the die across the wafer, and Fourier transform approaches used to isolate those components clustered around this fundamental frequency and its harmonics. An alternative method for die-level (or pattern dependent) modeling is a modified ANOVA (analysis of variance) approach. In Section 4, we will consider one important special case of layout pattern-dependent modeling, and examine methods for modeling *pattern density* effects. The final component of variation often considered is the wafer-die interaction. This corresponds to any perturbation in the die pattern as a function of position on the wafer. For example, if a substantial within-die variation is present, one often finds that the edge of the wafer will attenuate or accentuate the die pattern. This covariance between the wafer and die is of clear concern for process control; process modifications in an attempt to improve or control the wafer level variation may also have an unintentional effect on the die-level pattern dependencies.

Two implications or extensions of statistical metrology are addressed here. First, modeling of wafer scale uniformity must be undertaken with some caution when within-die variation is also present. In particular, sampling of a single (identical) structure from each die across the wafer can give a misleading impression of wafer level variation. Second, in a number of cases, long-range pattern density dependencies (acting across several mm) are important. In such cases, we find that signal processing concepts, including the notion of a density "impulse" or "step" response, can be effective for the characterization and modeling of such dependencies. Future control strategies will need to account for not only wafer scale uniformity, but the effect of control decisions on within-die uniformity as well.

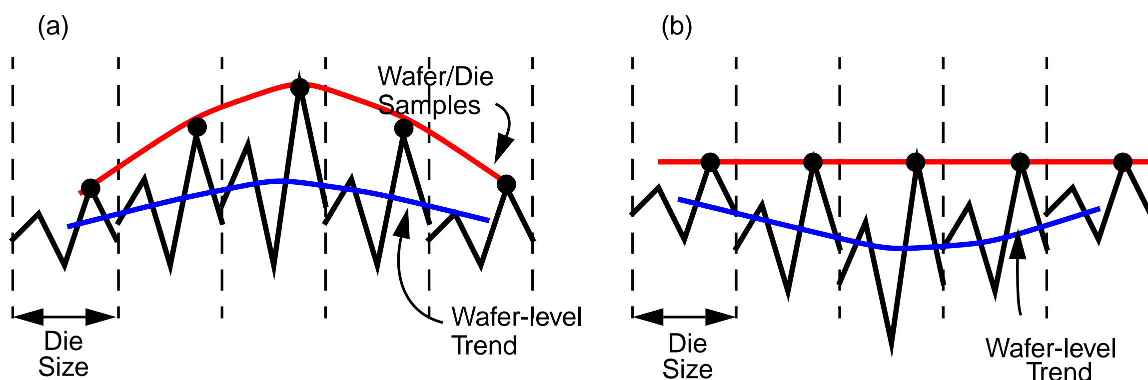
3. WAFER-LEVEL VARIATION MODELING

Several processes show layout pattern sensitivities; the calculation of wafer level uniformity in such cases must be performed with great care. Well-known pattern sensitivities in photolithography include linewidth variation as a function of nearby line distances; in plasma etch proximity and microloading effects are a concern; in CMP, planarization depends substantially on layout induced surface topography. In all of these cases, a systematic die-dependency is present: different structures within the die will display similar responses from one die to the next, even if those structures have radically different responses within the die.

In a typical process control scenario, a relatively small number of samples are often taken across the wafer. Film thickness measurements, for example, may be consistently gathered from a large capacitor or pad structure on the die for some number of die on the wafer. In this sampling, however, the observed nonuniformity is a combination of the true

underlying wafer level trend and any wafer-die interaction. This is illustrated in Figure 1(a), where we see a hypothetical one-dimensional cross-section through the wafer. The dies near the edge of the wafer are more uniform, and those near the center exhibit larger variation. While the die mean (or wafer-level trend) across the wafer shows a small curvature, the enclosing "envelope" of wafer and die variation is larger. A sampling of only one structure on each die may erroneously assign both die and wafer variation to the wafer scale uniformity, as illustrated in Figure 1(a). A control strategy which seeks to make these sampled values more uniform might be ineffective (if the control action only shifts the die means rather than reduce the die range), as illustrated in Figure 1(b).

