Title
Intrinsic and Systematic Variability in Nanometer CMOS Technologies

Author
Patel, Kedar

Publication Date
2010

Peer reviewed|Thesis/dissertation

Intrinsic and Systematic Variability in Nanometer CMOS Technologies

by

Kedar Kantilal Patel

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Costas Spanos, Chair
Professor Tsu-Jae King Liu
Professor Sandrine Dudoit

Fall 2010

Intrinsic and Systematic Variability in Nanometer CMOS Technologies

Abstract

Intrinsic and Systematic Variability in Nanometer CMOS Technologies

by

Kedar Kantilal Patel

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Costas Spanos, Chair

It has been widely recognized that variability is one the most important challenges to scaling of nanoscale CMOS devices. Intrinsic sources of variation such as discretization effect of dopant atoms, metal-gate work-function variation, and line width roughness threaten an end to scaling realized in the past decades. Line width roughness (LWR) is of great importance as it is a significant fraction of the minimum feature size for nanoscale devices, and it does not scale at the same pace as the minimum feature size. According to previous studies, a complete description of LWR can be provided by three parameters: root-mean square (RMS) roughness (σ), correlation length (ξ), and roughness exponent (α).

A robust method of estimating line width roughness parameters is presented. Specifically, the proposed method provides a *better* unbiased estimate of roughness amplitude σ than existing methods. It also provides an estimate of error in LWR parameters. The proposed method also allows for more flexibility in capturing SEM images in that we do not need a special test structure with all lines with same designed CD; any IC layout region with straight lines and arbitrary CDs would suffice. As an application of this method, LWR characteristics of many next-generation lithography processes are explored. LWR parameters are also incorporated in the FinFET device framework, and useful physical insights are provided in regards to its impact on device performance.

Variability can also be systematic in nature. Systematic spatial variation can occur at the wafer- or die-level. Accurate estimation of various variability components is necessary for robust circuit design. To this end, a hierarchical decomposition of semiconductor process variation is performed. A holistic discussion on all components of process variation is provided. Specifically, global (inter-die) variation is modeled in a multivariate normal framework. The same framework is extended to enable wafer-selection for model estimation. Least angle regression and agglomerative hierarchical clustering are proposed for selecting wafers for model estimation. Methodologies to model systematic local (intra-die) variation and spatial correlation are provided. Spatial correlation in intra-die observations is extracted using the variogram, and issues in variogram estimation are discussed in detail.

To my beloved wife, Neha
and
my twin bundles of joy, Arav and Anika

Contents

List of Figures v

List of Tables vii

1 Introduction 1
- 1.1 Semiconductor Technology . . . 1
- 1.2 Process Variability . . . 2
- 1.3 Role of Modeling . . . 3
- 1.4 Modeling Challenges . . . 7
- 1.5 An Overview of This Work . . . 8

2 Robust Estimation of LWR Parameters 10
- 2.1 Introduction . . . 10
  - 2.1.1 Motivation and Background . . . 10
  - 2.1.2 Our Work . . . 12
- 2.2 Line Width Roughness Model . . . 13
  - 2.2.1 LWR Parameters . . . 13
  - 2.2.2 A Variogram Model . . . 14
- 2.3 Estimation of Model Parameters . . . 16
  - 2.3.1 Estimation of ξ and α . . . 16
  - 2.3.2 Estimation of σ^2 . . . 19
- 2.4 Robust Estimation of LWR Parameters . . . 20
  - 2.4.1 Procedure . . . 21
  - 2.4.2 Optimal Block Length . . . 25
- 2.5 Results and Discussion . . . 26
  - 2.5.1 Simulated Data . . . 26
  - 2.5.2 Validation Using Simulated Data . . . 26
  - 2.5.3 Experimental Data . . . 30
- 2.6 Summary . . . 31

3 LWR in Next-Generation Lithography Processes **32**
3.1 Introduction . . . 32
3.2 Estimation of LWR Parameters . . . 33
3.2.1 SEM Image Acquisition and Processing . . . 33
3.2.2 Estimation Procedure . . . 34
3.3 Next-Generation Lithography (NGL) Processes . . . 34
3.3.1 Double Patterning Lithography (DPL) . . . 35
3.3.2 Self-Aligned Double Patterning Lithography (SADP) . . . 36
3.3.3 Extreme Ultraviolet Lithography (EUV) . . . 37
3.3.4 Directed Self-Assembly Lithography (DSA) . . . 37
3.3.5 Nano-imprint Lithography (NIL) . . . 38
3.4 Results and Discussion . . . 39
3.5 Summary . . . 45

4 Gate LER Model for FinFET Performance **46**
4.1 Introduction . . . 46
4.2 Line Edge Roughness . . . 47
4.2.1 Background . . . 47
4.2.2 Spacer vs. Resist Lithography . . . 48
4.3 Simulation Details and Model Formulation . . . 49
4.3.1 FinFET Structure . . . 49
4.3.2 Simulation Details . . . 52
4.3.3 Model Formulation . . . 53
4.4 Results and Discussion . . . 56
4.5 Summary . . . 61

5 Decomposition of Semiconductor Process Variation **64**
5.1 Introduction . . . 64
5.2 Semiconductor Process Variation . . . 65
5.3 Spatial Correlation . . . 68
5.4 Decomposition of Process Variation . . . 70
5.4.1 Nomenclature . . . 70
5.4.2 Modeling Choices . . . 70
5.5 Wafer-Level Variation . . . 72
5.5.1 Multivariate Normal Model . . . 72
5.5.2 Normality Test and Outlier Rejection . . . 72
5.6 Wafer Selection . . . 73
5.6.1 Least Angle Regression Using PCA Basis Functions . . . 73
5.6.2 Cluster Analysis . . . 76
5.7 Die-Level Variation . . . 77
5.8 Residual Analysis . . . 77

5.8.1 Variogram of Residuals . . . 78
5.8.2 Variogram Estimation . . . 79
5.9 Results and Discussion . . . 82
5.9.1 Wafer Selection . . . 83
5.9.2 Model Estimators . . . 87
5.10 Summary . . . 91

6 Conclusion **92**
6.1 Summary of Contributions . . . 92
6.2 Suggested Future Work . . . 93

A Bias in Finite Length Variance **95**

B Validity of LLE Approach **97**

Bibliography **98**

List of Figures

1.1 Moore's Law scaling trend . . . 2
1.2 SNM distribution of 6T SRAM . . . 3
1.3 Example of systematic wafer-level spatial variation . . . 4
1.4 Sources of intrinsic variation . . . 4
1.5 Levels of abstraction in modeling parameters . . . 5
1.6 Schematic view of planar MOSFET shown with 2D slice approximation . . . 6

2.1 Biased estimate $\hat{\sigma}_{LWR}$ as a function of length of the line . . . 11
2.2 A sample scanning electron micrograph (SEM) image . . . 12
2.3 Illustration of roughness descriptors . . . 13
2.4 Generalized framework of simulated and digitized SEM data . . . 17
2.5 Graphical illustration of *block of blocks bootstrap* . . . 22
2.6 Graphical illustration of *moving block bootstrap* . . . 23
2.7 Mean-squared error (MSE) in α and ξ as a function of the block length . . . 27
2.8 Optimal block length computed using Politis and White method . . . 27
2.9 Comparison of two methods of estimating WLS weights . . . 28
2.10 Comparison of four different estimates of σ . . . 29
2.11 Sample fit of variogram using WLS and BBB . . . 30
2.12 Comparison of $\hat{\sigma}_{BBB}$ and $\hat{\sigma}_{LLE}$ for various NGL processes. . . . 31

3.1 SEM images at intermediate SADP steps . . . 33
3.2 *Litho-freeze-litho-etch* (LFLE) process flow . . . 35
3.3 Self-aligned Double-Patterning process flow . . . 36
3.4 Directed Self-Assembly process flow . . . 37
3.5 Nano-imprint lithography process flow . . . 38
3.6 LWR (3σ) in nm for various NGL processes . . . 40
3.7 Normalized LWR for various NGL processes . . . 41
3.8 Correlation length (ξ) for various NGL processes . . . 42
3.9 Roughness exponent (α) for various NGL processes . . . 42
3.10 Power Spectral Density (PSD) for LFLE and EUV . . . 44
3.11 Power Spectral Density (PSD) for DSA and SADP . . . 44

3.12 Results of process optimization from 64nm pitch EUV 45

4.1 Comparison of LER and LWR 47
4.2 Illustration of spacer and resist lithography methods 49
4.3 Illustration of LER components in FinFET 50
4.4 Schematic views of a DG FinFET 51
4.5 Non-ideal 2-D DG-FET structure 52
4.6 Definition of model parameters 54
4.7 Fin width dependence of saturation threshold voltage 57
4.8 Fin width dependence of $I_{d,sat}$ and I_{off} 57
4.9 Threshold voltage variation over ΔL and δ space 58
4.10 Threshold voltage dependence on δ and ΔL 58
4.11 Monte-Carlo comparison of $V_{t,sat}$ for simulated and experimental grid 60
4.12 Variability in saturation threshold voltage $V_{t,sat}$ for resist-defined gate 61
4.13 Variability in drive and off-state current as function of LWR amplitude . . . 62
4.14 Variability in drive and off-state current as function of correlation length . . 62
4.15 Variability in saturation threshold voltage $V_{t,sat}$ for resist-defined gate 63

5.1 Overview of variability components and attributes 66
5.2 Illustration of semiconductor process variation hierarchy 67
5.3 Source of spatial correlation due to wafer-level variation 68
5.4 Factors influencing our ability to estimate spatial correlation 69
5.5 Graphical illustration of variogram parameters 81
5.6 Probability density plot of raw data 83
5.7 Mahalanobis distance before and after rejecting outlier wafers 84
5.8 Example of Least Angle Regression 85
5.9 Dendrogram of hierarchy discovered in our data set 86
5.10 Procedures to determine the number of clusters 87
5.11 Multivariate normal representation of ω_{ijk} 88
5.12 Systematic spatial variation average die 88
5.13 Probability plot of residuals ϵ 89
5.14 Estimated variogram in X and Y directions 90
5.15 Parametric variogram fitted using weighted-least squares 90

List of Tables

3.1 Summary of LWR parameters for various NGL technologies 43

4.1 2-D Device Simulation Parameters 52
4.2 2-D Nominal Device Performance Parameters 56

5.1 Subscript notations used in decomposition of variability 71
5.2 Estimation of model terms 71

Acknowledgments

I left Berkeley in 1998 with B.S and M.S degrees. Returning back to Berkeley in 2006 to pursue a Ph.D. degree was not easy. There are many people at Berkeley who deserve recognition and my gratitude for supporting my endeavor. Foremost among them is Prof. Costas Spanos. Costas challenged my understanding by posing intriguing questions. In spite of his remarkable knowledge, he has always been open to new ideas. On a personal level, Costas has been very supportive, patient, and friendly. No journey is devoid of its metaphorical ups and downs, and my endeavor was no exception. I'd like express my sincere gratitude to Costas for being a pillar of support, for his inspiration and encouragement during the rough patches of my journey.

I would also like to thank Prof. Tsu-Jae King Liu for chairing my qualifying exam committee and serving on dissertation committee. I am especially grateful to Tsu-Jae for providing the clarity in my research on FinFET. Her remarkable knowledge and responsiveness have been instrumental in the work presented in chapter 4. I would also like to thank Prof. Sandrine Dudoit for serving on my qualifying exam and dissertation committees and Prof. Elad Alon for serving on my qualifying exam committee. I would also like to thank Prof. Soumendra Lahiri at Texas A&M University for guiding the statistical aspects of my research. His masterful insights on bootstrap was instrumental in work presented in chapter 2.

The work presented in chapter 3 would not have been possible without the help of Dr. Thomas Wallow at Global Foundries, and I am indebted to him for his support. I would also like to thank Dr. Vassilios Constantoudis (IMEL) for invigorating discussions of line width roughness. Vassilios has evangelized the three parameter description of line width roughness in the scientific community, and in spite of his immense contributions on the subject of line width roughness, he remains modest and humble. I would like to thank Dr. Lee Smith at Synopsys for help with Sentaurus and Dr. Patrick Naulleau (LBNL) and Prof. Tyrone Vincent (Colorado School of Mines) for helpful discussions.

Finally, I would like to express my deepest gratitude to my family for their love, compassion and support in my endeavor. My wife, Neha, deserves to share my degree as had it not been for her unbounded love and constant support, my journey would not have even started. She always remained cheerful despite of enduring the extreme hardships due to my absence from home on weekends for four years. I am also grateful to my twin bundles of joy, Arav and Anika, for foregoing four years of their weekend play time with me so that I could pursue my calling. I would also like to deeply thank my parents for inspiring me through their achievements in higher education, which were accomplished against the backdrop of immense familial, societal, and cultural hardships.

Ithaca

(Constantine P. Cavafy, 1863 - 1933)

When you set out on your journey to Ithaca,
pray that the road is long,
full of adventure, full of knowledge.
The Lestrygonians and the Cyclops,
the angry Poseidon – do not fear them:
You will never find such as these on your path,
if your thoughts remain lofty,
if a fine emotion touches your spirit and your body.
The Lestrygonians and the Cyclops,
the fierce Poseidon you will never encounter,
if you do not carry them within your soul,
if your soul does not set them up before you.

Pray that the road is long.
That the summer mornings are many,
when, with such pleasure, with such joy
you will enter ports seen for the first time;
stop at Phoenician markets,
and purchase fine merchandise,
mother-of-pearl and coral, amber, and ebony,
and sensual perfumes of all kinds,
as many sensual perfumes as you can;
visit many Egyptian cities,
to learn and learn from scholars.

Always keep Ithaca on your mind.
To arrive there is your ultimate goal.
But do not hurry the voyage at all.
It is better to let it last for many years;
and to anchor at the island when you are old,
rich with all you have gained on the way,
not expecting that Ithaca will offer you riches.

Ithaca has given you the beautiful voyage.
Without her you would have never set out on the road.
She has nothing more to give you.
And if you find her poor, Ithaca has not deceived you.
Wise as you have become, with so much experience,
you must already have understood what these Ithacas mean.

Chapter 1

Introduction

1.1 Semiconductor Technology

Semiconductors are ubiquitous these days. There is hardly a part of our life that has not been improved by semiconductors. New applications are being invented to exploit the miniaturization of integrated circuits and the performance improvement at that scale. Broadly speaking, the production of the integrated circuits can be classified into two parts, namely, circuit design and circuit fabrication.

Traditional circuit design has three primary and competing objectives: maximize performance, minimize power consumption, maximize yield. A circuit's performance is typically evaluated in terms of speed or delay in the transmission of a signal. The delay in signal transmission can be reduced with higher transistor drive currents. Higher drive currents typically result in higher leakage currents which in turn increases the power consumption. Yield loss of a circuit circuit can be as a result of a defect during fabrication or due to the failure of circuit to meet its parametric objectives (higher drive currents, lower leakage current). Other secondary objectives of integrated circuit (IC) industry are cost, level of integration, functionality, and form factor [1].

Circuit fabrication is enabled by semiconductor technology or process development. Fabrication is realized through a series of processing steps, with each step having a specific objective such as creating a lithographic pattern, depositing or etching a film, etc.. The increased demand for complex circuits with increased performance at very low cost has driven semiconductor technology development to a record pace, and it has made achieving the aforementioned design objectives more difficult than ever before [1]. The minimum feature size on a die or chip has decreased exponentially over the years. The scaling trend of minimum feature size is usually referred to as Moore's Law [2]. It states that the number of components per chip doubles approximately every 24 months. Figure 1.1 shows the Moore's Law scaling for the past several decades [3]. Since its inception in 1992 by the Semiconductor Industry Association (SIA), the International Technology Roadmap for Semiconductors (ITRS) has

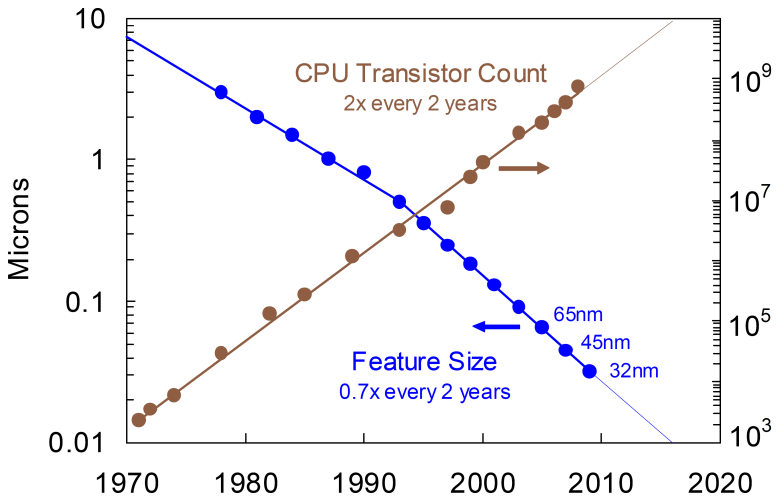


Figure 1.1: Moore's Law scaling trend. After [3]

published the technical capabilities that are required to remain on Moore's Law trend. According to the ITRS, the scaling of the industry workhorse—the planar CMOS MOSFET, will face significant challenges in the near-term [1]. New materials and new device architectures, in addition to process control of unprecedented proportion, will be required to break the scaling barriers. The complexity of process flow also has increased dramatically over the past decade. The increased number of processing steps (that are commensurate with the process complexity) have introduced new sources of variability such as strain or stress in thin films. Variability in physical and electrical characteristics has been further compounded by discretization of charge and matter at the nanoscale [4].

1.2 Process Variability

Variation is the deviation of realized values from its design intent. In semiconductor processing, variability arises from the multitude of processing steps it takes to fabricate a wafer (the unit of production), and it causes undesired variation in the performance of electrical circuits on a die or chip (unit of merit). In a nutshell, variation can result in yield loss or degraded circuit performance. Figure 1.2 shows the simulated distribution of 6T static random-access memory (SRAM) signal-to-noise margin (SNM) over an ensemble of 200 SRAM cells for 25-, 18- and 13-nm generations [4]. SNM is a key metric for the performance and reliability of SRAM. Notice the dramatic widening of distribution as the gate-length is scaled down.

Variation may be *systematic* or *random* in nature. Systematic spatial variation can occur at large (lot or wafer) or small (die) scale. Wafer-level systematic spatial variation is

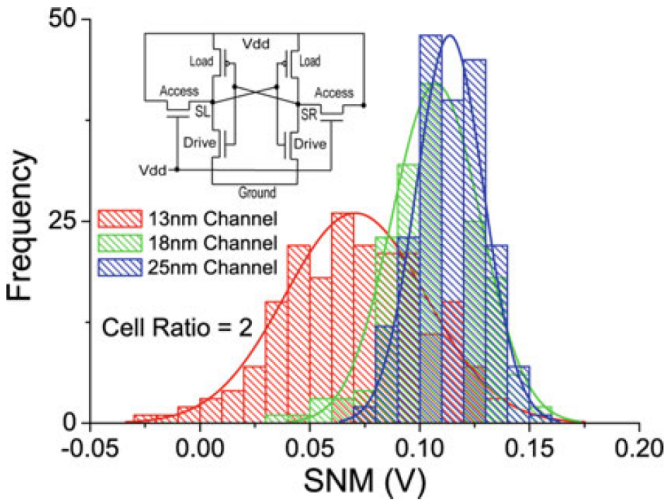


Figure 1.2: SNM distribution of 6T SRAM over an ensemble of 200 SRAM cells for 25-, 18- and 13-nm generations. After [4]

typically found to be slowly varying and smooth function across the wafer. Figure 1.3 shows an example of wafer-level systematic spatial variation. Systematic spatial variation can cause spatial dependencies or correlation between collection of structures on a die. Devices or structures that are in close proximity behave much more similarly than those that are spaced farther apart. *Random* variation represents the uncertain component of variation. As shown in Figure 1.4, random dopant, line edge (or width) roughness, metal gate granularity, high-k granularity, interface roughness, etc., are examples of random variation [4]. These sources of variation are *intrinsic* to semiconductor processing, and as such, the random variation is also commonly referred to a *intrinsic variation*.