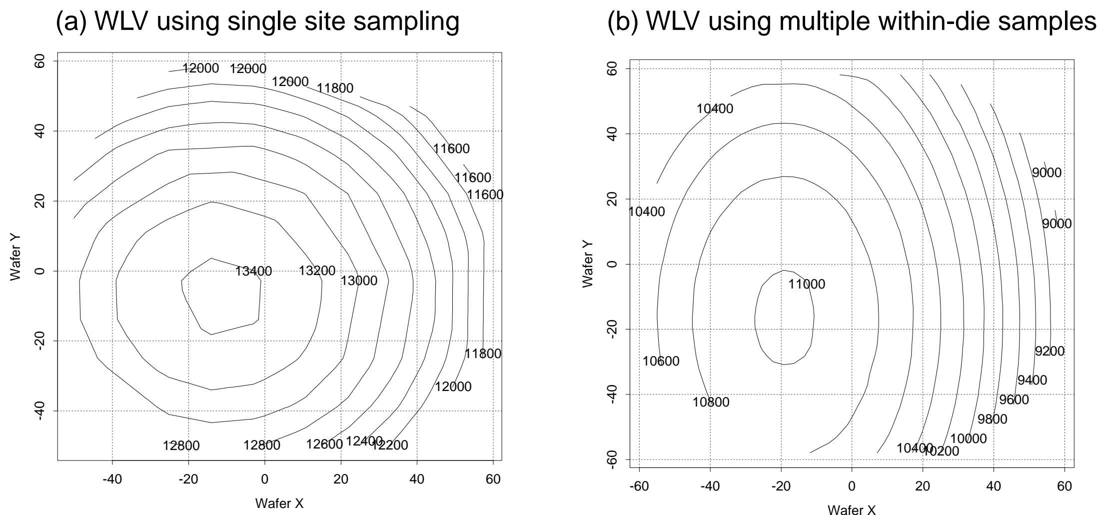
Figure 1: (a) Wafer-level trend generated by single-point sampling within each die can be very different than the mean surface. (b) Control based on sampled surface may achieve erroneous "uniformity."



Two methods can be used to overcome what is essentially a sampling and aliasing problem. First, intensive sampling of the structures within measured die, in addition to sampling of several die across the wafer, can be performed. For example, in Figure 2 we see that the shape of a regression surface fit to all available data (i.e. multiple within-die samples) is substantially different from the surface fit using single site per die sampling in a typical CMP process. This intensive sampling, of course, has a substantial cost in gathering the measurements. A simple but effective strategy is to measure both a "high" and a "low" spot within each die that is sampled across the wafer. In CMP, for example, one might measure both a "dense" and a "sparse" region on the die. The within-die range then becomes a simple estimate of the die variance, and can be used to check for behavior such as in Figure 1.

A deeper question also emerges when we fit regression surfaces for wafer-level variation in the face of within-die variation, and in particular when wafer-die interactions may be present. In the typical regression surface fit, an inherent assumption of the ordinary least squares (OLS) regression is that the variance is uniform across the space of the fit. As illustrated schematically in Figure 1, this may not be the case when substantial within-die variation is present, thus casting doubt on the use of OLS.

Figure 2: Wafer-level variation surfaces for a CMP process, generated from (a) single site sampling (88 die); (b) multiple within-die samples (15 points within each die).