1.3 Role of Modeling

Models help us gain insight into the physical aspects of a phenomenon. In the IC industry, modeling has been the foundation of circuit simulation. Figure 1.5 shows the levels of abstraction in model parameters. Model parameters at each level, at least in theory, can be expressed in terms of parameters at lower level of abstraction. For instance, expressing physical parameter in terms of process parameters is not trivial as typically each physical parameter is influenced by a number of process parameters spanning many steps.

Using LWR as an example, we can demonstrate the relationship of parameters across different levels of abstraction. LWR is the undesired variation in width of a line, and a complete description of LWR can be provided by three parameters: root-mean square (RMS)

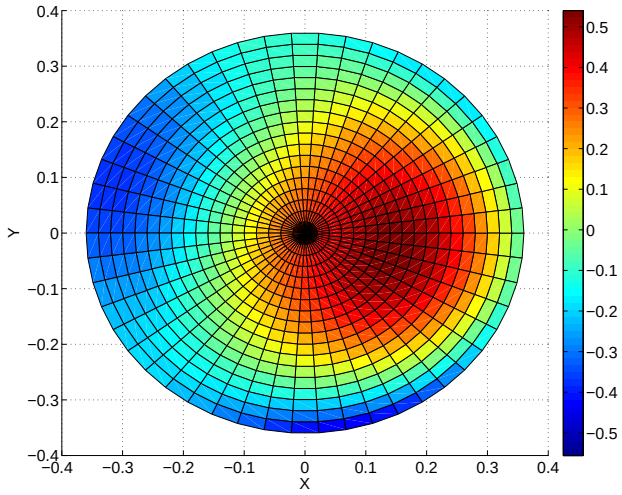


Figure 1.3: Example of systematic wafer-level spatial variation in ring oscillator frequency.

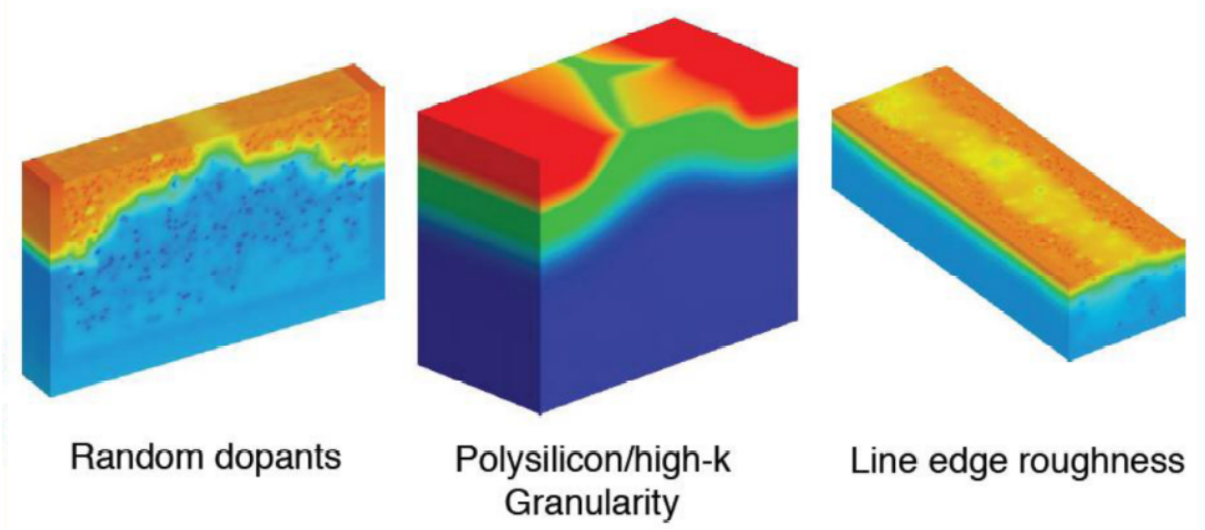


Figure 1.4: Sources of intrinsic variation. After [4].

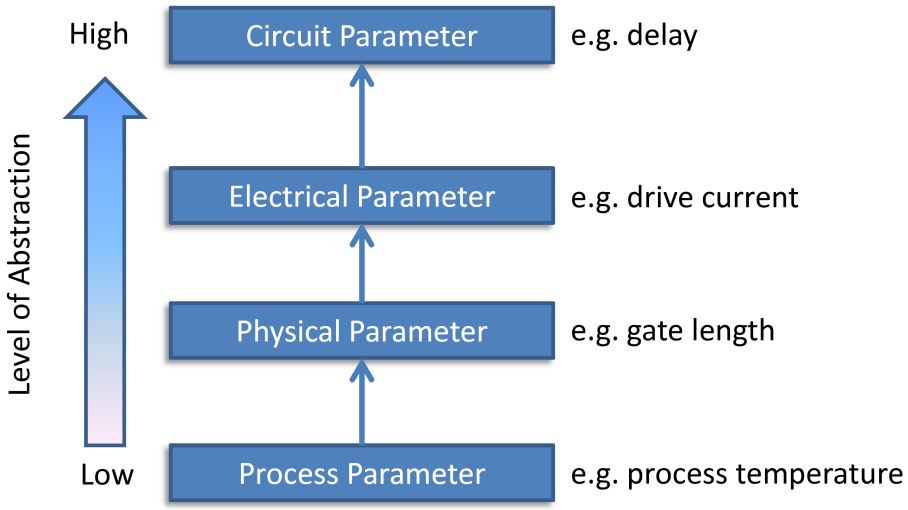


Figure 1.5: Levels of abstraction in modeling parameters

roughness (σ), correlation length (ξ), and roughness exponent (α) [5, 6]. For self-affine and isotropic surfaces, LWR can be characterized by a specific functional form of auto-covariance [7],

$$C(h;\sigma,\xi,\alpha) = \sigma^2 \exp\left(-\left[\frac{|h|}{\xi}\right]^{2\alpha}\right) \quad h \in \mathbb{Z}^d. \tag{1.3.1}$$

Now consider the line width roughness in gate of planar CMOS transistor. It causes variation in the gate length along the width. The roughness can typically be approximated by a slice approach (as shown in Figure 1.6). The gate is approximated by many narrow slices, where each slice is assumed to be ideal with no roughness. Let $\partial L(x)$ denote the fluctuation of gate length at position x along the width W, and let $\bar{L}$ denote the gate length averaged over W. After first-order Taylor expansion around $\bar{L}$, the drive current I of this transistor can be given as

$$I = \int_0^W J\left(\bar{L} + \partial L(x)\right) dx = J(\bar{L})W + \frac{\partial J(\bar{L})}{\partial \bar{L}} \int_0^W \partial L(x) dx + \text{H.O.T.} \tag{1.3.2}$$

In (1.3.2), $J(L)$ represents the current density for gate length L. The variation in I can be

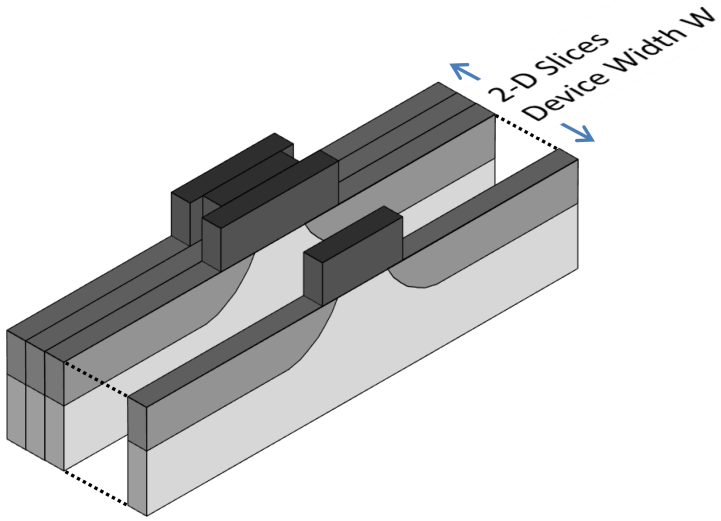


Figure 1.6: Schematic view of planar MOSFET shown with 2D slice approximation for gate LWR.

given as

$$\begin{aligned}\sigma_I^2 = E(I-\bar{I})^2 &= \left(\frac{\partial J(\bar{L})}{\partial \bar{L}}\right)^2 \int_0^W \int_0^W E\left[\partial L(x)\partial L(x')\right] dxdx' \\ &= \left(\frac{\partial J(\bar{L})}{\partial \bar{L}}\right)^2 \int_0^W \int_0^W C(x-x')dxdx' \\ &= \left(\frac{\partial J(\bar{L})}{\partial \bar{L}}\right)^2 \sigma^2(2\xi W),\end{aligned} \tag{1.3.3}$$

where we used $\bar{I} = J(\bar{L})W$ and the double integral was evaluated using (1.3.1) under the assumptions $\alpha = 0.5$ and $W \gg \xi$. Now, consider the intrinsic delay of a transistor $d = CV/I$, where C is the total gate capacitance, V is the supply voltage, and I is the drive current. We can express the variability in d as

$$\sigma_d^2 = \sigma_I^2 \left(\frac{\partial d}{\partial I}\right)^2 = \left(\frac{\partial J(\bar{L})}{\partial \bar{L}}\right)^2 \sigma^2(2\xi W) \left(\frac{CV}{I^2}\right)^2. \tag{1.3.4}$$

Statistical static timing analysis (SSTA) is an analysis of delay between various circuit nodes [8, 9, 10, 11, 12, 13]. The variability in intrinsic transistor delay (1.3.4) can subsequently be incorporated into SSTA along with delay due to interconnects. Doing so allows us to draw meaningful conclusions regarding the impact of each LWR parameter on the speed of the circuit. Our simple LWR example demonstrates how fundamental roughness parameters can

be linked to the variability in delay at the circuit level. This approach can be extended on a grander scale through the use of SPICE or BSIM models. Key physical parameters such as gate length, gate dielectric thickness, etc., are captured in the transistor models. These models are validated against silicon data, and they are subsequently used to simulate the circuit behavior.

Semiconductor IC fabrication involves multitude of processing steps, and each step contributes a degree of uncertainty in some physical parameter. Variability, which originated during processing, propagates to higher levels in Figure 1.5. An estimate of variability at higher levels of abstraction can be made through the use of SPICE models; variability in physical parameters such as transistor gate length can be transformed into variability of electrical characteristics such as threshold voltage or drive current.

1.4 Modeling Challenges

Models can be used for describing a phenomenon (such as LWR) or even to describe variability. For instance, knowledge of how the variance of a particular physical or electrical parameter decomposes across the semiconductor hierarchy (lot, wafer, die, and within-die) is essential in robust circuit design. Models of variability can be stipulated at any level of abstraction shown in Figure 1.5.

Irrespective of what is being modeled, there are fundamental challenges associated with the problem of modeling:

- *Existence of suitably compact representation and methods for characterization*

  Our earlier example of LWR was based on the premise that the LWR phenomenon can be completely described by only three parameters. However, not all phenomena can be represented compactly. Random dopant fluctuation is a good example of one such case. Dopants are impurities implanted in CMOS transistors to tailor electrical characteristics. The implantation process involves an energetic beam of either n- or p-type impurity atoms. With highly scaled transistors, an incorrect final resting position of even a singular impurity atom can alter the electrical performance of a transistor. Randomness in the location of these dopants does not have a compact representation. There is also no non-destructive method of accurately characterizing the post-implantation positions of these dopants. In cases such as random dopant fluctuations, the physical to electrical transformation can only be accomplished by a *brute force* Monte-Carlo simulation [14]. However, further propagation of variability up to the circuit-level becomes prohibitively expensive and intractable due to computational limitations.

- *Availability of data at a level of abstraction conducive to being used for circuit simulation*

Even when models and methods of their estimation do exist, the data is not always available at a level of abstraction that is conducive to be used in circuit simulation. For instance, if the modeled parameter is an electrical parameter like the threshold voltage of a transistor, then it can be incorporated much more easily in the circuit simulation than a material parameter such as film stress or annealing temperature. The latter choice would involve more elaborate simulation setups, and in many cases may require simplifying assumptions. Generally speaking, electrical parameters are more desirable choice for modeling purposes as they tend to incorporate all underlying physical and material phenomena.

- *Scarcity of rich enough spatial and temporal data to accurately estimate model parameters*

  In early stages of the process development cycle, there is often scarcity of rich enough spatial and temporal data to accurately model process variation. Accuracy or reliability of any variation model is directly proportional to the amount of data available during the extraction phase. Electrical data collected from test structures on production chips or test chips can provide substantial view of variability, and can readily be used in circuit simulation.

With increasing process complexity due to aggressive scaling, the set of model parameters at each level has exploded. Incorporating variability information from *all* processing steps for *all* model parameters is intractable. However, through a judicious choice of model parameter, preferably estimated at the electrical level of abstraction, we can use the variability information to create a robust circuit design.

1.5 An Overview of This Work

According to previous studies ([5, 6]), a complete description of LWR can be provided by three parameters: root-mean square (RMS) roughness (σ), correlation length (ξ), and roughness exponent (α). Estimation of LWR parameters is necessary for semiconductor process optimization, comparison of next-generation lithography (NGL) processes as well as device performance simulation.

In chapter 2, we propose a robust method of estimating LWR parameters. Using a vectorized block or *block of blocks* bootstrap technique for dependent data and a *weighted least squares* (WLS) fitting procedure, we fit a specific form of a variogram model. Block of blocks bootstrap is used to estimate the variance of a variogram, which in turn provides the WLS weights. Additionally, the bootstrap approach also allows us to estimate the error in the estimated LWR parameters, a vital requirement that has not been addressed by any of the previously reported procedures on this subject. Our procedure is shown to work even in the presence of some unknown local CD variation or if there is a systematic difference in

CD (by design or otherwise) between the lines. We validated our procedure with simulated roughness profiles with deterministic LWR parameters.

In chapter 3, we conduct a comprehensive comparative study of LWR in NGL processes using the estimation method developed in chapter 2. In this chapter, we investigate mainstream lithography options such as double patterning lithography (DPL), self-aligned double patterning (SADP), and extreme ultra-violet (EUV), as well as alternatives such as directed self-assembly (DSA), and nano-imprint lithography (NIL). The correlation length indicates the distance along the edge beyond which any two line width measurements can be considered independent. For NGL processes, it is found to range from 8 to 24 nm. It has been observed that LWR decreases when transferred from resist into the final substrate; all NGL technology options produce $< 10\%$ final LWR. We make meaningful comparison with LWR values stipulated by ITRS [15]. Additionally, spatial frequency transfer characteristics for DSA and SADP are also reported. Based on our study, the roughness exponent (which corresponds to local smoothness) is found to range from $\sim$0.75-0.98; it is close to being ideal ($\alpha = 1$) for DSA.

Earlier in section 1.3, we demonstrated how LWR parameters can be related to electrical parameter such as the transistor drive current of a planar CMOS transistor. However, the planar CMOS transistor may no longer be a viable device architecture in the near-term [1]. New device architecture such as multi-gate field effect transistor (MuGFET) and fully-depleted silicon-on-insulator (FDSOI) are expected to replace planar CMOS architecture [1]. FinFET is one such leading candidate for MuGFET. In chapter 4, we conclude our work on LWR by presenting a model for estimating the impact of gate line width roughness on the performance of double-gate (DG) FinFET devices. In this chapter, we present a framework to link device performance attributes (such as threshold voltage, drive current and off-state leakage) to LWR descriptors σ, ξ, and α. We provide physical insight into how LWR impacts device performance for 13nm gate length DG FinFETs, and we demonstrate that our modeling approach is more efficient than Monte-Carlo TCAD simulations, and it provides comparable results with appropriately selected input parameters. The FinFET device architecture is found to be robust to gate LWR effects. Furthermore, a spacer-defined gate electrode (vs. a resist-defined gate electrode) provides for reduced variability in performance, indicating that the gate-length mismatch has more impact than lateral offset between the front and the back gates.

In chapter 5, we perform a hierarchical decomposition of semiconductor process variation, and we closely examine the inter-relationship of across-wafer and across-die components specifically in terms of spatial correlation. The multivariate framework proposed for wafer-level variation is capable of handling scenarios where the across-wafer spatial variation has some unknown distribution. We also present method for rejecting outlier wafers and strategy for selecting wafer for model estimation.

Chapter 6 summarizes the contributions of this work, and suggestions for further research are offered therein.

Chapter 2

Robust Estimation of Line Width Roughness (LWR) Parameters

2.1 Introduction

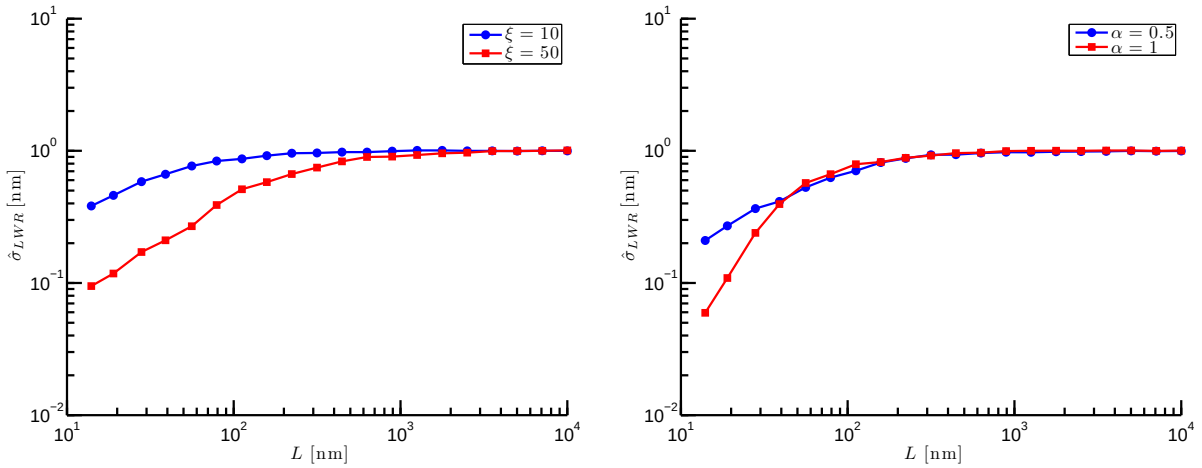
2.1.1 Motivation and Background

The formation of line edge roughness is a stochastic phenomenon. It has been shown to originate from many sources such as shot noise, mask roughness, resist diffusion statistics, chemical statistics, and quantum mechanics. During resist development, polymer aggregates along the edge are non-uniformly dislodged from their surrounding polymer matrix due to their different dissolution rates [16, 17]. Line *edge* roughness from two coupled edges leads to line *width* roughness. From a practical standpoint, line *width* is a more useful physical parameter than line *edge* for gate and interconnect definition. For convenience, we will use the term LER to describe the phenomenon that causes LWR.

With aggressive scaling of technology, LWR is increasingly becoming a larger component of the total variation [15]. It has generally been accepted that a complete description of LWR can be provided by three parameters: root-mean square (RMS) amplitude or standard deviation of line width (σ), correlation length (ξ), and roughness exponent (α) [5, 6]. These parameters can then be used for process characterization, transistor performance modeling or defining a roadmap [18, 19, 20, 15].

Scanning electron micrograph (SEM) image is the most practical source of data for estimating LWR parameters. However, the estimation of the aforementioned LWR parameters from the SEM image is a non-trivial task. The primary challenge in the estimation process is the limited availability of data in the SEM image. The SEM image presents us with few lines of finite length, and our objective is to estimate the parameters of the underlying LER process that generated this sample of data. It has been recognized that in presence of correlation between line widths at a given separation, the estimate of σ for a finite length of line can be significantly biased [5, 21]. Let $\hat{\sigma}$ and $\hat{\sigma}_{LWR}$ denote the *unbiased* and *biased*

estimators of σ respectively. Figure 2.1 shows a plot of $\hat{\sigma}_{LWR}$ as a function of line length L for simulated data. Note that $\hat{\sigma}_{LWR} \to \hat{\sigma}$ as length of the line $L \to \infty$, but there is a



(a) (b)

Figure 2.1: Biased estimate $\hat{\sigma}_{LWR}$ shown as a function of length of the line for (a) different correlation length values with $\alpha = 0.5$, and (b) different roughness exponent values with $\xi = 20$. The roughness profiles were simulated with $\sigma = 1$.

significant bias in $\hat{\sigma}_{LWR}$ for short line lengths.

The Semiconductor Industry Association (SIA) is semiconductor industry's leading consortium of global chip manufacturers, equipment suppliers and research communities. According to the guidelines provided by SIA in the International Technology Roadmap for Semiconductors (ITRS), $\hat{\sigma}_{LWR}$ is determined as the biased estimate of line width variation over a length greater than or equal to 2 μm measured at less than or equal 4 nm intervals [15]. However, most SEM images are captured at much higher resolution in order to accurately measure the average line width or critical dimension (CD) of lines. For example, figure 2.2 shows an SEM image of 40 nm full-pitch lines created by *double patterning lithography* (DPL). It shows 8 lines of approximately 288 nm length. The resolution of the image shown in figure 2.2 is 0.65 nm/pixel. Thus, contrary to the ITRS guidelines, the typically available SEM image has *shorter* lines scanned at *higher* resolution.

Leunissen, Lawrence and Ercken proposed a two-parameter (σ and ξ) model (henceforth referred to as 'LLE'); they assumed $\alpha = 0.5$ based on the data available to them at the time [21]. A two parameter model is less desirable, because it assumes some value of α (typically $\alpha = 0.5$ or $\alpha = 1$). The value of α may vary depending on the type of lithography process used to generate the lines. Additionally, in regards to the estimation of $\hat{\sigma}$, the LLE approach suffers from two drawbacks: (1) local non-LER related variations in CD are attributed to $\hat{\sigma}$ and (2) the estimate of $\hat{\sigma}$ rapidly deteriorates with reduction in number of lines in SEM

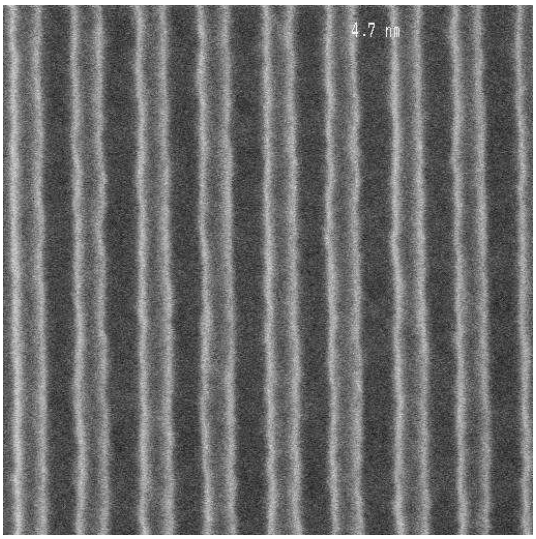


Figure 2.2: SEM image of lines generated by *litho-freeze-litho-etch* DPL process [22]

image [21]. Independently, Constantoudis et al developed a more flexible model that allows α to be fitted to data [5, 6], but the estimation of $\hat{\sigma}$ suffered from the same drawback as the LLE method. We will have a detailed discussion of the LLE method in section 2.3.

In summary, there is need for a robust extraction procedure that can provide a complete and accurate description of LWR for arbitrarily *short* and *fewer* number of lines.

2.1.2 Our Work

In this chapter, we present a procedure that provides a robust estimate of LWR parameters—$\hat{\sigma}$, $\hat{\xi}$, and $\hat{\alpha}$. Our estimation procedure is a confluence of spatial statistics and fractal concepts developed to understand the surface growth phenomena. Using the *block of blocks* bootstrap technique for dependent data and *weighted least squares* (WLS) fitting procedure, we fit a specific form of variogram model. Our procedure is shown to perform robustly in practical scenarios with limited data. *Block of blocks* bootstrap is used to estimate the variance of variogram, which in turn is used as weights in the WLS procedure. We first validate our procedure with simulated roughness profiles with deterministic LWR parameters, and then we test the robustness of the procedure using actual profiles from variety of different lithographic processes.

We use the term *robust* here to describe the stability of the proposed inference procedure in the presence of limited data *without* the central assumption of a Gaussian process. Moreover, our procedure works even in the presence of some unknown local CD variation or if there is a systematic difference in CD (by design or otherwise) between the lines. This aspect of our procedure (a) prevents non-LER sources of variation from being attributed to

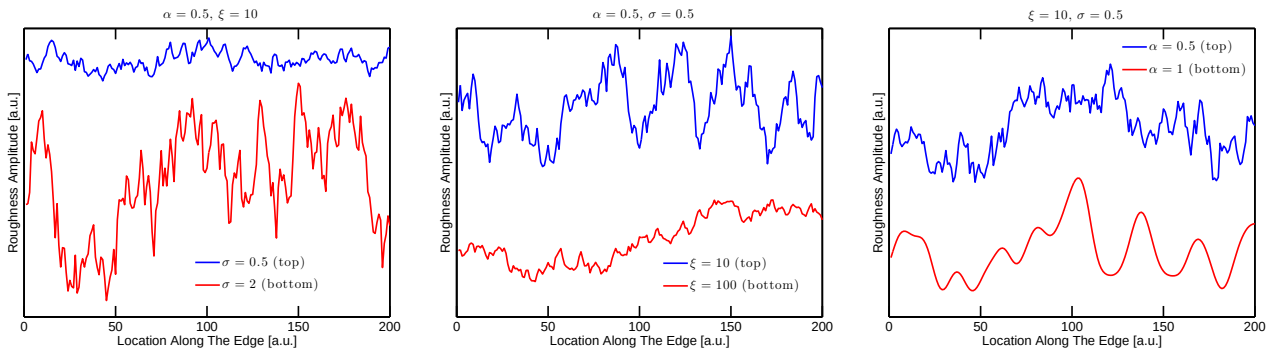
LER, and (b) it allows for more flexibility in capturing SEM images in that we do not need a special test structure with all of the lines having the same designed CD; any IC layout region with straight lines and arbitrary CDs would suffice. Additionally, the bootstrap approach also enables us to estimate the standard error in the estimated LWR parameters; to our knowledge, this vital and basic need has not yet been addressed by any of the previously reported procedures on this subject.

The rest of the chapter is organized as follows: In section 2.2, we provide a brief introduction to the LWR parameters and describe the LWR model. In section 2.3, we discuss the estimation of LWR parameters—$\hat{\sigma}$, $\hat{\xi}$, and $\hat{\alpha}$. A detailed outline of the complete procedure can be found in section 2.4. In section 2.5 we discuss the application of our procedure to simulated and actual data. Concluding remarks are made in section 2.6.

2.2 Line Width Roughness Model

2.2.1 LWR Parameters

A brief introduction to the LWR parameters is provided here for completeness. The RMS roughness or standard deviation (σ) is the most fundamental parameter to characterize LWR. Figure 2.3(a) shows two simulated roughness profiles with different values of σ, and it is fairly evident that larger values of σ correspond to greater roughness of the line. However, the statistic σ only provides a measure for transverse (to the line) fluctuations, and it does not describe any correlations between different lateral or longitudinal (along the line) locations. For instance, both simulated profiles shown in Figure 2.3(b) have the same value of σ. The



(a) (b) (c)

Figure 2.3: Illustration of the impact of (a) Roughness amplitude, RMS roughness or standard deviation of line width from its mean (σ), (b) auto-correlation length (ξ), and (c) roughness exponent (α) on line edge roughness. Note the differences in the oscillatory behavior of the peaks in (b). In (c), the higher value of α results in a smoother profile.

top (blue) roughness profile gives the appearance of being more rough than the bottom (red) profile. The lateral scale of fluctuations is noticeably different between the two profiles. In order to describe the spatial variation, we need to relate line widths at two different locations along the edge. The autocorrelation function ρ is a useful tool in describing such correlation. Generally for LER, ρ is described by a monotonically decreasing function expressed in terms of distance between the two locations along the line (defined as the lag h). It is characterized by a single parameter called correlation length ξ, defined as $\rho(\xi) \equiv 1/e$. ξ is a representative lateral dimension used to describe the roughness profile. For $h < \xi$, the line widths at these locations are considered to be correlated. Conversely, for $h \gg \xi$, the two locations can be considered independent. Unlike σ (lower the better), the desired trend in ξ is not obvious. However, as shown in chapter 1, a shorter correlation length is desirable as it leads to lower variability in the electrical parameters.

One final parameter, α, that is used to describe LWR is illustrated in Figure 2.3(c). It shows two simulated roughness profiles with the same values of σ and ξ. The profile pictured on top (blue) appears to be more jagged than the one shown on the bottom (red). This demonstrates that the two-parameter description of LWR is insufficient. To complete the description of roughness, fractal concepts from thin-film materials science have been employed [23]. A self-affine object remains scale invariant under an affine transformation. That is, one can re-scale the roughness profile laterally and longitudinally, and obtain a new profile that is statistically identical to the original profile. A self-affine roughness profile has the property [23]

$$f(x) = \epsilon^{-\alpha} f(\epsilon x), \tag{2.2.1}$$

where α is the roughness exponent, ϵ is a scale factor, and $f(x)$ is the line width at location x. Thus, for a self-affine function f, the function values scale as a power law. The roughness exponent α (also known as the *Hurst exponent*) characterizes *short-range* roughness and for a d-dimensional surface it is directly related to the fractal dimension D by the identity $\alpha = d - D$ [23]. Physically, α can be thought of as a local smoothness parameter. It can be shown that $0 \leq \alpha \leq 1$ [23]. As shown in Figure 2.3(c), higher value corresponds to a locally smooth roughness profile.

2.2.2 A Variogram Model

In this section we use concepts from spatial statistics to formulate a model using the LWR parameters discussed above. Subsequently, we will estimate the LWR parameters by fitting this model to the roughness data.

The line width roughness profile can be generalized as a spatial sequence of random variables. A spatial sequence is considered *intrinsically stationary* if its finite dimensional joint distributions do not change when shifted in position. Consider an intrinsically stationary spatial sequence in region $\mathcal{R}$, i.e., let $\mathcal{X}_s = \{X_i : i \in \mathbb{Z}\}$, be a collection of random variables

with an unknown mean $\mu \in \mathbb{R}$ such that

$$E(X_i - X_{i+h}) = 0, \tag{2.2.2}$$

and

$$\mathrm{Var}(X_i - X_{i+h}) = \mathrm{Var}(X_0 - X_h) = 2\gamma(h) \tag{2.2.3}$$

for all $(i, h) \in \mathbb{Z}$. (2.2.2) implies that the mean is constant everywhere in region $\mathcal{R}$. (2.2.3) implies that the variance of the *difference* is constant everywhere in region $\mathcal{R}$, and that it depends only on h. In spatial statistics, 2γ is known as the *variogram*, and γ is known as the *semi-variogram*. In thin film material science, the variogram is also known as the *height-height correlation function* [23, 24]. Additionally, if $\mathcal{X}_s$ is *second-order or weakly stationary*, then (2.2.2) holds, and $\mathcal{X}_s$ has a common *auto-covariance function* $C(h) = \mathrm{Cov}(X_0, X_h)$ such that

$$\mathrm{Var}(X_i - X_{i+h}) = \mathrm{Var}(X_i) + \mathrm{Var}(X_{i+h}) - 2\mathrm{Cov}(X_i, X_{i+h}) \tag{2.2.4}$$

$$= 2\left[C(0) - C(h)\right]. \tag{2.2.5}$$

Thus, second-order stationarity implies intrinsic stationarity, and we have

$$2\gamma(h) = 2\left[C(0) - C(h)\right] \tag{2.2.6}$$

for $h \in \mathbb{Z}$. Previously, we introduced auto-correlation function ρ as a tool to describe the spatial correlation between line widths at two locations along a line. The auto-correlation function can be defined in terms of the auto-covariance function as

$$\rho(h) \equiv \frac{C(h)}{C(0)}, \tag{2.2.7}$$

where $C(0) = \sigma^2$ is the variance of spatial sequence $\mathcal{X}_s$. Using (2.2.6) and (2.2.7), we can define the variogram in terms of the auto-correlation function as

$$2\gamma(h) = 2\sigma^2\left[1 - \rho(h)\right] \tag{2.2.8}$$

The auto-correlation function ρ has the following important properties:

1. $\rho(0) = 1$
2. $\rho(h) = \rho(-h)$
3. $|\rho(h)| \leq \rho(0)$
4. $\lim_{h\to\infty} \rho(h) = 0$

The last relation holds for a wide class of stationary processes, including the spatial process considered here, but not in general. As previously mentioned, the correlation length ξ represents a characteristic lateral dimension of the roughness profile. Several analytic forms exist that satisfy (2.2.1) [24]. For self-affine and isotropic surfaces, a specific functional form was proposed by [7],

$$\rho(h;\xi,\alpha) = \exp\left(-\left[\frac{|h|}{\xi}\right]^{2\alpha}\right),\ h \in \mathbb{Z}^d. \tag{2.2.9}$$

Let $\theta \equiv (\xi,\alpha)' \in (0,\infty)\times[0,1]$ denote the structural parameter of the auto-correlation function defined by (2.2.9). A phenomenologically correct self-affine form for ρ was proposed by [25], but we prefer (2.2.9) because it lends itself better for fitting purposes. Note that the choice of a particular form for ρ does not alter our procedure other than a simple change of fitting function. Using (2.2.9), we can now rewrite (2.2.8) as

$$2\gamma(h;\sigma^2,\theta) = 2\sigma^2\Big[1-\exp\Big(-\Big[\frac{|h|}{\xi}\Big]^{2\alpha}\Big)\Big],\ h \in \mathbb{Z}. \tag{2.2.10}$$

Using the Taylor expansion of the exponential function, it can be shown that the variogram (2.2.10) satisfies the self-affine property (2.2.1), and exhibits power law behavior $2\gamma \sim h^{2\alpha}$ for $h \ll \xi$.

2.3 Estimation of Model Parameters

Both sources of roughness data employed in our work—the simulated data and the SEM image, generate data on a regular grid. We can generalize both of these sources into a common framework as follows. Consider M independent realizations of a *stationary* LER process:

$$\mathcal{X}_n \equiv \{X_{is} : s = 1,\ldots,L,\ i = 1,\ldots,M\} \tag{2.3.1}$$

with unknown *line-wise* means $\mu_i \in I\!R$ and common auto-covariance function $C(h)$

$$E[(X_{is}-\mu_i)(X_{i(s+h)}-\mu_i)] = C(h) = \sigma^2\rho(h). \tag{2.3.2}$$

Here σ^2 and $\rho(h)$ denote the variance and the auto-correlation function of the underlying LER process respectively, M denotes the number of available lines, and Ld is the physical length of each line with grid spacing $d \in I\!R$. L is the length of the line normalized to grid spacing, and will be used interchangeably with Ld. The generalized framework is graphically illustrated in figure 2.4.

2.3.1 Estimation of ξ and α

We adapt ideas from spatial statistics and estimate the structural parameter $\theta \equiv (\xi,\alpha)'$ of the variogram (2.2.10). We use a version of Matheron's *classical* variogram estimator [26]

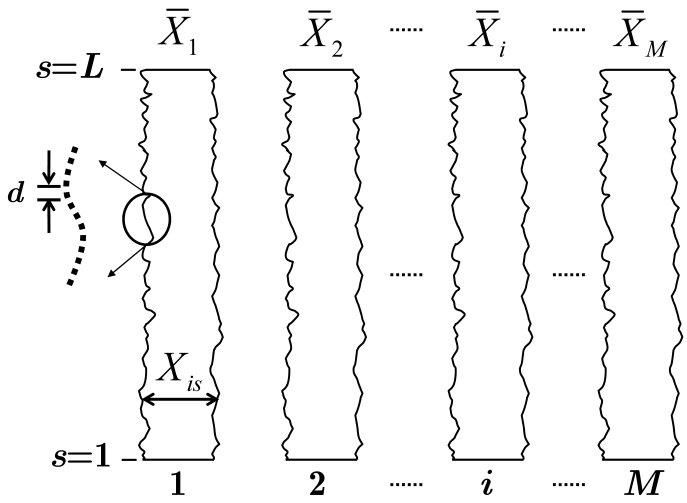


Figure 2.4: Generalized framework of simulated and digitized SEM data. A typical SEM image can be generalized to consist of M lines where each line has L number of measurements collected at a spacing d. X_{is} denotes the line width at location s on line i and $\bar{X}_i$ denotes the average line width (or CD) of the i-th line.

to fit our model (2.2.10) using the weighted least squares (WLS) method. Our estimation procedure closely follows method described in [27]. A version of the *classical* variogram estimator based on all ML observations can be given as [26]

$$2\hat{\gamma}(h) = \frac{1}{N_h} \sum_{i=1}^{M} \sum_{s=1}^{L-h} \left(X_{is} - X_{i(s+h)} \right)^2, \tag{2.3.3}$$

where

$$N_h = M(L-h) \tag{2.3.4}$$

denotes the number of pairs available at lag h.

Recall that the variogram and covariogram are related by (2.2.6). We prefer using variogram estimation over covariogram estimation for two reasons: (i) it completely avoids the estimation of the mean parameters μ_i's (that are of little interest in this case), and (ii) it ensures a higher level of accuracy (a mean squared error of order $O([ML]^{-1})$ that is not masked by the error of estimating the individual mean parameters based on observations from a single line (giving MSEs of the order $O(L^{-1})$). Additional statistical arguments in favor of the variogram have also been made [28].

Note that (2.3.4) indicates that fewer pairs are available at higher lag values. Thus, the variogram estimation error increases with the lag. The WLS method automatically provides

heigher weights for early lags and lower weights for the lags at which number of contributing pairs is low [27]. Several methods have been proposed for fitting variogram models [28], but the robustness of WLS, and the absence of any distributional assumptions, makes it the most practical method for fitting variogram [29]. Assuming *heteroskedasticity*, the WLS criterion is to minimize

$$(2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta}))' \boldsymbol{V}^{-1} (2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta}))$$

where $\boldsymbol{V}$ is diagonal matrix of variances of variogram. Thus, we can define the WLS estimator of (σ^2, θ) as

$$(\hat{\sigma}^2_{WLS}, \hat{\theta}) = \underset{\sigma^2,\theta}{\operatorname{argmin}} \sum_{h=1}^{h_0} w(h) \left[2\hat{\gamma}(h) - 2\gamma(h; \sigma^2, \theta)\right]^2, \tag{2.3.5}$$

where h_0 is a user specified upper range of lag values, $2\hat{\gamma}(h)$ and $2\gamma(h; \sigma^2, \theta)$ are given by (2.3.3) and (2.2.10) respectively, and $w(h)$ are the weights given to observations with lag h such that

$$w(h) = \frac{1}{\operatorname{Var}(2\hat{\gamma}(h))}. \tag{2.3.6}$$

Bootstrap and subsampling methods for dependent data can be employed to estimate the variances of the variogram at different lags [30, 31]. However, at the time of Cressie's publication in 1985, such methods had not yet been developed. Cressie suggests the following approximation for the variance of $2\hat{\gamma}$ [27]

$$\operatorname{Var}(2\hat{\gamma}(h)) \approx \frac{2[2\gamma(h; \sigma^2, \theta)]^2}{N_h}. \tag{2.3.7}$$

Using this approximation, (2.3.5) can be rewritten as

$$(\hat{\sigma}^2_{WLS,CA}, \hat{\theta}_{CA}) = \underset{\sigma^2,\theta}{\operatorname{argmin}} \sum_{h=1}^{h_0} N_h \left[\frac{2\hat{\gamma}(h)}{2\gamma(h; \sigma^2, \theta)} - 1\right]^2. \tag{2.3.8}$$

As an alternative to the Cressie approximation (2.3.7), in section 2.4 we will use *block of blocks bootstrap* method to compute $\operatorname{Var}(2\hat{\gamma}(h))$. We will compare the bootstrap weight based $\hat{\theta}$ from (2.3.5) with the Cressie approximated $\hat{\theta}_{CA}$ in the discussion section.