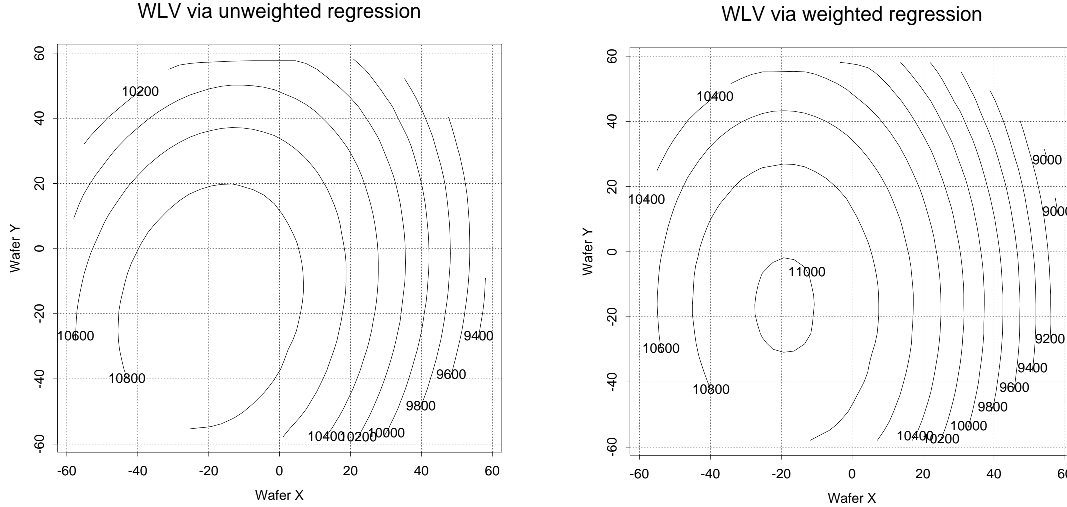
One approach is to stabilize the variance using a weighted least squares fit. For example, one may fit a regression surface to all available data, but weight each data point by the inverse of the variance of the die from which that data comes. As shown in Figure 3, the surface produced by weighted regression can be substantially different than that produced by unweighted regression. Indeed, in Figure 3 we find that the weighted regression is more successful in capturing a larger range of wafer level variation (9000 to over 1100) compared to the unweighted case (9400 to under 11000). In the additive linear model of Equation (1) however, it is not entirely clear if weighted or unweighted regression for $f_{\mathrm{WLV}}(x, y)$ is most appropriate, and how one should account for the components of the variance in the face of weighted regression. This is closely related to the suggestion of Davis et al. [3], who propose that both the mean and covariance be estimated simultaneously using a maximum likelihood approach. A variant of this approach which estimates wafer-level, die-level, and interaction terms simultaneously should be pursued in future research.

4. PATTERN DENSITY MODELS

Some processes appear to exhibit a long-range dependence on the surface topography (or presence of layout-induced structures) on the wafer. For example, CMP is known to exhibit a strong dependence on the amount of “raised” material (or “pattern density”) to be planarized in some local region (where “local” may span several mm). Because the raised features are closely related to, or caused by, the specifics of the layout, the modeling and control of such pattern density effects can be particularly troublesome. From the point of view of control, this introduces a product dependency — the layout is different for each product, and the effect of the process may be slightly different for each layout. The model-

ing of pattern-dependent effects will be increasingly necessary to overcome such non-idealities in process technology.

Figure 3: CMP regression wafer surfaces constructed using unweighted and weighted least squares regression.



The modeling of pattern density effects is interesting in that the length scale and model structure (e.g. spatial autocorrelation function [4]) of the dependence is not known a priori and must be established. In some cases, an output z of interest can be modeled as a function g of an "effective density" $d(x,y)$ that is positionally dependent:

$$f_{\mathrm{DLV}}(x, y) = z(x, y) = g(d(x, y)), \tag{2}$$

and in many cases, a relatively simple functional relationship between the effective density and the output variable exists [5], e.g.

$$z(x, y) = \alpha + \beta d(x, y). \tag{3}$$

The key problem then reduces to appropriate modeling of the effective density $d(x,y)$.

One approach is to consider a square "window" over which the density is calculated, as shown in Figure 4, and determine the window dimensions based on the best fit to available experimental or manufacturing data [6]. The density can be thought of as the convolution of the window $w(x,y)$ and the layout pattern $l(x,y)$

$$d(x, y) = w(x, y) \otimes l(x, y). \tag{4}$$

In the case of a square window, the width of the window provides a very compact characterization of the length scale of the response. For example, in CMP processes one can think of this parameter as the "planarization length" over which the polishing pad has effect. As shown in Figure 5, even this simple square window approach can be very effec-

tive in relating observed polishing dependencies with computed or extracted densities [7].

Figure 4: Local density determination using a moving window of size w_L.