The WLS estimator for (σ^2, θ) defined in (2.3.5) can be solved by any non-linear optimization procedure. Most mathematical packages such as MATLAB provide built-in functions for constrained non-linear optimization [32]. Lastly, we would like to remark on an important practical consideration—the choice of maximum lag h_0. In our framework, for lines of length L, the largest possible lag is $H = (L - 1)$. A practical choice of h_0 is then [27]

$$h_0 = \operatorname{argmax}\{h : h \leq H/2 \text{ and } N_h \geq 30\}. \tag{2.3.9}$$

2.3.2 Estimation of σ^2

Pursuant to the framework defined earlier, the sample mean of each line (commonly referred to as the critical dimension or CD) can be given by

$$\bar{X}_i = L^{-1} \sum_{s=1}^{L} X_{is}. \tag{2.3.10}$$

We can estimate the variance in CD or variance in (2.3.10) as

$$\hat{\sigma}^2_{CD} = \frac{1}{M-1} \sum_{i=1}^{M} (\bar{X}_i - \bar{X})^2, \tag{2.3.11}$$

where $\bar{X}_i$ is given by (2.3.10), and $\bar{X}$ is the grand average defined as $\bar{X} = M^{-1} \sum_{i=1}^{M} \bar{X}_i$.

The mean-adjusted sample variance (an estimator of σ^2) is given by

$$\widehat{\sigma}^2_{LWR} = \frac{1}{ML} \sum_{i=1}^{M} \sum_{s=1}^{L} (X_{is} - \bar{X}_i)^2. \tag{2.3.12}$$

Figure 2.1 shows a plot of $\widehat{\sigma}^2_{LWR}$ as a function of length L using simulated data. Note that $\widehat{\sigma}^2_{LWR}$ is a poor estimator of σ^2, and it has a significant bias at short lengths when ξ is large or α is small [5, 21]. The guidelines provided in the ITRS roadmap stipulate that $\widehat{\sigma}^2_{LWR}$ should be measured over lengths greater than or equal to 2 μm [15]. For typical values of ξ and α, $\widehat{\sigma}^2_{LWR}$ would be *approximately unbiased* if $L \geq 2\mu m$. However, for $L \sim 200 - 1000$ nm, we need a explicit bias correction in $\widehat{\sigma}^2_{LWR}$. One such bias correction method was proposed by Leunissen, Lawrence and Ercken (henceforth referred to as 'LLE')[21]. It was empirically observed that adding (2.3.11) to (2.3.12) provided the necessary bias correction, and it was proposed that [6, 21]

$$\hat{\sigma}^2_{LLE} \equiv \widehat{\sigma}^2_{LWR}(L) + \hat{\sigma}^2_{CD}(L) = \hat{\sigma}^2. \tag{2.3.13}$$

In other words, the sum of variances defined by (2.3.11) and (2.3.12) is *invariant* with the length of line L, and that it equals the *unbiased* variance of an infinitely long line. The relationship (2.3.13) was demonstrated to hold by using simulated data [6]. Under independence, this is a well-known result in the Linear Models theory in Statistics [33]. However, to the best of our knowledge, a mathematical proof for the assertion (2.3.13) has not yet been provided in the literature for dependent variables. We provide a proof for relation (2.3.13) in Appendix B. Equation (2.3.13) does not perform well when the number of lines in the SEM image M is greatly reduced. Indeed, (2.3.13) was demonstrated to hold with $M = 500$ [6]. However, for typical values of $M \sim 10 - 12$, (2.3.13) provides a poor estimate. This poses a real problem for advanced lithographic technologies such as DPL. There are two lithographic sequences in DPL, and every alternate line belongs to the same lithographic sequence. LER characteristics of each lithographic sequence can be significantly

different, and they need to be analyzed separately [22]. As such, in case of DPL, only half of the lines are available for estimating $\hat{\sigma}$ of each lithographic step. The LLE method can be modified for DPL; the $\hat{\sigma}^2_{CD}$ term in (2.3.13) can be separated into two terms based on alternate lines and for each of these terms the value of M could be as low as 3! Thus, although a simple modification of the LLE procedure circumvents this limitation for DPL, it comes at a cost of reduced accuracy, because only half the number of lines can now be used for each lithographic step. Thus, $\hat{\sigma}^2_{LLE}$ is not a robust estimator of σ^2. But even more importantly, the use of (2.3.13) to estimate the bias corrected value of σ^2 can fundamentally be flawed. The $\hat{\sigma}^2_{CD}$ term in (2.3.13) incorporates all local *non-LWR* sources of variability such as mask CD variation (among other factors). This variation can also be lithography process induced systematic or random. An estimate of σ^2 based on (2.3.13) would be inflated by the amount of local variability present in the SEM lines.

We propose a more explicit bias correction to $\widehat{\sigma}^2_{LWR}$. It is straightforward to show (see Appendix A) that

$$E[\widehat{\sigma}^2_{LWR}(L)] = \sigma^2 f(L) \tag{2.3.14}$$

where $f(L)$ represents the bias factor. For the special case of $\alpha = 0.5$, Leunissen et al did recognize the need for explicit bias correction, but they did not develop any procedure to use it further [21]. For the assumed form of the auto-correlation function given by (2.2.9), it is possible to derive a closed form expression for $f(L)$ for $\alpha = 0.5$ and $\alpha = 1$ (see Appendix A). However, it is better to use the numerical representation

$$f(L;\theta) \equiv \left(1 - \frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L}\rho(s-t)\right), \tag{2.3.15}$$

because it applies for any choice of α and ξ. Here ρ is defined by (2.2.9) with a structural parameter $\theta \equiv (\xi, \alpha)'$. Substituting the estimates of the structural parameter found in the previous by WLS, we can define our estimator of σ^2 as

$$\hat{\sigma}^2_{BBB} = \widehat{\sigma}^2_{LWR}\left[\hat{f}(L;\hat{\theta})\right]^{-1}. \tag{2.3.16}$$

In summary, we have four available estimators of σ^2: $\hat{\sigma}^2_{LWR}$, $\hat{\sigma}^2_{WLS}$, $\hat{\sigma}^2_{LLE}$, and $\hat{\sigma}^2_{BBB}$. $\hat{\sigma}^2_{LWR}$ is a *biased* estimator of σ^2, and $\hat{\sigma}^2_{WLS}$, $\hat{\sigma}^2_{LLE}$, and $\hat{\sigma}^2_{BBB}$ are the *unbiased* estimators of σ^2. We will compare the performance of each estimator in section 2.5.

2.4 Robust Estimation of LWR Parameters

In a 1979 seminal paper, Efron introduced a non-parametric resampling technique in the context of *independently and identically distributed* (IID) data (henceforth referred to as the IID bootstrap) [34]. The IID bootstrap is a conceptually simple way to make inferences regarding the distributional properties of an unknown distribution. In this method, the

empirical distribution is resampled with replacement to create replicates of the original observations. This process is commonly referred to as *bootstrapping*. Bootstrap replicates can be used to estimate the standard error, bias, and confidence interval of parameters. A very nice introduction to the bootstrapping techniques can be found in [35]. The IID bootstrap quickly became popular due to its simplicity and the advancements in computational power. It found applications in large variety of statistical problems. Shortly after Efron introduced the IID bootstrap, in 1981 Singh described its limitation for dependent data [36], and in 1986 Carlstein proposed the non-overlapping block method to deal with weakly dependent data [37]. Independently, Kunsch [38] and Liu and Singh [39] formulated a new resampling scheme called the *moving block bootstrap* in 1989 and 1992, respectively. The name *moving block bootstrap* (MBB hereafter) was coined by Liu and Singh [39]. In contrast to the IID bootstrap, where only *single* observations are resampled, in the MBB approach *blocks* of consecutive observations are resampled at random and with replacement. The resampled blocks are then concatenated together to form a bootstrapped version of the original series. Note that the blocks defined in the MBB method are *one-dimensional*, and thus, they cannot be used for estimating statitics that are based on sample lag correlations. For example, for estimating the variogram we need to incorporate the lagged version of the *original* sequence into the definition of the blocks. A variant of MBB that covers such statistics was proposed by Kunsch [38], and it was further explored by Politis and Romano [40]. The modified blocking scheme was called *vectorized block bootstrap* [38] or *block of blocks* [40]. We will refer to the block of blocks bootstrap method as BBB hereafter. In the following, we illustrate the difference between the MBB and BBB. Let the original sequence be denoted as $\{R_n\}$. In the BBB method, we define a *new sequence* that includes the lagged version of the original sequence such that

$$Y_j = (R_j, \ldots, R_{j+h}),\ 1 \leq j \leq (n-h).$$

We then define blocks of in terms of consecutive Y-values instead of R, and subsequently, resample and concatenate these blocks as in the MBB method. A graphical illustration of block of blocks method is shown in Figures 2.5 and 2.6. In the MBB method, the blocks are defined as shown in Figure 2.6, albeit directly for the original sequence $\{R_n\}$ instead of $\{Y_n\}$. A detailed exposition of the bootstrap methods and their properties for dependent data can be found in [30].

2.4.1 Procedure

In the following discussion, we will describe the application of the BBB method for estimating the weights used in fitting the variogram as well as estimating the standard error (or variance) in the LWR parameters. Consider the framework defined in the beginning of section 2.3. Our objective here is to estimate the LWR parameters ($\hat{\sigma}$, $\hat{\xi}$, and $\hat{\alpha}$) and their respective standard error or variance. The steps in our extraction algorithm using BBB are as follows:

1. Let $e_{is} = X_{is} - \bar{X}_i,\ s = 1, \ldots, L,\ i = 1, \ldots, M$ denote the residuals. Define a new

sequence of *centered* residuals

$$\mathcal{R}_n \equiv \{R_{is} = e_{is} - \bar{e} : \ s = 1, \ldots, L, \ i = 1, \ldots, M\}.$$

where $\bar{e} \equiv (ML)^{-1} \sum_{i=1}^{M} \sum_{s=1}^{L} e_{is}$ is the grand average of all e_{is}'s.

2. Given $\mathcal{R}_n$ and lag h, define a new sequence

$$Y_{ip} \equiv \Big(R_{ip}, \ldots, R_{i(p+h)}\Big)', \ p = 1, \ldots, (L-h), \ i = 1, \ldots, M.$$

For later use, we also denote the components of Y_{ip} (that is R_{is}'s) as $Y_{ip,1}, \ldots, Y_{ip,(1+h)}$. Figure 2.5 shows a graphical representation of the new sequence Y_{ip} formed based on lagged version of the sequence R_{is}.

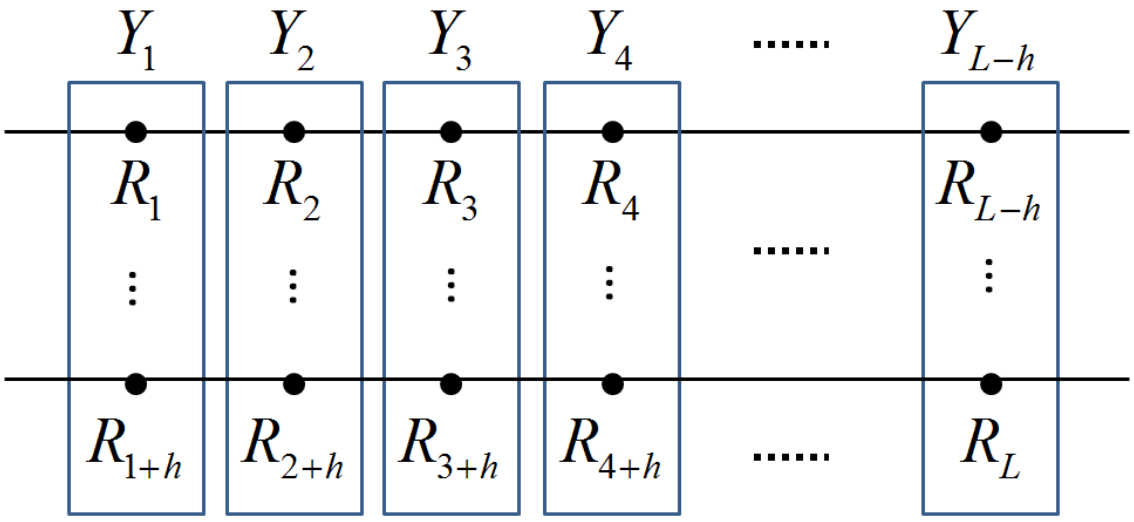


Figure 2.5: Graphical representation of the new sequence Y_{ip} formed based on lagged version of the sequence R_{is}. Note that the i subscripts are dropped from Y_{ip} and R_{ip} for clarity.

3. For a block length ($\ell : 1 < \ell < (L - h_0)$), let $k \equiv \lceil L/\ell \rceil$ where $k \in \mathbb{Z}$, and h_0 is the maximum lag defined in (2.3.9). Here, $\lceil x \rceil$ denotes the ceiling of x, i.e. the smallest integer greater than or equal to x.

4. Define $\mathcal{B}(i, j)$ to be the block of ℓ consecutive observations of Y_{ip} within each line such that

$$\mathcal{B}(i, j) \equiv \Big(Y_{ij}, \ldots, Y_{i(j+\ell-1)}\Big), \ j = 1, \ldots, (L - h - \ell + 1), i = 1, \ldots, M. \qquad (2.4.1)$$

Figure 2.6 shows a graphical representation of the block definition.

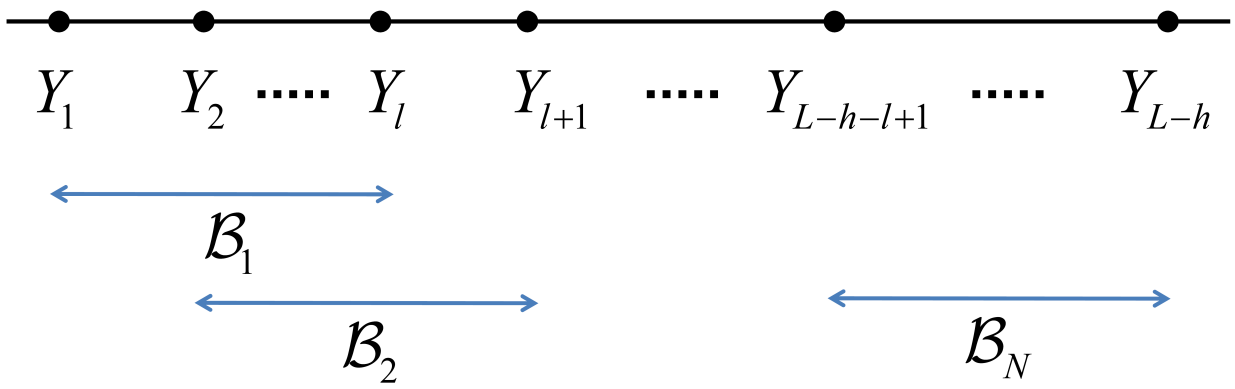


Figure 2.6: Graphical representation of the blocks under the *block of blocks* method. Note that the i subscripts are dropped from Y_{ip} and $\mathcal{B}(i,j)$ for clarity.

5. Resample Mk-many blocks at *random and with replacement*, and reconstruct the MBB version of the Y_{ip} series by concatenating the values in the resampled blocks. Following the discussion in [41], note that it is not always possible to define the block size ℓ that satisfies $L = kl$ where $k \in \mathbb{Z}$. In such cases, we let $k \equiv \lceil L/\ell \rceil$, delete the last $(kl - L)$ observations, and retain only the first L resampled values for the reconstruction of each line [41].

6. From step 5, we have a bootstrapped version of Y_{ip} series for which we can compute the bootstrapped version of sample variance (2.3.12)

$$(\sigma^2_{LWR})^* = \frac{1}{ML}\sum_{i=1}^{M}\sum_{p=1}^{L}(Y^*_{ip,1} - \bar{Y}^*_{i,1})^2. \tag{2.4.2}$$

   Here $\bar{Y}^*_{i,1} = L^{-1}\sum_{p=1}^{L} Y^*_{ip,1}$.

7. To define the BB version of $2\hat{\gamma}(h)$ of (2.3.3), note that in terms of Y_{ip}'s, we can write

$$2\hat{\gamma}(h) = \frac{1}{M(L-h)}\sum_{i=1}^{M}\sum_{p=1}^{L-h}\left(Y_{ip,1} - Y_{ip,(1+h)}\right)^2.$$

   Hence, we define the bootstrapped variogram $2\gamma^*$ as

$$2\gamma^*(h) = \frac{1}{M(L-h)}\sum_{i=1}^{M}\sum_{p=1}^{L-h}\left(Y^*_{ip,1} - Y^*_{ip,1+h}\right)^2, \tag{2.4.3}$$

obtained by replacing $\{Y_{ip} : p = 1, \ldots, L-h, i = 1, \ldots, M\}$ by the resampled values $\{Y^*_{ip} : p = 1, \ldots, L-h, i = 1, \ldots, M\}$

8. Repeat steps 5 through 7 a large number of times (say, $B = 500+$ times), independently (i.e., select the Mk blocks from the collection (2.4.1) at random, every time). B denotes the number of bootstrap samples. Thus, we have B bootstrap sets of (2.4.2) and (2.4.3).

9. Compute the variance of $\hat{\gamma}(h)$ using

$$\widehat{\text{Var}}(2\hat{\gamma}(h)) = [w(h)]^{-1} = \frac{1}{B-1}\sum_{b=1}^{B}\left[2\gamma^{*b}(h) - 2\bar{\gamma}^*(h)\right]^2, \tag{2.4.4}$$

where $2\gamma^{*b}$ is the estimate of 2γ computed using the b-th bootstrap sample data set in step 7, and $\bar{\gamma}^*$ is the average of $\{2\gamma^{*b} : b = 1, \ldots, B\}$.

10. Using (2.4.4) and (2.3.5), compute $\theta^* \equiv (\alpha^*, \xi^*)'$ for each of the B bootstrap sets of variogram from step 8.

11. From step 10, we have $B = 500+$ values of α^* and ξ^* that can be used for setting confidence limits for the parameters, as well as for estimating the standard errors of our estimators. Estimator of the variance of α is given by the sample variance of the B replicates as

$$\widehat{\text{Var}}(\hat{\alpha}) = \frac{1}{B-1}\sum_{b=1}^{B}\left[\alpha^{*b} - \bar{\alpha}^*\right]^2, \ \{\alpha^{*b} : b = 1, \ldots, B\}, \tag{2.4.5}$$

where α^{*b} is the estimate of α computed using the bth resampled data set in step (c), and where $\bar{\alpha}^*$ is the average of $\{\alpha^{*b} : b = 1, \ldots, B\}$. Similarly, the estimator of variance of $\hat{\xi}$ is given by

$$\widehat{\text{Var}}(\hat{\xi}) = \frac{1}{B-1}\sum_{b=1}^{B}\left[\xi^{*b} - \bar{\xi}^*\right]^2, \ \{\xi^{*b} : b = 1, \ldots, B\}, \tag{2.4.6}$$

where ξ^{*b} is the estimate of ξ computed using the bth resampled data set in step (c), and where $\bar{\xi}^*$ is the average of $\{\xi^{*b} : b = 1, \ldots, B\}$.

12. To compute the standard error in $\hat{\sigma}^2_{BBB}$, we proceed similarly. For each of the B bootstrap replicates, we compute $(\hat{\sigma}^2_{BBB})^{*b}$ using the corresponding values of α^{*b}, ξ^{*b} and $(\sigma^2_{LWR})^{*b}$ in (2.3.16). Then, the block bootstrap estimator of the variance of $\hat{\sigma}^2_{BBB}$ is given by

$$\widehat{\text{Var}}(\hat{\sigma}^2_{BBB}) = \frac{1}{B-1}\sum_{b=1}^{B}\left[(\sigma^2_{BBB})^{*b} - (\bar{\sigma}^2_{BBB})^*\right]^2, \tag{2.4.7}$$

where $(\bar{\sigma}^2_{BBB})^*$ is the average of $\{(\sigma^2_{BBB})^{*b} : b = 1, \ldots, B\}$.

2.4.2 Optimal Block Length

The accuracy of the block bootstrap (BB) methods critically depends on the choice of the *block length* [30]. Large block sizes reduce the bias, but inflate the variance of the block bootstrap estimator, and smaller block sizes tend to have the opposite effects. The *optimal block length* is chosen such that the mean-squared error or MSE (that takes into account the combined effect of a block length on both the bias and the variance parts) of the block bootstrap estimator is minimized. In the literature, more than one data dependent methods for choosing the optimal block size has been proposed; see, for example, Chapter 7 of [30]. A computationally simple method for block length selection is given by Politis and White ([42]). We now provide a description of the main steps of this method.