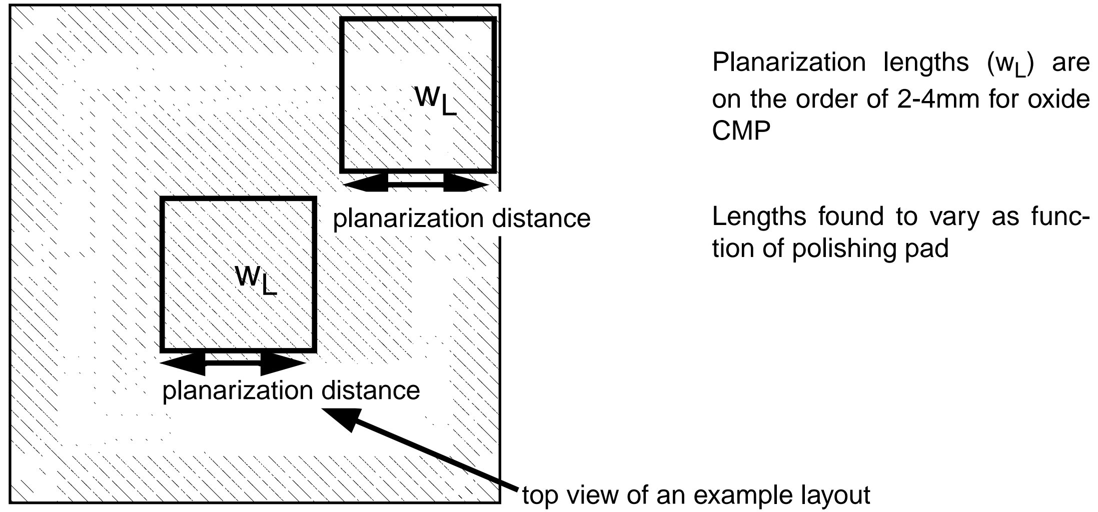
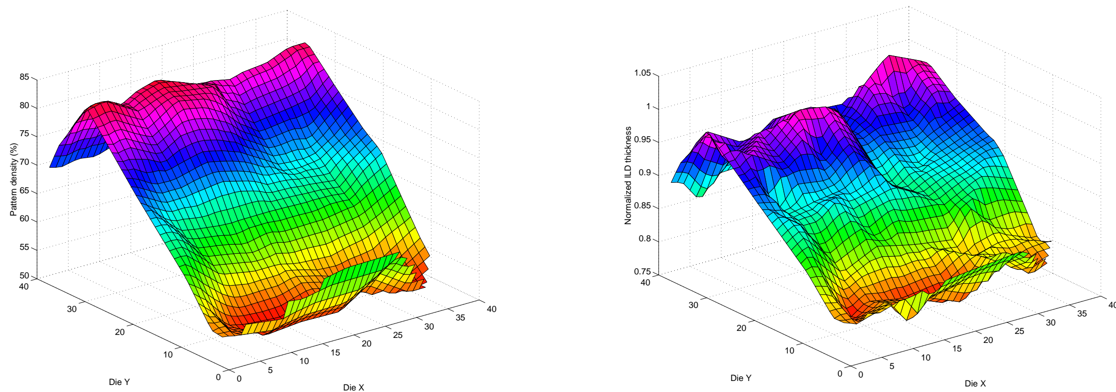


Figure 5: CMP die pattern modeling: the final polished oxide thickness shows a strong dependence on the underlying layout density.



(a) Die-level density computed using a 4mm moving average window over layout.

(b) Extracted die pattern (final oxide thickness) from experimental data.

On the other hand, the simple square window approach is unsatisfying in two respects. First, the square window, while convenient, is not directly connected to the expected physical behavior of the system. Indeed, treating the pad as an elastic material, one expects a much more narrowly shaped "impulse" response (window shape $w(x,y)$) than a square window. Second, the determination of the window size by the best "goodness of fit" to different computed window sizes is somewhat indirect and inexact (a number of different window sizes are often found with very nearly similar regression R^2 goodness of fit values); it would be helpful if a more direct approach could be employed.

An alternative based on a signal processing approach is to determine the window characteristics using the data in the frequency transform space:

$$W(k_x, k_y) = \frac{D(k_x, k_y)}{L(k_x, k_y)}$$
$$w(x, y) = F^{-1}(W(k_x, k_y)) \tag{5}$$

where F^{-1} is the inverse Fourier transform, and D and L are the Fourier transforms of d and l, respectively. In principal, it should be possible to directly extract the window shape and size from experimental layouts and resulting measurements (assuming the function g is also linear in the density). In practice, however, the layout frequency transform often has zeroes in it which make the extraction difficult.

The density response function or moving average window is likely to be non-rectangular, i.e. to have a "tapered" magnitude as a function of distance from center. Direct experimental determination of the density "impulse" response $w(x,y)$ is difficult. A closely related alternative, again borrowing from a signal processing approach, is direct experimental determination of the density "step" response. Our goal is the extraction of the response $w(x)$ of the polishing process to an infinitesimal impulse in density which we can use with new layouts (having different feature level layouts) to compute the expected final polished oxide thickness or planarity over the entire die. In one dimension, the resulting effective density can be computed as

$$\int_{-\infty}^{\infty} l(\xi) w(x-\xi) d\xi = d(x) \tag{6}$$

from which the resulting oxide thickness can be computed using Equation (3). The effective density impulse response $w(x)$ is then easily retrieved by differentiation of the spatial step response (the observed thickness $z(x)$) with respect to the spatial variable:

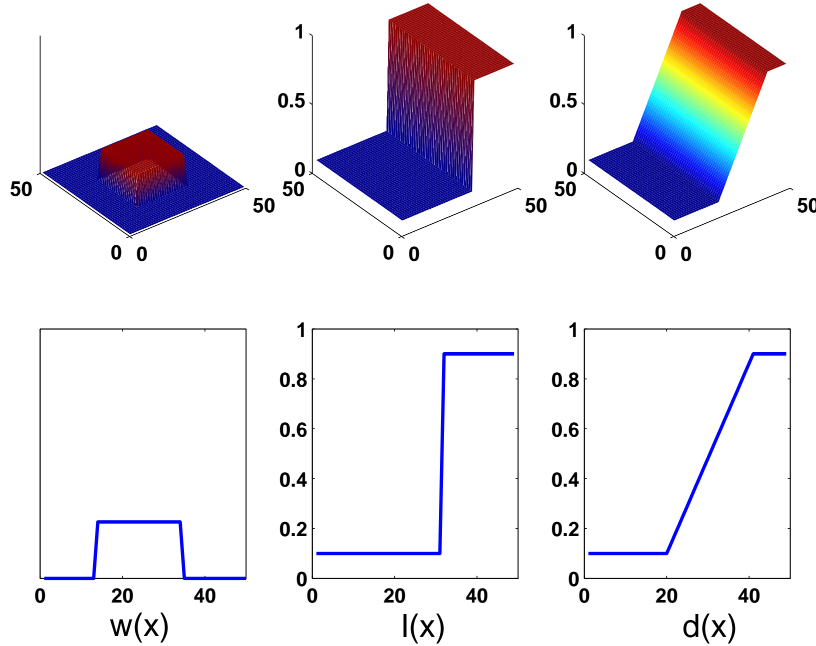
$$\frac{d}{dx} z(x) = \frac{d}{dx}[\alpha + \beta d(x)] = \beta \frac{d}{dx} d(x) . \tag{7}$$

Differentiating the step $l(x) = u(x)$ inside the integral of Equation (6):

$$\frac{d}{dx} z(x) = \beta \frac{d}{dx} d(x) = \beta \int_{-\infty}^{\infty} \delta(\xi) w(x-\xi) d\xi = \beta w(x) \tag{8}$$

which recovers our impulse response $w(x)$ as desired.

Figure 6: Quasi-two dimensional response *d(x)* for a rectangular window *w(x)* or “pad response” function applied to a quasi-1D step *l(x)*.



In two-dimensions (corresponding to the surface of the wafer), we must instead consider a quasi-1D step in density (a step in the x direction and infinite in the y direction) as illustrated in Figure 6. For the case of a rectangular window, the measurement $z(x, 0)$ of the result across the polished step is identical to the result we would see if we were to consider a simple one-dimensional system and step response $z(x)$. In general, this simple correspondence does not hold and it is not possible to uniquely recover the shape and extent of an arbitrary symmetric two-dimensional impulse response from the quasi-1D step response. In two dimensions, the response $z(x,y)$ is the convolution, as described in Equation (4) of the window function and an arbitrary layout function. Writing the convolution in integral form, the effective density is

$$d(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} l(\xi, \varphi)w(x-\xi, y-\varphi)d\xi d\varphi\,. \tag{9}$$