Main Steps:

1. Compute
$$\hat{C}(k) = \frac{1}{M(L-|k|)} \sum_{i=1}^{M} \sum_{p=1}^{L-|k|} (Y_{ip,1} - \bar{Y}_i)(Y_{ip,(1+|k|)}) - \bar{Y}_i),$$
a version of the sample autocovariance of $\{Y_{ip}\}$ at lag k, where
$$\bar{Y}_i = \frac{1}{(L-|k|)} \sum_{p=1}^{L-|k|} Y_{ip,1}.$$

2. Choose k_0 such that $\hat{C}(k) \approx 0$ for all $k > k_0$. In other words, find the smallest integer k_0 beyond which $\hat{C}(k)$ is not significantly different from zero.

3. Compute
$$\hat{G} = \sum_{k=-2k_0}^{2k_0} \lambda(k/[2k_0])|k|\hat{C}(k),$$
where $\lambda(\cdot)$ is the flat-top kernel of Politis and Romano (1995):
$$\lambda(t) = \begin{cases} 1 & \text{if } |t| \in [0, 1/2) \\ 2(1-|t|) & \text{if } |t| \in [1/2, 1] \\ 0 & \text{otherwise.} \end{cases}$$

4. Compute $\hat{D} = \sum_{k=-2k_0}^{2k_0} \lambda(k/[2k_0])\hat{C}(k)$.

5. Compute the estimated optimal block size by
$$\ell_{opt} = \left[\left(\frac{3\hat{G}^2}{2\hat{D}^2} \right) L \right]^{1/3}$$

2.5 Results and Discussion

Equation (2.2.10) provides us a functional form of the variogram model using LWR parameters. In order to extract the LWR parameters from (2.2.10), we need roughness profiles to estimate the variogram. In this work, simulated roughness data was used to test and validate our procedure. A brief description of the procedure used to generate simulated data is provided below. Once the procedure was validated using simulated data, we applied it to actual roughness profiles obtained by inline metrology. Details regarding SEM image acquisition and post-processing will be discussed in the next chapter.

2.5.1 Simulated Data

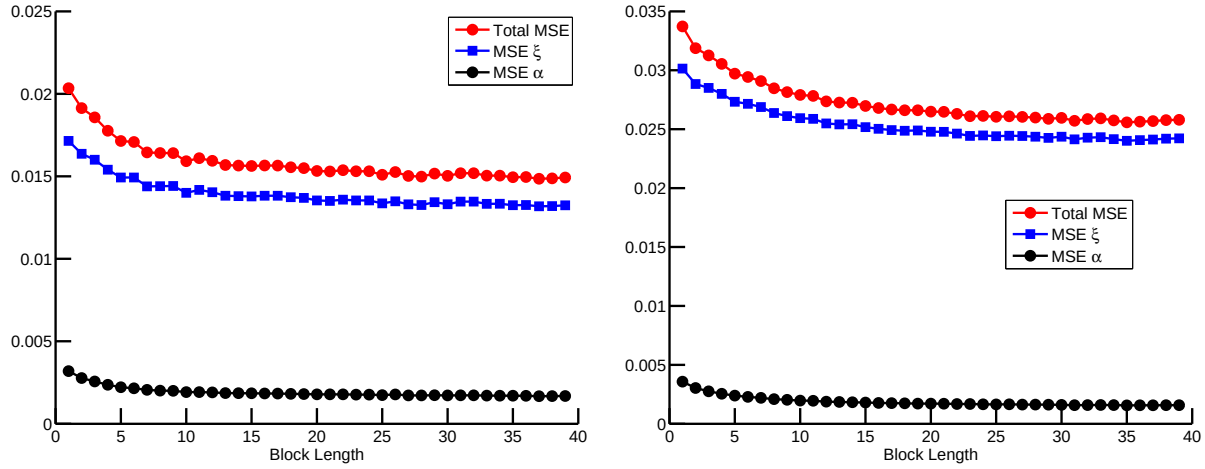
We generate simulated roughness profile with the desired LWR parameter using the convolution method described in [24]. The outline of the procedure is roughly as follows:

1. Compute $\rho(h)$ using (2.2.9) for desired values of ξ and α.
2. Compute the power spectrum $P(k)$ by taking the Fourier transform of $\rho(h)$ computed in step 1.
3. Compute the amplitude $p_n = \sqrt{P(k)}$.
4. Numerically generate a Gaussian white noise sequence X_n and compute its Fourier transform x_n.
5. Compute the Fourier transform of the roughness profile $y_n = x_n p_n$.
6. Compute the inverse Fourier transform of y_n to obtain the roughness profile Y_n with unit variance.

The roughness profile thus obtained has a unit variance. To obtain a *sampled* roughness profile of length L from a process with a desired variance, we generate a unit variance profile of length $N \gg L$ using the method described above. The resulting sequence can subsequently be scaled to have the desired variance, and L values can be sampled from the middle to avoid any discrete FFT edge artifacts on a finite series.

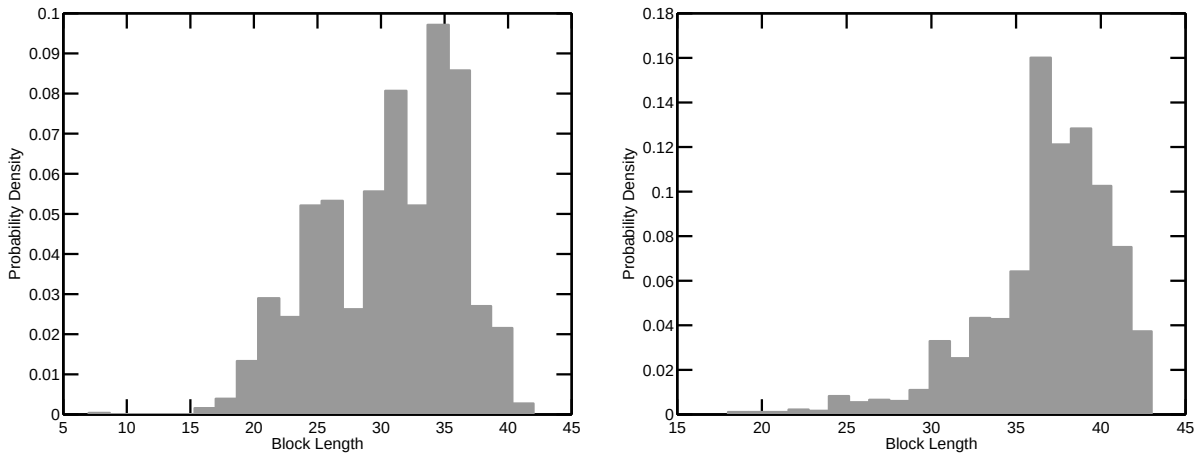
2.5.2 Validation Using Simulated Data

The choice of optimal block length is of paramount importance in determination of the shape parameters ξ and α. First, we investigate the MSE of each shape parameter as well as the total MSE in terms of the choice of block length. For $\alpha = 0.5$, we can see in Figure 2.7 that the total MSE stabilizes to a low value for a block length of approximately 25. The results are based on 200 sample Monte-Carlo simulation. For longer correlation length, MSE



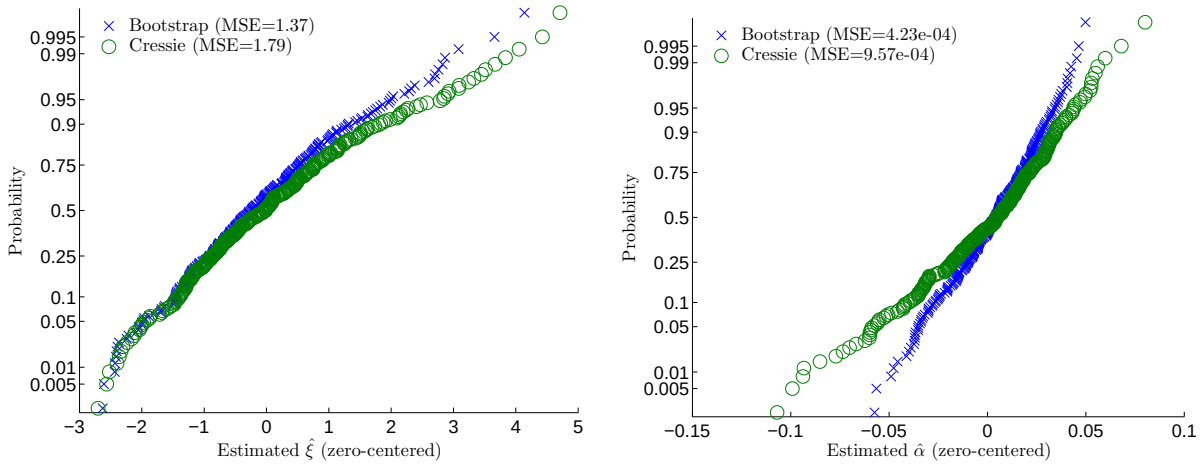
(a) (b)

Figure 2.7: Mean-squared error (MSE) in α and ξ as a function of the block length for (a) $\xi = 10$, and (b) $\xi = 20$. Total MSE is the normalized sum of MSE of α and ξ. The roughness profiles were simulated with $\alpha = 0.5$, $\sigma = 1$, $M = 8$, and $L = 500$. The results shown here are from 200 sample Monte-Carlo simulation.



(a) (b)

Figure 2.8: Optimal block length computed using Politis and White method ([42]) for (a) $\xi = 10$, and (b) $\xi = 20$. The roughness profiles were simulated with $\alpha = 0.5$, $\sigma = 1$, $M = 8$ and $L = 500$. The block length was computed using the same data set as in Figure 2.7.



(a) (b)

Figure 2.9: Comparison of two methods of estimating WLS weights for shape parameters (a) ξ and α based on 200 sample Monte-Carlo experiment. The roughness profiles were simulated with $\xi = 10$, $\alpha = 0.5$, $\sigma = 1$, $M = 8$, and $L = 500$.

in ξ stabilizes at longer block lengths as expected. Similar results were observed for $\alpha = 1$. Figure 2.8 shows the optimal block length computed using Politis and White method [42]. In this instance, the block length was computed using the same data set as was used for computing the MSE in Figure 2.7. We can see that Politis and White method consistently finds optimal block length in the region of minimum total MSE. Similar results were observed for other values of ξ and α. For all subsequent discussions, we will use the Politis and White method to compute the optimal block length.

In the WLS method, estimation of the shape parameters ξ and α is also critically dependent on the weights given to each lag. In section 2.3, we defined the weights for the WLS method as in (2.3.6). In this context, we discussed two methods of estimating $\mathrm{Var}(2\hat{\gamma}(h))$: the Cressie approximation (2.3.7) and the bootstrap method (2.4.4). Figure 2.9 shows a 200 sample Monte-Carlo comparision of the shape parameters ξ and α. We can see that the Cressie approximation has a larger variance in the estimate of α than the bootstrap method. The bootstrap method for estimating weights has a lower MSE for both ξ and α. Similar results were observed for other values of α and ξ. Going forward, the $\mathrm{Var}(2\hat{\gamma}(h))$ will be estimated using the bootstrap method.

Based on 200 sample Monte-Carlo simulation, in Figure 2.10 we compare the four available estimators of σ. In an ideal scenario (Figure 2.10(a)), when no local systematic variation is present, performance of all estimators is roughly similar. The bias in $\hat{\sigma}_{LWR}$ is clearly visible, and all of the unbiased estimators ($\hat{\sigma}_{BBB}$, $\hat{\sigma}_{WLS}$, and $\hat{\sigma}_{LLE}$) appear to perform equally well. However, if there is any *unknown* local variation in the lines of the SEM image, the

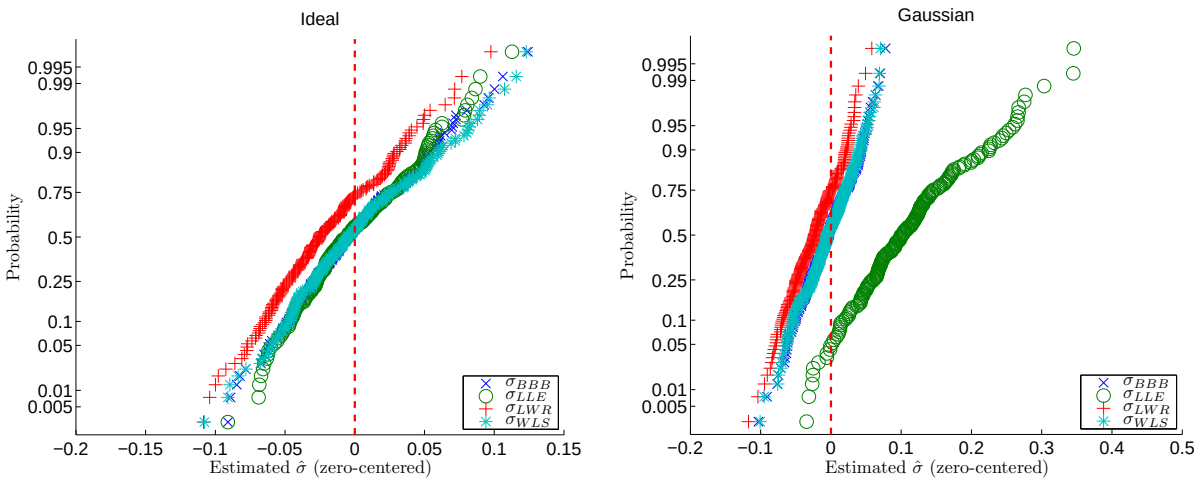


Figure 2.10: Comparison of four different estimates of σ based on 200 sample Monte-Carlo simulation in two scenarios: (a) in absence of any local CD variation ("Ideal"), and (b) in presence of a local variation $\mathcal{N}(0, 0.25)$ ("Gaussian"). The roughness profiles were simulated with $\sigma = 1$, $\xi = 10$, $\alpha = 0.5$, $M = 8$, and $L = 500$. In terms of estimating the σ, $\hat{\sigma}_{WLS}$ and $\hat{\sigma}_{BBB}$ appear to perform equally well, whereas $\hat{\sigma}_{LLE}$ estimate is inflated due to local variation.

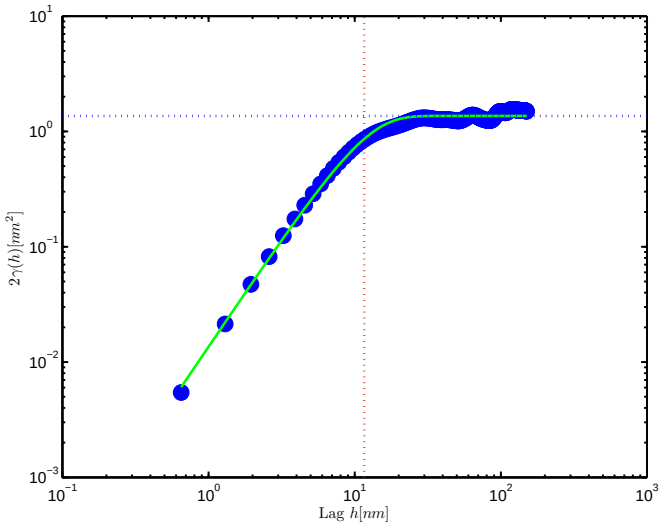


Figure 2.11: Sample fit of variogram using WLS and BBB. The solid line represents the fitted variogram (2.2.10). The dotted horizontal and vertical lines indicate estimated values of $\hat{\sigma}_{WLS}$ and $\hat{\xi}$, respectively.

choice of estimator becomes important. The source of such variation can be from the lithography mask or proximity related. Furthermore, the nature of the variation can be random or systematic. Although process optimization and OPC can minimize such variation, residual variation is unavoidable. In Figure 2.10(b), we explore the possibility of a local Gaussian CD variation. The estimators $\hat{\sigma}_{WLS}$ and $\hat{\sigma}_{BBB}$ perform equally well since they are both bias-corrected values derived from the variogram. The estimator $\hat{\sigma}_{LLE}$ does not perform as well as the estimators $\hat{\sigma}_{WLS}$ and $\hat{\sigma}_{BBB}$. The LLE estimator of σ is sensitive to these unwanted variation due the second term in (2.3.13). The BBB estimator performs robustly under such scenarios.

2.5.3 Experimental Data

In the preceding discussion, we validated the optimal block length and provided a method to estimate WLS weights (i.e. $\mathrm{Var}(2\hat{\gamma}(h))$). We also validated the choice of estimator of σ^2 by minimizing total MSE of LWR parameters using simulated data. Figure 2.11 shows a sample fit of variogram using the WLS and BBB using actual roughness profile. In Figure 2.10, using simulated data, it was shown that the estimate of σ_{LLE} can be corrupted by the presence of some local CD variation. Figure 2.12 compares $\hat{\sigma}_{BBB}$ and $\hat{\sigma}_{LLE}$ for variety of different NGL processes such as *litho-freeze-litho-etch* (LFLE) double patterning lithography (DPL), self-aligned double patterning (SADP), EUV, directed self-assembly (DSA), and nano-imprint lithography (NIL). We observe that in most cases, $\hat{\sigma}_{LLE}$ is significantly higher

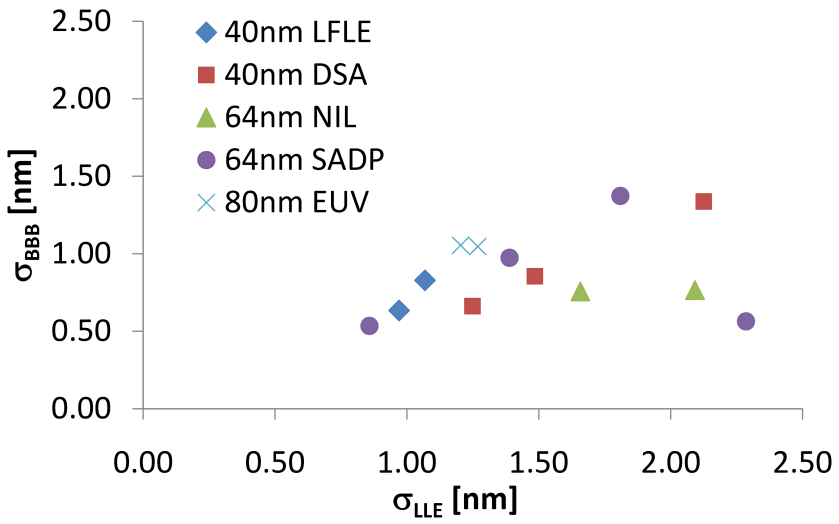


Figure 2.12: Comparison of $\hat{\sigma}_{BBB}$ and $\hat{\sigma}_{LLE}$ for various NGL processes.

that $\hat{\sigma}_{BBB}$. In certain cases (SADP, NIL, DSA), we observed odd and even effects in CD of adjacent lines in SEM image; these effects were unrelated to LWR. In such instances, the $\hat{\sigma}_{LLE}$ estimator attributed the systematic effect to LWR, and it provided an inflated estimate compared to $\hat{\sigma}_{BBB}$. A brief discussion on each of the NGL processes as well as their detailed results will be presented in the next chapter.

2.6 Summary

In summary, we presented a robust method to estimate LWR parameters—$\hat{\sigma}$, $\hat{\xi}$, and $\hat{\alpha}$. Our procedure is shown to perform robustly in practical scenarios with limited data. The proposed method is stable in the presence of limited data without the central assumption of a Gaussian process. Moreover, our procedure works even in the presence of some unknown local CD variation or if there is a systematic difference in CD (by design or otherwise) between the lines. This aspect of our procedure, (a) prevents non-LER sources of variation from being attributed to LER, and (b) it allows for more flexibility in capturing SEM images in that we do not need a special test structure with all lines with same designed CD. The latter aspect allows one to use any IC layout region with straight lines and arbitrary CDs (as opposed to test structures in the kerf area) for LWR parameter extraction. Using experimental data from a variety of next-generation lithography processes, it was shown that σ can erroneously be over-estimated by over $2X$ if local variation is not treated properly.