For the quasi-1D step response, our spatial layout function is a step in the x direction:

$$l(x, y) = u(x)\,. \tag{10}$$

We can differentiate the measured trace

$$\frac{d}{dx}z(x, 0) = \frac{d}{dx}[\alpha + \beta d(x, 0)] = \beta\frac{d}{dx}d(x, 0) \tag{11}$$

to retrieve the impulse response: Differentiating Equation (9) for the case of the step in

Equation (10) now yields the quasi-1D impulse response. Once again we perform the differentiation of the layout function inside the integral to generate the quasi-1D impulse response:

$$\begin{aligned}\beta \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(\xi) w(x-\xi, y-\varphi) d\xi d\varphi &= \beta \int_{-\infty}^{\infty} w(x, y-\varphi) d\varphi \\ \frac{d}{dx} z(x, 0) &= \beta \int_{-\infty}^{\infty} w(x, y) dy\end{aligned} \quad . \tag{12}$$

We see that the quasi-1D impulse response in Equation (12) is different than the simple one-dimensional impulse response of Equation (8). The functional form of the 1D and quasi-1D impulse responses will only be comparable (to within a scale factor) if the impulse response *w(x,y)* is separable in *x* and *y* so that the *y* dependency integrates out to a constant independent of *x*.

If we make the additional approximation that the shape of the impulse response is gaussian, the separability condition is satisfied and we can extract a simple shape parameter for the impulse response. This shape parameter can then be used to characterize the process pattern dependency. That is to say, a gaussian impulse response

$$w(x, y) = e^{-\frac{(x^2 + y^2)}{2\sigma^2}} \tag{13}$$

has a simple parameter σ which characterizes the width of the response. This case is illustrated in Figure 7, in which we find that the differentiated quasi-1D impulse response is simply

$$\frac{d}{dx} z(x, 0) = \beta C \int_{-\infty}^{\infty} \delta(\xi) w(x-\xi) d\xi = \beta C e^{-\frac{x^2}{2\sigma^2}} . \tag{14}$$

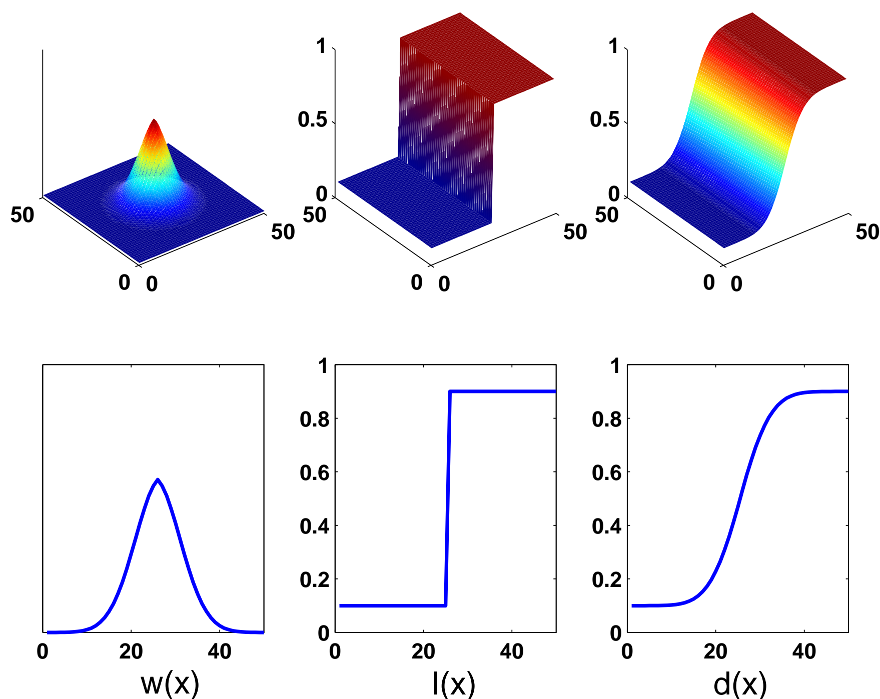
The quasi-1D step response thus offers a simple experimental approach to determine appropriate window functions for use in modeling of pattern density dependent die-level variation. One can envision an experimental design in which the σ parameter is extracted for different combinations of process conditions, and simple regression or other models of this planarization length as a function of the process can be generated.

5. CONCLUSION AND FUTURE WORK

We now have the basic elements needed to attack the problem of both wafer-level and die-level spatial dependencies for process control. Wafer-level extraction and spatial modeling can be achieved, given that sufficient care is taken in sampling and modeling. The first stage is to develop the spatial model; the second stage required for process control is

to also capture the process dependencies of that spatial variation. Approaches exist for such process/spatial modeling. For example, Mozumder and Lowenstein [8] suggest modeling spatial regression coefficients as a function of the process parameters. Alternatively, the "multiple response surface" (or "site models") approach of Guo and Sachs [9] considers the modeling of individual sites or locations on the wafer as a function of the process, and then the combination of these site models for flexibly form spatial models. Some additional research into the use of site modeling in the face of within-die variation is needed (e.g. hierarchical modeling of "die" and "wafer" sites).

Figure 7: The response of a two-dimensional gaussian "pad impulse" function to a quasi-1D step *l(x)*.



In this paper, we have also considered steps that need to be taken to make die-level spatial modeling amenable to the needs of process control. Here again, we need to be able to extract the die-level "signature" or pattern dependencies. The first step — creating the spatial pattern or model — must again be augmented by the ability to model the pattern as a function of process parameters (which will be the control variables in any feedback control scheme). Here, we have examined compact parametric descriptions of the pattern density dependency that may be appropriate for some kinds of fabrication processes, such as CMP. A "planarization length" parameter can be defined, and we suggest that experimental "step response" approaches can be used to extract this parameter. Once reduced to a parametric dependency, that parameter can itself be modeled as a function of process conditions.

Further work is needed to verify the notion of the pattern density step response, and to consider the process dependency of such a parameter. An exciting new challenge for process control will then be development of strategies that utilize both wafer-level and die-level spatial and process dependency models to achieve the difficult goals of both die- and wafer-level uniformity. Process control which achieves multiple objectives including wafer and die uniformity, as well as throughput and environmental goals, is an important area for further research.

ACKNOWLEDGMENTS

The contribution of many students in the statistical metrology and process control groups at MIT is gratefully acknowledged. Examples are drawn from several collaborations, including those with HP, LSI, Sandia National Laboratories, and TI. This work has been supported in part by DARPA under contract #DABT-63-95-C-0088 and AASERT grant #DAAHA04-95-I-0459, and by the NSF/SRC Engineering Research Center for Environmentally Benign Semiconductor Manufacturing.

REFERENCES

1. D. Boning and J. Chung, "Statistical Metrology - Tools for Understanding Spatial Variation," *SPIE 1996 Symp. on Microelectron. Manuf.*, Austin TX, Oct. 1996.
2. B. Stine, D. Boning, and J. Chung, "Analysis and Decomposition of Spatial Variation in Integrated Circuit Processes and Devices," *IEEE Trans. Semi. Manuf.,* pp. 24-41, Feb. 1997.
3. J. C. Davis et al., "Improved Within-Wafer Uniformity Modeling Through the Use of Maximum-Likelihood Estimation of the Mean and Covariance Surfaces," *ECS Proc. Vol. 95-4*, p. 412, 1995.
4. B. D. Ripley, Spatial Statistics, Wiley, New York, 1981.
5. B. Stine et al., "A Closed-Form Analytic Model for ILD Thickness Variation in Oxide CMP Processes," *Proc. CMP-MIC*, Feb. 1997.
6. B. Stine, et al., "Rapid Characterization and Modeling of Pattern Dependent Variation in Chemical Mechanical Polishing," submitted to *IEEE Trans. Semi. Manuf.*, Nov. 1996.
7. R. Divecha et al., "Effect of Fine-Line Density and Pitch on Interconnect ILD Thickness Variation in Oxide CMP Processes," *Proc. CMP-MIC*, Feb. 1997.
8. P. Mozumder and L. Lowenstein, "Method for Semiconductor Process Optimization Using Functional Representations of Spatial Variations and Selectivity," *IEEE CPMT*, Vol. 15, No. 3, pp. 311-316, June 1992.
9. R.-S. Guo and E. Sachs, "Modeling, Optimization, and Control of Spatial Uniformity in Manufacturing Processes," *IEEE Trans. Semi. Manuf.*, pp. 41-57, Vol. 6, No. 1, Feb. 1993.