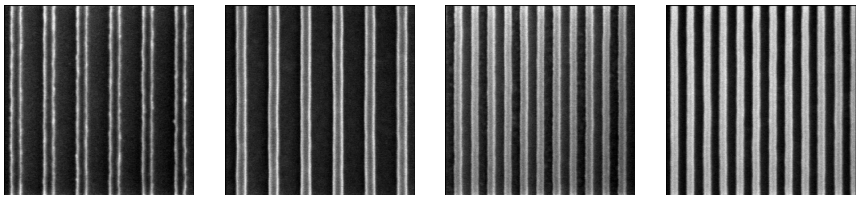
Chapter 3

Line Width Roughness (LWR) in Next-Generation Lithography (NGL) Processes

3.1 Introduction

In the previous chapter we introduced line width roughness (LWR) parameters, and provided physical interpretation to each parameter. We also developed a robust method to estimate the LWR parameters. In this chapter, our objective is to use the newly developed estimation method to explore the LWR characteristics of many next-generation lithography (NGL) processes. We investigate mainstream lithography options such as *litho-freeze-litho-etch* double patterning (DPL), self-aligned double patterning (SADP), and extreme ultra-violet (EUV), as well as alternatives such as directed self-assembly (DSA) and nano-imprint lithography (NIL). Using EUV as an example, we will demonstrate the importance of process optimization.

Although LWR is an intrinsic phenomenon to resist processing, roughness is transferred to the underlying layers in subsequent process steps. It has been recognized that during the transfer process, the characteristics of the roughness, namely the RMS amplitude and frequency are altered [43, 44]. Amongst the NGL options considered, the exact transfer mechanisms for LWR differ widely. It is, therefore, also intriguing to compare these processes in terms of LWR characteristics at the initial, intermediate, and final stages. The rest of the chapter is organized as follows: Detailed discussion on image analysis and estimation methodology can be found in section 3.2. Each NGL technology is briefly described in section 3.3. Our results are discussed in section 3.4.



(a) (b) (c) (d)

Figure 3.1: Representative SEM images for 64nm pitch SADP: (a) Resist, (b) APF Mandrel, (c) Post-spacer definition, and (d) Nitride substrate.

3.2 Estimation of LWR Parameters

There are two widely used tools available for LWR metrology: scanning electron microscope (SEM) and atomic force microscope (AFM). AFM measurements are made by using a specialized probe tip to physically scan the feature sidewalls. SEM measurements are made by rastering an electron beam to recreate a high-resolution top-down image of the feature lines. A detailed comparison between the two techniques has shown that LER quantification through SEM image analysis is a reliable method [17, 45]. Estimation of LWR parameters through SEM image analysis is commonly referred to as *off-line analysis*. Off-line estimation of LWR parameters can broadly be described as a two-step process: (1) SEM image acquistion and processing, and (2) estimation of LWR parameters from the processed SEM image.

3.2.1 SEM Image Acquisition and Processing

Standard line and space arrays were used as image targets for this work. Automated CD-SEM tools were employed for acquiring images for LFLE, EUV, SADP, and NIL. Fields of view with edge lengths ranging from 0.7 to 0.8 μm were used, with corresponding pixel edge lengths ranging from 1.36 to 1.66 nm. Typical accelerating voltages between 500 and 800 V were employed. Integration times were adjusted to produce images with SNR ranging from 10-15 [46]. A Hitachi S4800 analytical SEM was employed for some EUV resist, NIL, and DSA image acquisition. A 1.2 μm horizontal field of view was used, with corresponding pixel edge length of 0.94 nm. Images were captured at 2 mm working distance with 2 kV nominal accelerating voltage. 128 frame captures yielded uncompressed TIFF files with SNR ranging from 15-20. Figure 3.1 shows sample images for SADP process flow.

Image quality poses numerous practical challenges for analysis of LWR. Ideally, these challenges may be addressed by acquiring images for all patterns being compared using a single SEM with stable and known imaging performance. In this study, features are produced by a large variety of patterning techniques that are at varying stages of process maturity. As such, a single SEM approach proved untenable. Nevertheless, we present one test case

(64nm pitch NIL) in which, with acquisition and analysis optimization, consistent results are obtained using an analytical SEM and an automated CD-SEM. We follow the previously described guidelines for optimal image acquisition and processing [45, 47, 46, 48].

Following the acquisition, the images were processed through the SuMMIT software package for line edge extraction [49]. Numerous systematic distortions are typical in CD-SEM images. Accordingly, we pre-screened images for LWR characterization and eliminated images with uncorrectable edge artifacts, tilt, intensity skew, defocus, feature charging distortions, etc.. Correctable global image curvature and tilt were removed prior to LWR analysis. We pre-filtered images to further enhance SNR using a pseudo-Lorentzian filter adapted from resist diffusion studies [50]. Similar image filters are used in image reconstruction to enhance edge contrast while attenuating white noise without causing ringing artifacts [51]. Metrology noise was removed for spatial frequencies beyond the Nyquist frequency. Optimally pre-filtered CD-SEM images had SNR of approximately 20, and pre-filtered analytical SEM images had SNR of approximately 30. Edges were defined at line edge intensity threshold of 50% using a linear intensity edge fit interpolation algorithm [49]. After image selection and pre-filter optimization, a sufficient number (8 to 15) of images were acquired for each process step.

3.2.2 Estimation Procedure

The line edge extraction procedure described above transforms the SEM image into a numeric array of roughness data. The data is available in the same framework as the one previously described in Figure 2.4. LWR parameters were estimated using the robust estimation method developed in the previous chapter.

3.3 Next-Generation Lithography (NGL) Processes

LWR has been recognized as a major challenge to be overcome for all NGL options [15]. It is understood that a number of factors affect LER—photomask, ILS, resist composition, etch process, etc., are just a few such factors. Sidewall protrusions of polymer aggregates have been shown to block the etch [52]. This causes the roughness from the resist pattern is transferred to the underlying substrate during the etch process. In the following discussion, we will discuss each NGL technology in detail from the perspective of LWR. As we shall see, many of the NGL options go through a number of intermediate processing steps before the final pattern in the substrate is defined. LWR characteristics have been reported to change through these processing steps [43, 44].

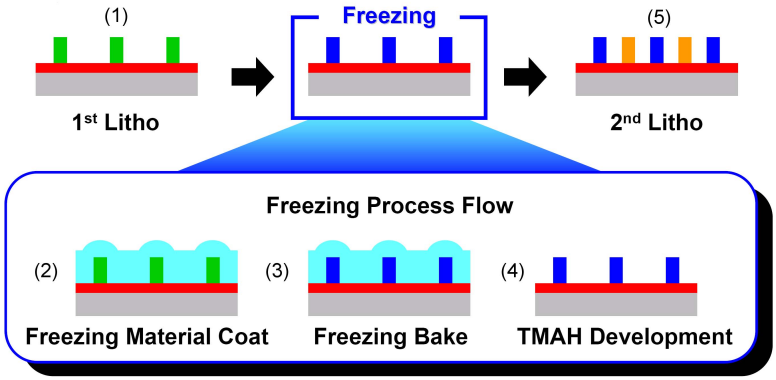


Figure 3.2: *Litho-freeze-litho-etch* (LFLE) process flow using chemical freezing agent. (after [22])

3.3.1 Double Patterning Lithography (DPL)

Generally speaking, DPL comprises two lithographic processes in which every alternate line belongs to the same lithographic process. There are several variations of DPL documented in the literature [15]. For discussion purposes here, we broadly categorize them in two types—processes with two coupled patterning and etch steps (*litho-etch-litho-etch*; LELE), and processes with two patterning steps that are transferred in a single etch (*litho-freeze-litho-etch*; LFLE). For our study of DPL, we have chosen the LFLE process.

Due to the high overall LELE cost, methods that simplify integration of lithographic and etch processes are of growing importance. Efforts in this area center on the LFLE processes in which independent lithographic patterns are generated in sequentially applied resist films prior to etch transfer. In the LFLE process flow, a stabilization step that *freezes* the first photoresist pattern, is required to protect the first pattern from being damaged during the second resist pattern. Implementation of the LFLE as a successful NGL for semiconductor manufacturing is dependent on demonstration of superior cost effectiveness and layout compatibility. In this study, we present pre- and post-etch LWR analysis of a 40 nm final pitch LFLE process that employs a stabilizing polymeric coating to freeze the first resist pattern [22]. Figure 3.2 shows a graphical illustration of the main steps of this flow. After the first lithography step, the wafer is coated with a freezing agent composed of resin and cross-linker, along with the appropriate casting solvent. After the unreacted freezing agent is developed away, the second lithographic step is performed. Materials to support this LFLE method have been developed by JSR Micro [22], and more recently by Dow Chemical as well [53].

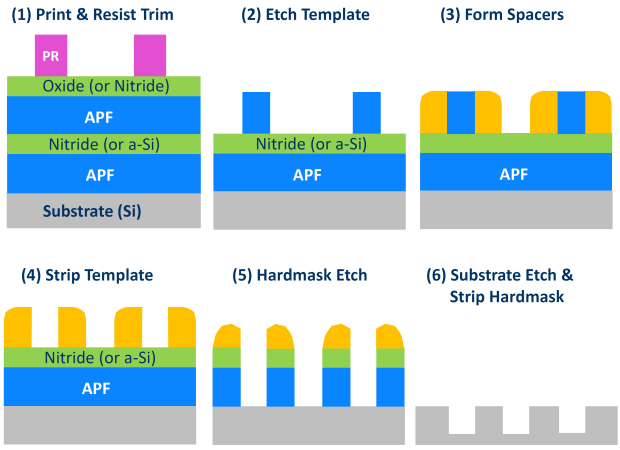


Figure 3.3: Self-aligned Double-Patterning process flow. (*Courtesy: Chris Bencher, Applied Materials*)

3.3.2 Self-Aligned Double Patterning Lithography (SADP)

The SADP process uses deposition of conformal coatings to generate frequency doubled features at the sidewalls of pre-patterned features. Process integration plays a key role in determining the success of SADP: although SADP eases the lithographic burden associated with sub-80 nm patterning, much of the burden is transferred to increased film stack and deposition complexity.

Here, we present post-lithographic and through-etch LWR analysis for a 64 nm pitch baseline SADP process as implemented by Applied Materials [54]. LWR is analyzed at several intermediate stages in the etch transfer process. Figure 3.3 provides a graphical illustration of the main steps in the SADP process flow. For our study, we acquired images at steps (1), (2), (4), and (6). We note that the SADP process creates spacer patterns that are fundamentally different in their LWR behavior compared to patterns created using double patterning and EUV. The difference arises from the conformal spacer deposition which creates highly correlated spacer edges. As a result, patterning tone becomes an important consideration. If spacers are used to define the final product features, highly correlated, low LWR features will result. Alternatively, a tone inversion, in which trenches in between spacers are used to define the final product features, will produce features with more conventional uncorrelated edges. Coupled with integration and CD uniformity considerations, this unique LWR behavior of SADP is important for choice of patterning tone, particularly for BEOL applications.

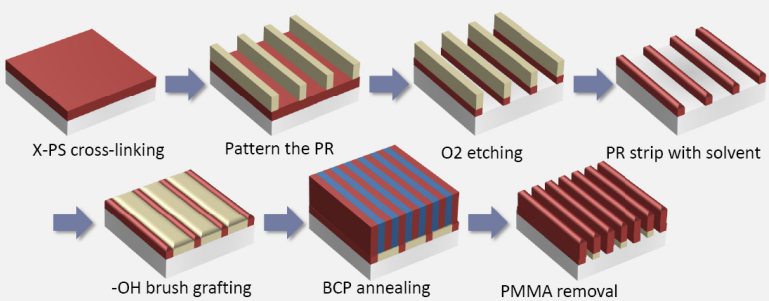


Figure 3.4: Directed Self-Assembly process flow. (*Courtesy: Chi-Chun Liu, Adam Welander and Paul Nealey, Univ. of Wisconsin*)

3.3.3 Extreme Ultraviolet Lithography (EUV)

EUV lithography continues to promise a relatively constraint-free single patterning solution for pitches below the 193 nm immersion limit. As a result, process flows are very conventional in comparison with other NGL candidates. Overall high infrastructure development cost, lithographic tool cost, and technical issues centering on reticle defectivity, source power and stability, and limits of photoresist performance, continue to pose challenges for adoption in high volume manufacturing [15].

In this study, we present pre- and post-etch LWR analysis for 80 nm pitch baseline EUV process implemented using the full-field 0.25 NA ASML alpha-demo tool located in Albany, NY. Additionally, we present post-lithography LWR analysis of 64 nm pitch resist processes implemented on both the ASML alpha-demo tool and the LBNL/SEMATECH 0.3 NA micro-exposure tool (MET).

3.3.4 Directed Self-Assembly Lithography (DSA)

The DSA process used in this study is described in detail elsewhere [55]. We briefly describe the DSA process here for completeness. Figure 3.4 provides a graphical illustration of the main steps in SADP process flow. DSA employs a thin blend of immiscible diblock copolymer (e.g. PMMA and polystyrene) that, upon coating and annealing, forms spontaneously ordered and thermodynamically stable arrays of chemically distinct polymer domains. The microphase separation occurs as a result of surface and interfacial forces between the substrate surface and the two blocks of the copolymer. This process uses "conventional lithography" (EUV in the case presented here) to define a template or pre-pattern for forming lamellae that are oriented perpendicular to the substrate and aligned along predefined grooves. Subsequent domain separation occurs perpendicular to the grooves and results in pitch-splitting. A wide variety of array structures can be created at defined pitches with judicious choice of pre-pattern, block copolymer composition and molecular weight, and film thickness. One

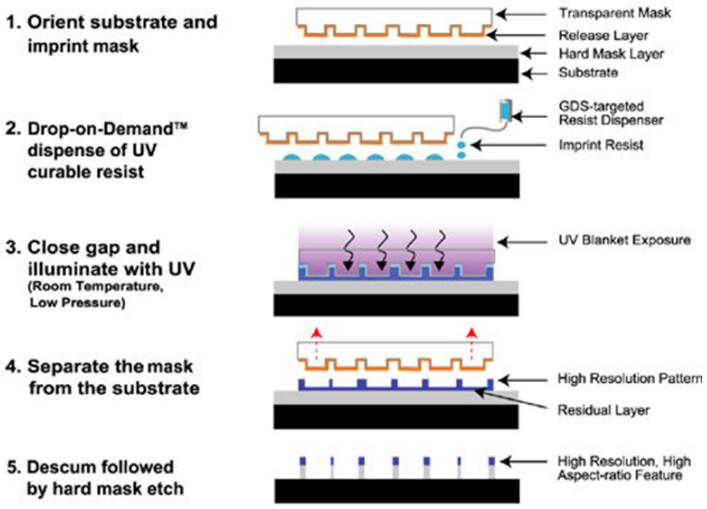


Figure 3.5: Nano-imprint lithography S-FIL^{TM} process flow. (*Source: Molecular Imprints Inc.*)

key feature of DSA is that high-fidelity self assembly is possible even if "sparse" pre-patterns are used. This allows the creation of frequency multiplied (double, triple, or even quadruple) DSA patterns relative to the pitch of the pre-pattern. Pre-patterns may be created using relief structures such as edges etched in a substrate or using thin films with chemical affinity for one of the block copolymer components. Recent progress in DSA patterning has resulted in growing interest in commercial applications such as patterned media.

Implementation of DSA as a successful NGL for semiconductor manufacturing is dependent on successful resolution of critical technical challenges including all aspects of process integration and maturity, as well as layout compatibility with geometric restrictions imposed by DSA. In this study, we present LWR analysis of DSA patterns at 40 nm pitch using poly(styrene)-b-poly(methyl methacrylate) diblock copolymer films, and track LWR through multiple process steps including pre-patterning at 80 nm pitch, self-assembly, and etch transfer into a silicon substrate.

3.3.5 Nano-imprint Lithography (NIL)

NIL has been considered a viable alternative to optical lithography [15]. In this study, we evaluate *step and flash imprint lithography* or S-FIL^{TM} process reported earlier at the 2008 SPIE for sub-64nm full pitch node [56]. Figure 3.4 provides a graphical illustration of the main steps in S-FIL^{TM} process flow. In NIL, a silica mask with patterned trenches is used to physically imprint the desired pattern into a thin layer of low viscosity resist. The resist is physically squeezed out of regions with clear pattern in the template. Subsequently, the

resist is exposed to UV light while it is embedded inside the template, and then the mask is removed. Upon removal of the mask, a thin residual layer of the imprint resist remains underneath the imprinted pattern. The thickness of this residual layer is controlled by the imprint process, and it is removed by a short trim etch. Typically, the resist layer is too thin to be of any practical use, and therefore, the pattern is then transferred to the underlying hardmask layer.

In contrast to optical lithography, NIL patterns are physically transferred to resist using a template or mold. The S-FIL process makes use of photo-polymerization of a low-viscosity resist precursor fluid inside the template. The template is created by using an anisotropic plasma etch to transfer a pattern created using electon-beam lithography. We note that NIL does not employ pattern reduction in either generation of the template or substrate patterning (i.e imprint mask is 1× demagnification). Since no frequency filtering due to a reduction optic occurs, the photo-polymerized substrate pattern approaches a perfect inverted tone replica of the template. Also, when the resist is exposed inside the template, unlike conventional optical lithography, there is no Gaussian diffusion of acid molecules across the pattern edge. The template pattern in turn is expected to retain all characteristics, including roughness, of the lithographic and etch transfer steps used in its creation.

NIL is finding commercial acceptance in creation of patterned media for hard disk storage as well as other applications. Implementation of NIL as a successful NGL for semiconductor manufacturing is dependent on successful resolution of critical technical challenges including overlay, template cleaning, defectivity, and throughput. Overlay limitations specifically have slowed the successful implementation of mix-and-match technology demonstrations. As a result, access to integration process flow data including etch transfer is limited. Here, we present LWR analysis of NIL resist patterns at 64 nm, but do not include results from subsequent etch transfer steps. NIL patterns studied here were created using a Molecular Imprints Imprio300^{TM} tool located at SEMATECH in Albany, NY. All materials and process conditions were standard for the Imprio^{TM} tool.

An outgrowth of these considerations is of practical concern for LWR analysis: repeat steppings across a wafer produce NIL patterns that are deterministic from stepping to stepping. Image sampling across a wafer must be implemented so as not to repeatedly capture features that were created by the identical template site. In this study, we have made use of repeated pattern macros present on the NIL template to generate sets of images in which each image represents a unique template site.

3.4 Results and Discussion

In our evaluation of various technology options, we took substantial measures to ensure that we did not introduce any undue bias in the results. The SEM images for each technology were consistently acquired and processed using the methodology discussed in section 3.2. In our reporting of LWR parameters, we adopt the following labeling convention: <*process*>–

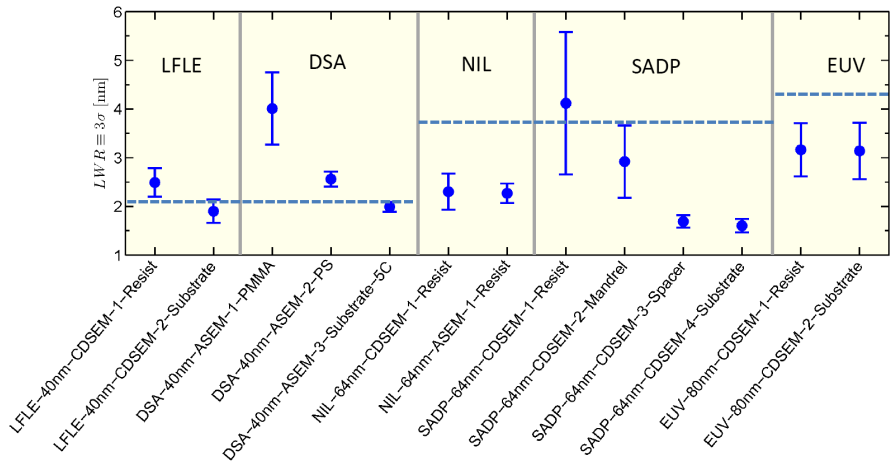


Figure 3.6: LWR (3σ) in nm for various NGL processes. The dotted lines indicate ITRS roadmap values [15].

<*pitch*>–<*type of SEM used for acquisition*>–<*process sequence*>. Automated CD-SEM and analytical SEM are abbreviated as CDSEM and ASEM respectively. With the exception of NIL, we monitored the evolution of LER at all intermediate process steps from resist processing to final substrate definition. With the exception of LFLE, 8 to 15 images were captured for each process step at constant process condition (dose, focus, etc.). In case of LFLE, minor differences in LWR between first and second lithography have been reported [22]. In our study, we did not find these differences to be statistically significant, and therefore, all LFLE LWR parameters reported below represent an aggregate value of the two lithography steps.

Conventionally, the RMS amplitude of LWR (σ) is reported as 3σ figure of merit, and it is commonly referred to as LWR. And as such, we use the same convention to report our results. Figure 3.6 shows LWR (in nm) for LFLE, DSA, EUV, NIL, and SADP processes at a variety of full pitch values. The general trend in each technology is LWR attenuation from resist to etch. Also shown in Figure 3.6 are ITRS roadmap values corresponding to *full pitch* and not *technology nodes* [15]. Figure 3.6 also indicates that the difference in LWR due to different SEM tools (but same wafers) is statistically insignificant. The processes reported here come from widely disparate stages in development—although some versions are implemented in mass production others show very promising results in the development phase. Given the fact that each technology is reported for different pitch values, we present normalized LWR results in Figure 3.7. ITRS roadmap stipulates LWR values $< 8\%$ at resist [15]. At resist, we observe that NIL and EUV processes are the only two that meet this criteria. However, note that each NGL process option has $< 10\%$ *final* LWR contribution; SADP has $\sim 5\%$ LWR.

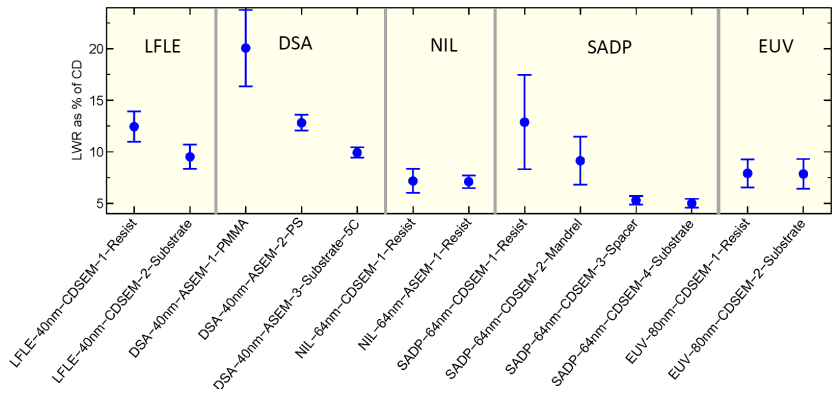


Figure 3.7: Normalized LWR (3σ as % of CD) for various NGL processes.

Correlation length was recently added as a figure of merit to the ITRS roadmap [15]. Figure 3.8 compares the correlation lengths observed for different NGL processes with ITRS roadmap values corresponding to *full pitch* and not to *technology nodes*. The correlation lengths range from approximately 8 to 24 nm, and it meets the requirements set forth by the ITRS. Note that the correlation length increases when LER is transferred through an etch of an inorganic film. This finding is consistent with an earlier report of increased correlation length [44]. For DSA, the correlation length was found to decrease from PMMA to polystyrene (PS) lines. For SADP, the change in correlation length is a bit peculiar. It is interesting to note the increase in ξ is very stark for SADP, where it approximately doubled from resist to mandrel etch, and also from spacer to substrate etch. The decrease in correlation length from mandrel to spacer requires further investigation; each spacer defined line inherits characteristics of one edge from the mandrel and the other from the conformal spacer deposition and anisotropic etch.

Figure 3.9 shows the roughness exponent for a variety of different NGL processes. As mentioned previously, α values close to 1 indicate a locally smooth roughness profile whose autocorrelation function has a Gaussian shape. Amongst the various processes considered here, DSA produces the most locally smooth profile. The DSA profile seems to undergo a transformation during the self-assembly phase of the diblock copolymer. The block copolymer assembly process has been reported to self-heal irregularities, because the thermodynamics of the block copolymer system determine the overall shape and size of the domains [55]. Also, the conformal spacer deposition brings about a dramatic change in α for SADP; a significantly smooth profile is found after spacer formation compared to that after Mandrel etch. However, when the spacer lines are used to define the substrate, the value of α drops significantly. This behavior needs to be investigated further. For LFLE, EUV, and SADP (resist to mandrel), although there appears to be a slight reduction in value of α from resist to etch, the change has not been found to be statistically significant.

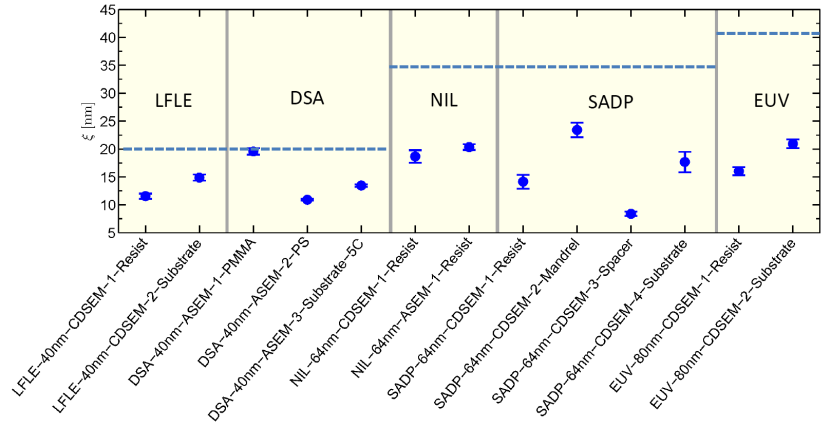


Figure 3.8: Correlation length (ξ) for various NGL processes. The dotted lines indicate ITRS roadmap values [15].

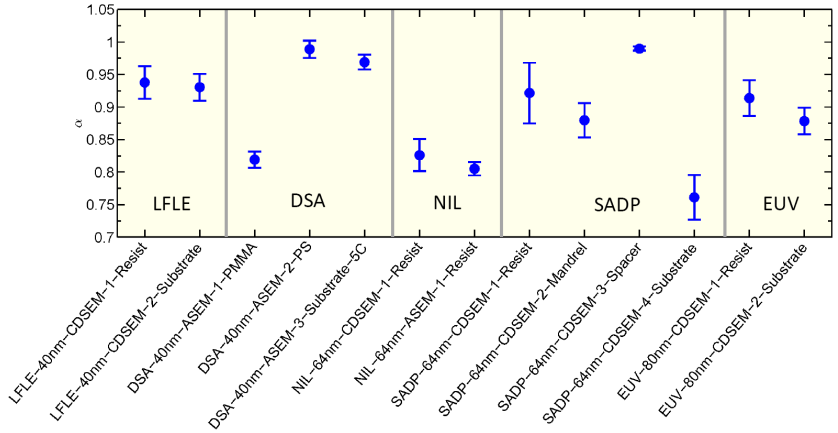


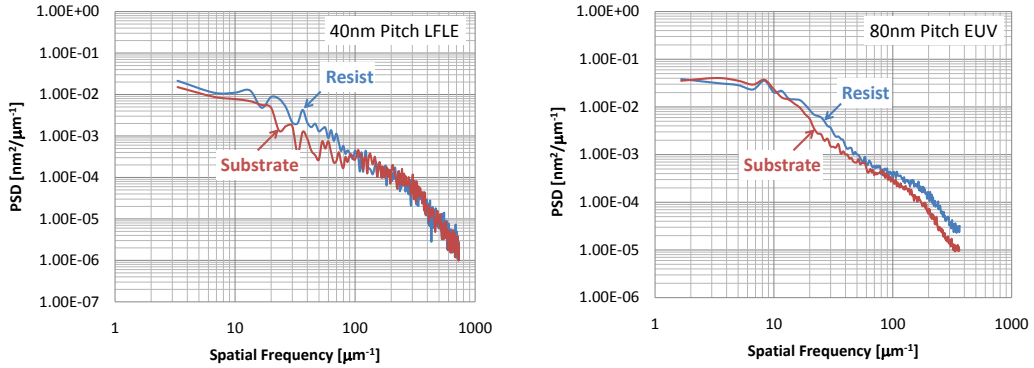
Figure 3.9: Roughness exponent (α) for various NGL processes. The ideal value for α is 1.

Full Pitch [nm]	NGL Process	LWR % of CD (Std. Error)		LWR [3σ, nm] (Std. Error)		Correlation Length [ξ, nm] (Std. Error)		Roughness Exponent [α, nm] (Std. Error)	
		Resist	Final	Resist	Final	Resist	Final	Resist	Final
40	LFLE	12.4 (0.7)	9.5 (0.6)	2.5 (0.1)	1.9 (0.1)	12 (0.2)	15 (0.3)	0.94 (0.01)	0.93 (0.01)
40	DSA	20.1 (1.9)	9.9 (0.3)	4 (0.4)	2 (0.1)	19.5 (0.3)	13.4 (0.1)	0.82 (0.01)	0.97 (0.01)
64	NIL	7.2 (0.6)	N/A	2.3 (0.2)	N/A	19 (0.6)	N/A	0.83 (0.01)	N/A
64	SADP	12.9 (2.3)	5 (0.2)	4.1 (0.7)	1.6 (0.1)	14.2 (0.6)	17.7 (0.9)	0.92 (0.02)	0.76 (0.02)
80	EUV	7.9 (0.7)	7.9 (0.7)	3.2 (0.3)	3.2 (0.3)	16 (0.4)	20.9 (0.4)	0.91 (0.01)	0.88 (0.01)

Table 3.1: Summary of LWR parameters in resist and final substrate for NGL technologies considered in this work. The numbers have been rounded to the nearest meaningful decimal place.

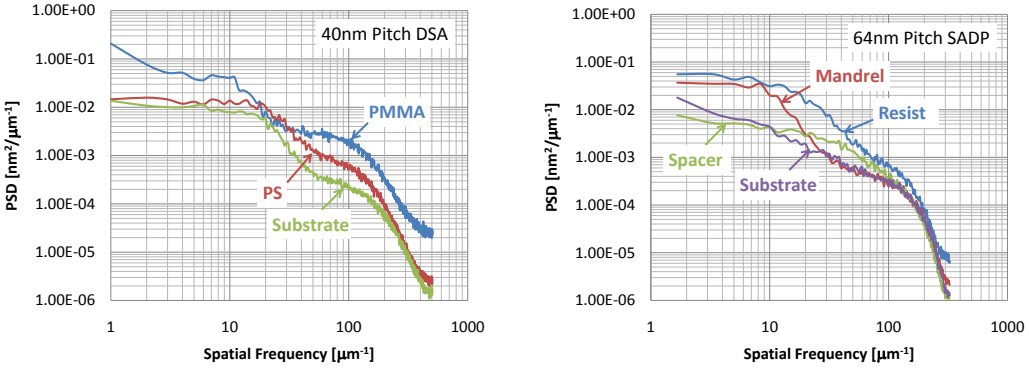
Numerical values of the LWR parameters for each of the NGL technologies is shown in Table 3.1. LER is a dynamic phenomenon; its characteristics evolve with processing steps. Although conventionally the parameters σ, ξ, and α provide a complete description of LWR, it is interesting to observe the changes in spatial frequency content of LWR with the progressive processing from resist to final substrate etch. Figure 3.10 shows the power spectral density plot for LFLE and EUV. The suppression of mid-frequencies (10-100 μm^{-1}) for LFLE and EUV following etch is consistent with earlier findings for conventional lithography [43, 44]. Figure 3.11 shows the power spectral density plot for DSA and SADP. To the best of our knowledge, this is the first time PSDs of DSA and SADP have been reported. In the case of DSA, we observed some low frequency meandering of lines as well as defects in the X-PS underlayer. There is clear suppression of power in the entire frequency range for DSA. For SADP, typical mid-range (10-100 μm^{-1}) frequency suppression can be seen for resist to mandrel and spacer to substrate transfers. Following the spacer formation, there is substantial reduction of power in the low frequency regime.

There are many factors that affect LWR such as lithography (resist type/thickness, focus/exposure, ILS, etc.), etch (wafer temperature, power, chemistry, pressure, etc.), lithography system (ILS, shot noise, mask LER, etc.), and process integration scheme. Significant reduction in LWR can be achieved by careful process optimization. Figure 3.12 shows the impact of process optimization on LWR for EUV. Experiments were conducted on 64nm pitch EUV to optimize the resist chemistry, and underlayer type and its thickness. Post-develop smoothing techniques were also employed. The learning was applied to 80nm pitch EUV. The improvements achieved with these process factors appear to be additive in nature.



(a) (b)

Figure 3.10: Power Spectral Density (PSD) for (a) 40nm pitch LFLE, and (b) 80nm pitch EUV.



(a) (b)

Figure 3.11: Power Spectral Density (PSD) for (a) 40nm pitch DSA, and (b) 64nm pitch SADP.

Process	% LWR (Std. Error)	LWR 3σ [nm] (Std. Error)	ξ[nm] (Std. Error)	α (Std. Error)
64nm Pitch MET	19.46 – 24.76 (3.26 – 7.46)	6.22 – 7.92 (1.04 – 2.39)	20.15 – 23.73 (0.37 – 0.69)	0.82 - 0.9 (0.009 – 0.013)
80nm Pitch ADT	7.9 (0.68)	3.16 (0.27)	16 (0.36)	0.91 (0.014)

Figure 3.12: Results of process optimization from 64nm pitch EUV (*unoptimized*) to 80nm (*optimized*).

3.5 Summary

In summary, we explored variety of next-generation lithography processes in terms of their LWR characteristics. NIL and EUV processes are the only two that meet the $< 8\%$ LWR criteria at resist; however, note that each NGL process option has $< 10\%$ *final* LWR contribution. Minimizing the roughness in resist is still a prerequisite to ultimately minimizing roughness in the substrate. In terms of the roughness exponent, technologies such as DSA that produce a more locally smooth profile ($\alpha \to 1$), are desirable. Each of the NGL process meets the ITRS roadmap values for correlation length. However, a fundamental understanding of process conditions causing the increase in correlation length during etch is required. Whereas the technologies considered here are promising in terms of their LWR performance, for these technologies to be adopted in mass production, many technical (process integration) and economic challenges must first be overcome.

Chapter 4

Gate Line Edge Roughness Model for Estimation of FinFET Performance

4.1 Introduction

Intrinsic process parameter fluctuations cause undesirable performance mismatch in identically designed transistors. As the dimensions of the transistors are scaled down, this mismatch increases, and hence it has greater impact on circuit performance and yield. The primary sources of transistor performance variability that have emerged are line edge roughness (LER), gate dielectric thickness (t_{ox}) variation, random dopant fluctuations (RDF), and metal-gate work-function (WFV) [14, 57]. Advanced transistor structures such as the double-gate (DG) FinFET [58] are more robust to t_{ox} variation and RDF because a thin body is used to suppress short-channel effects (SCE), without the need for channel/body doping. In a recent study, FinFETs have been found to have lower threshold voltage variability due to line edge roughness [59]. Due to the challenges with scaling planar bulk MOSFETs, advanced structures such as FinFET may be adopted as early as the 25nm CMOS technology node [60].

Earlier work on understanding the effects of the LER on device performance was either focused on planar bulk CMOS [61] or followed a computationally expensive Monte-Carlo (M-C) approach [62]. Due to the stochastic nature of LER, an accurate estimate of device performance variability can only be achieved through full M-C three-dimensional (3-D) device simulation. However, this computational approach is prohibitively expensive, and it does not provide any insight into how the LER impacts device performance. Our premise here is that LER manifests itself in the form of offset between the front-gate (FG) and the back-gate (BG) as well as difference in FG and BG critical dimensions. As such, we believe that the 2D transistor structure is sufficient to capture the effects due to the mismatched FG and BG. Therefore, in this chapter, we develop a computationally efficient statistical model that is formulated to link the characteristic LER descriptors to device performance variability.

The organization of this chapter is as follows: In section 2, we briefly describe how LER and LWR are related. In section 3, we describe the details of the 2-D device simulation and the formulation of our model. Our simulated device structure is designed to meet ITRS specifications for the 32nm high-performance (HP) CMOS technology node. Finally, in section 4, we discuss the results of our work. The impact of gate length variation and lateral offset between the FG and BG is studied. Sensitivity of key performance parameters such as saturation threshold voltage ($V_{t,sat}$), on-state saturation drive current ($I_{d,sat}$), and off-state leakage current (I_{off}) to the various LER parameters is discussed.

4.2 Line Edge Roughness

4.2.1 Background

Line edge roughness (LER) and line width roughness (LWR) are often used synonymously. Mathematically, they are related but different. As shown in Figure 4.1, LER refers to the

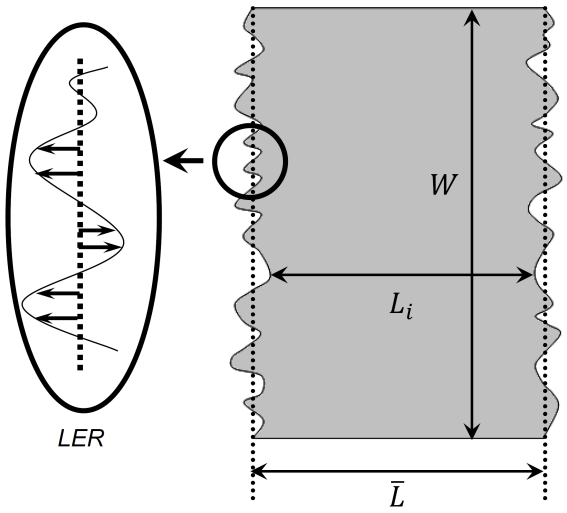


Figure 4.1: Comparison of LER and LWR. Line edge roughness (LER) is the fluctuation of a line about its mean value for a given edge. Line width roughness (LWR) is the fluctuation of line width L_i about its mean value $\bar{L}$ averaged over the width W.

fluctuations of a given line edge about its mean value while LWR corresponds to fluctuations in line width about its own mean value. For a line sampled at N points along the width W,

LWR is described by the variance in line width, as

$$\sigma_{LWR}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(L_i - \bar{L})^2. \tag{4.2.1}$$

LWR can also be described in terms of variability of each individual edges, as

$$\sigma_{LWR}^2 = \sigma_L^2 + \sigma_R^2 - 2\rho_X\sigma_L\sigma_R, \tag{4.2.2}$$

where the subscripts 'L' and 'R' refer to the left and right edges of a line respectively, and ρ_X is the cross-correlation coefficient between them. The value of ρ_X depends primarily on the method of line formation, as described later in this chapter. If we assume that $\sigma_L = \sigma_R \equiv \sigma_{LER}$, then we can simplify (4.2.2) to

$$\sigma_{LWR}^2 = 2\sigma_{LER}^2(1 - \rho_X). \tag{4.2.3}$$

As mentioned previously in chapter 2, LWR can be completely described by three parameters: correlation length (ξ), RMS amplitude or standard deviation (σ) of line edge from its mean value, and roughness exponent (α) [6, 63]. In order to capture that spectral content of roughness along the edge, we invoke the formulation of the auto-correlation function and use the same closed-form expression as (2.2.9)

$$\rho_A(y) = \exp\left[-\left(\frac{y}{\xi}\right)^{2\alpha}\right], \tag{4.2.4}$$

where y is the lag. It should be pointed out that (4.2.4) represents just one form of a plausible auto-correlation function. Other forms such as exponentially decaying sinusoid can also be used [64].

4.2.2 Spacer vs. Resist Lithography

In a FinFET fabrication process, the gate electrode can be defined in one of two ways: using resist as the mask ("resist-defined"), and using a spacer as the mask ("spacer-defined"). Conventional resist-defined lines produce edges with uncorrelated roughness, and so $\rho_X = 0$ can be assumed in (4.2.3). This is due to the fact that erosion of polymer aggregates is a random process for each resist edge. In contrast, spacer-defined lines have line edges that are well correlated. This is because a spacer mask is formed along the sidewall of a dummy resist-defined feature, via a conformal thin-film deposition process followed by highly uniform anisotropic etch process (Figure 4.2). If the spacer width (corresponding to the thickness of the deposited film) is much smaller than the inverse of the LWR spatial cutoff frequency, spacer-defined lines will have uniform width, and so $\rho_X = 1$ can be assumed in (4.2.3).

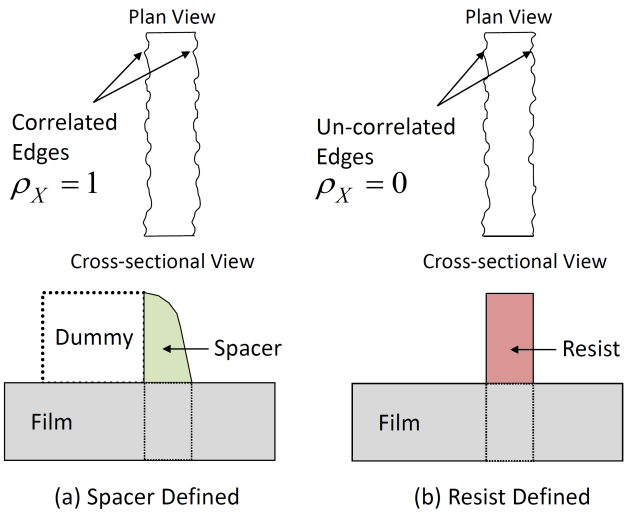


Figure 4.2: Illustration of methods of defining gates with (a) identical (and therefore correlated) edges, and (b) uncorrelated edges.

It should be noted that resist pattern transfer to an underlying layer acts a low-pass filter [65], so that LWR of a patterned film will have reduced high-spatial-frequency components as compared to the resist that was used to define it. For a bulk MOSFET structure, gate LWR affects device performance, because the gate length (L_g) is modulated along the width of the channel. Several approaches to modeling this effect have been reported in the literature; slice approximation [66] and full 3-D device simulation [61, 62, 67] are the most commonly used approaches. In the slice approximation approach, gate LWR is approximated by regularly sampling L_g along the width of the channel, and modeling the transistor as a parallel combination of individual transistors with channel width equal to the sampling interval and L_g values corresponding to the sampled values. (Gate LWR is zero for each individual transistor.) This approach can yield reasonably accurate estimations of performance parameters for planar bulk MOSFETs. Unfortunately, it is not applicable to the FinFET structure, because the channel length (along the fin sidewalls) is not impacted by gate LWR in the same manner.

4.3 Simulation Details and Model Formulation

4.3.1 FinFET Structure

A FinFET can be formed in a straightforward manner by first patterning a silicon-on-insulator (SOI) layer of thickness h_{fin} into a narrow fin of width t_{fin} and height h_{fin}. After

the gate stack layers are grown or deposited, either resist or a spacer is used to define the gate electrode that crosses over the active area (i.e. the fin). After the gate layer is etched using the resist or spacer mask, the resultant gate electrode straddles the fin, to gate the channels along the front and back fin sidewalls. Thus, the fin height h_{fin} determines the effective width of both the front and back channels of the transistor. Figure 4.3 shows how

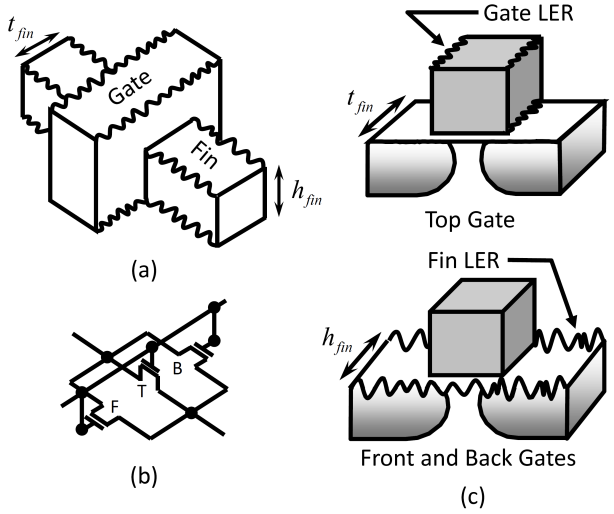


Figure 4.3: Illustration of fin LER and gate LER components in a tri-gate FinFET. The magnitude of LER is exaggerated here for illustration purpose. (a) FinFET with LER, (b) electrical diagram showing three transistors, and (c) bulk CMOS equivalent component transistors are shown separately to distinguish the difference in effects of the two LER components.

LWR affects both the fin and gate in a FinFET structure. If a thin gate dielectric (rather than a thick dielectric hard mask) exists between the gate and the top surface of the fin, a channel can also be formed along the top surface of the fin. In this case, the FinFET may be considered as a parallel combination of three field effect transistors (FET) with channels along the front, back, and top fin surfaces. The top FET has a smooth channel surface, but has non-uniform L_g due to gate LWR. In contrast, the front FET and back FET have a rough channel surface due to fin LWR, but relatively uniform L_g (dependent on the gate-etch process). Fin-sidewall roughness can significantly degrade carrier mobility due to surface scattering. Fortunately, the sidewall surfaces and fin corners can be smoothed prior to gate-stack formation by a suitable thermal anneal to improve carrier mobility, reduce gate leakage current, and improve device reliability [68, 69]. Additionally, it has been shown that fin LWR primarily affects the device performance by changing the average fin width in the channel region [62]. Thus, in this work, we focus primarily on gate LWR. The fin width must

be smaller than the effective channel length in order to suppress short channel effects (SCE) without the need for heavy fin/body doping. Indeed, light fin/body doping is desirable to minimize variability due to RDF effects. In this case, the volume of the fin is inverted when the FinFET is turned on [70] so that current flows in the body of the fin, rather than at the fin surfaces. Consequently, gating of the top fin surface (i.e. the top FET) contributes negligibly to off-state leakage and on-state drive current [71]. Therefore, in this study, we focus only on DG FinFET performance. Figure 4.4 illustrates how gate LWR can result in

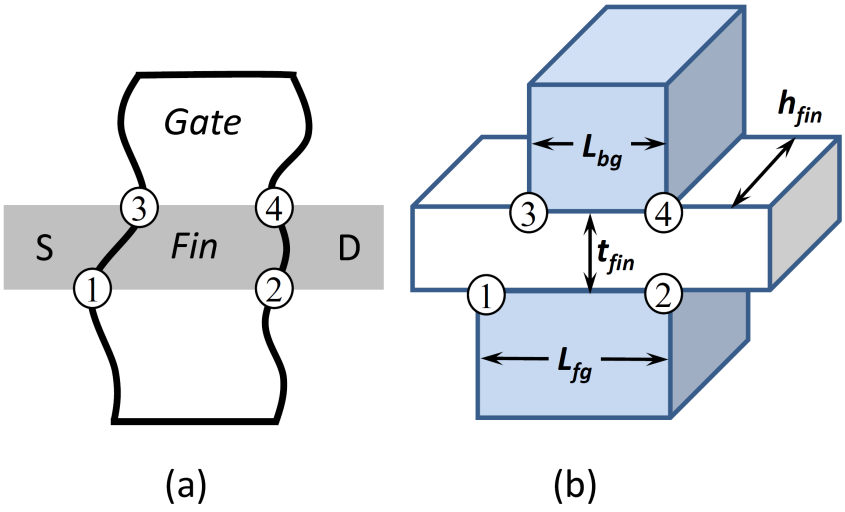


Figure 4.4: Schematic views of a DG FinFET: (a) top view of a FinFET, (b) 3-D top view of a FinFET illustrating various gate-electrode features. The FG and BG gate lengths, and their placements, are defined by the points labeled 1-4.

different L_g values and misalignment between the FG and BG. Gate length values for the FG and BG (L_{fg} and L_bg, respectively) are determined by "sampling" the auto-correlated LWR function along each edge of the gate electrode at the front and back surfaces of the fin; thus the locations of points 1-4 are affected by the fin width, since it determines the sampling distance. Although the primary criterion for the choice of fin width (t_{fin}) is SCE control, mitigation of gate LWR effects to reduce variability may be an important secondary consideration. As discussed earlier, spacer-defined lines have highly correlated edges so that gate-length variations are negligible if spacer lithography is used to pattern the gate electrode. Nevertheless, the FG and BG can be misaligned. Thus, it is important to also study the case where the FG and BG have the same gate length, but are offset by some distance. If a highly anisotropic and uniform etch is used to form the gate electrode, the locations of points 1-4 (as determined by gate LWR and fin width) are transferred uniformly from the top of the fin to the bottom of the fin. In reality, the etch bias can vary from the top of the fin

Electrical/Doping	Structural
V_{dd}=0.9V	L_g=13nm
ϕ_m=4.62eV	t_{ox}=6A
N_B=1e15 cm^{-3}	L_{sp}=7.2nm
$N_{s/d}$=1e20 cm^{-3}	t_{fin}=7.5nm
$\sigma_{s/d}$=4nm/dec	t_{poly}=13nm

Table 4.1: 2-D Device Simulation Parameters

to the bottom of the fin, resulting in a tapered profile. The gate sidewall along the fin height may have a rough profile, and it has been shown that this behavior is adequately modeled by using gate length that has been averaged along the height of the fin [62]. Moreover, the fin itself can have a tapered profile; this has been studied by other researchers [72, 73]. These aforementioned non-idealities of gate and fin profile are not considered in this work.

4.3.2 Simulation Details

Table 4.1 lists the values of process and device parameters that were used, generally following ITRS HP 32nm node specifications. Figure 4.5 shows the simulated 2-D device structure

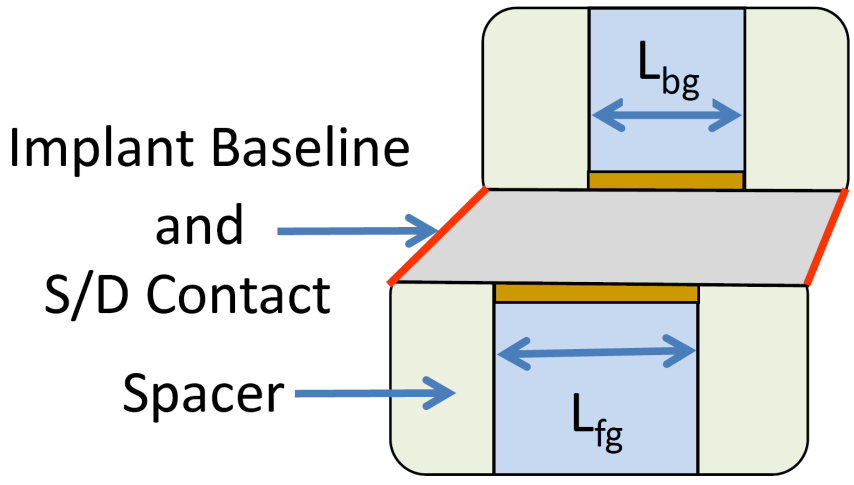


Figure 4.5: 2-D simulated device cross-section of DG-FET structure shown with non-idealities (misaligned and with gate length difference between FG and BG).

with FG and BG non-idealities. Source and drain doping profiles are Gaussian, peaked at the edges of the gate-sidewall spacers (defined by the implant baseline in Figure 4.5), and assumed to have lateral S/D doping gradient S/D=4nm/dec [74]. This implant profile produces a gate-underlapped source/drain structure, which has been found to be optimal for the sub-20nm physical L_g regime [75]. Assuming inversion carrier density of 1×19 cm^{-3}, the effective gate length of the nominal device is 23.4nm. Ideal metallic contacts are made to the surfaces of the uniformly doped S/D regions. All simulations were performed using the Sentaurus device simulator [76], with coupled Poisson, quantum, and high-field saturation models. In hydrodynamic (HD) simulations, the carrier velocity is assumed to depend on the local carrier temperature, and in near ballistic regime, it tends to overestimate velocity overshoot and drain current. In a study performed by Granzer et al. [77], it was found that for 20nm gate length double gate devices, the on-current and sub-threshold leakage current from hydrodynamic simulation were both overestimated by 80% compared to Monte-Carlo simulation. In order to accurately relate simulation data to experimentally determined values of on-current and sub-threshold leakage current, one would be required to carefully calibrate the HD model parameters such as the energy relaxation time (among other parameters). Nayfeh et al. [78] calibrated hydrodynamic parameters using full-band Monte-Carlo simulation. In our simulation, we used the energy relaxation time (τ_E) of 0.14ps and energy flux parameter (r_n) of 0.3 [74].

4.3.3 Model Formulation

First, we formulate a simple statistical model to describe the variability in geometrical parameters in terms of characteristic LWR descriptors. Consider the model parameters illustrated in Figure 4.6. Using point u_2 as the reference, we need to describe the relationship of points u_1, u_3, and u_4 in terms of the characteristic LWR descriptors. Misalignment between the FG and BG can occur due to presence of an offset (between the points u_1 and u_3 and/or between the points u_2 and u_4), with or without a difference in the gate critical dimension of the FG and BG. Therefore, the geometry depicted in Figure 4.5 can alternately be described by our choice of three parameters: front gate length (L_{fg}), offset between the FG and BG (δ), and gate length difference between the FG and BG (ΔL). By definition, the variability in L_{fg} is identically equal to the line width variability given by (4.2.3).

For any linear combination of n correlated Gaussian random variables,

$$U = \sum_{i=1}^{n} a_i u_i, \tag{4.3.5}$$

the variance of the linear combination can be given by [79]

$$\mathrm{Var}(U) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i} a_i a_j \sigma_i \sigma_j \rho_{ij}. \tag{4.3.6}$$

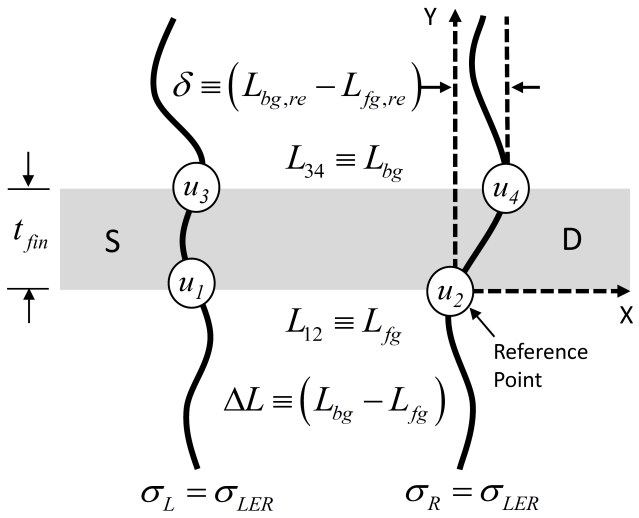


Figure 4.6: Definition of model parameters. The bold lines represent the left and right edges of the gate electrode. Points u_1-u_4 are the locations where gate electrode intersects the fin. The drain is arbitrarily assumed to be on the right side.

Indices i and j are points on the LWR profile as described in Figure 4.4, σ is their respective standard deviation, and ρ_{ij} is correlation between points i and j. Let us first define the offset parameter δ as the difference between the right edges of FG and BG, namely, points u_2 and u_4 in Figure 4.6.

$$\delta = (L_{bg,re} - L_{fg,re}) = (u_1 - u_2). \tag{4.3.7}$$

Therefore, using (4.3.6), we can write

$$\sigma_\delta^2 = a_2^2\sigma_2^2 + a_4^2\sigma_4^2 + 2a_2a_4\sigma_2\sigma_4\rho_{24}. \tag{4.3.8}$$

Substituting, $a_2 = -1$, $a_4 = 1$, $\sigma_2 = \sigma_4 = \sigma_{LER}$ and $\rho_{24} = \rho_A(t_{fin})$, we can express the variation in the offset parameter δ as

$$\sigma_\delta^2 = 2\sigma_{LER}^2\left(1 - \rho_A(t_{fin})\right). \tag{4.3.9}$$

As mentioned previously, the fin thickness (t_{fin}) determines the sampling distance in the auto-correlated LWR function along each edge of the gate electrode as defined in (4.2.4). The difference in gate length (L) between the FG and BG is given by

$$\Delta L \equiv (L_{bg} - L_{fg}) = (u_3 - u_4) - (u_1 - u_2). \tag{4.3.10}$$

The locations of points u_1, u_3, and u_4 relative to point u_2 are random, but related variables. Again, we invoke the use of (4.3.6) by substituting $a_1 = -1$, $a_2 = 1$, $a_3 = 1$, $a_4 = -1$,

$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_{LER}$, and:

$$\begin{aligned} \rho_{12} = \rho_{34} = \rho_X(0), \\ \rho_{14} = \rho_{23} = \rho_X(t_{fin}), \\ \rho_{13} = \rho_{24} = \rho_A(t_{fin}), \end{aligned} \tag{4.3.11}$$

where $\rho_X(0)$ and $\rho_X(t_{fin})$ are the cross-correlation terms between the left and the right edges evaluated at lag 0 and t_{fin}, respectively. $\rho_A(t_{fin})$ is the auto-correlation term defined in (4.2.4) and evaluated at lag t_{fin}. For resist-defined gate electrode, we have $\rho_X(0) = 0$ and $\rho_X(t_{fin}) = 0$, and the variation in ΔL is given by

$$\sigma^2_{\Delta L} = 4\sigma^2_{LER}\left(1 - \rho_A(t_{fin})\right). \tag{4.3.12}$$

It should be noted that for a given σ_{LER}, the variability in ΔL is twice the variability in δ. Similarly, for a spacer-defined gate electrode, we have $\rho_X(0) = 1$ and $\rho_X(t_{fin}) = \rho_A(t_{fin})$. The latter equality holds true because for spacer-defined gate electrode the left and right edges are assumed to be identical. Thus, for spacer-defined gate electrode, the variation in ΔL is zero:

$$\sigma^2_{\Delta L} = 0. \tag{4.3.13}$$

Overall variability in device parameter P depends on many process factors; gate and fin geometries are two important factors. It has been previously shown that fin LWR primarily affects the device performance by changing the average fin width in the channel region [62]. Therefore, to the first order, the variability in device parameter P due to fin LWR can be modeled as

$$\sigma^2_{P,F} = \left(\frac{\partial P}{\partial t_{fin}}\right)^2 \sigma^2_{F,LWR}, \tag{4.3.14}$$

where $\sigma^2_{F,LWR}$ is the variance in the fin width due to LWR. The variability in device parameter P, purely in terms of gate and fin geometries, can be written as

$$\sigma^2_P = \sigma^2_{P,F} + \sigma^2_{P,G}. \tag{4.3.15}$$

Here the subscripts 'F' and 'G' refer to the fin and gate contributions to device parameter variance respectively. Since fin and gate electrodes are formed independently, their variance can be assumed to be statistically independent. In this chapter, we focus primarily on the contribution of gate to device performance variability. In the following section, we estimate the device parameter sensitivity to the model parameters L_{fg}, δ, and ΔL via 2-D device simulations using a deterministic grid of values for these parameters. Variability in these geometrical model parameters is transformed into variability in device parameters via probability density functions generated from the deterministic set.

Parameter	Units	Value
$V_{t,sat}$	mV	210
SS	mV/dec	69
DIBL	mV/V	30
$g_{m,sat}$	mA/V	6.75
$I_{d,sat}$	mA/μm	2.48
I_{off}	pA/μm	94.4

Table 4.2: 2-D Nominal Device Performance Parameters

4.4 Results and Discussion

Nominal transistor performance parameters obtained from 2-D device simulation are shown in Table 4.2. They roughly matches the ITRS roadmap values for 32nm HP node [60]. Hereafter, device parameters will be referenced to the nominal device where no offset or gate length difference exists between the FG and BG. The saturation threshold voltage refers to the value of V_{gs} corresponding to 100nA/ m, for V_{ds}=0.9V.

Let us first understand the fin LWR contribution to device parameter variability. The ITRS does not specify any LWR requirements for fin width [60]. We assume that the fin LWR budget for the 32nm node (7.5nm fin width) is the same as the gate LWR budget for the 18nm node (7nm physical gate length). Thus, $3\sigma_{F,LWR}$ is assumed to be 1nm. Figure 4.7 shows the fin width dependence of saturation threshold voltage. It should be noted that at 28mV/nm, the threshold voltage is quite sensitive to the fin width thickness variation. Thus, using (4.3.14), the fin LWR is estimated to contribute 28mV (3σ) to the total variation in $V_{t,sat}$. Fin width dependence of saturation current and sub-threshold leakage current are shown in Figure 4.8. Fin LWR is estimated to induce 82μA/μm and 0.5 log(A)/μm variation in $I_{d,sat}$ and I_{off}, respectively.

Given the complex statistical nature of LWR, M-C approach is an obvious choice. However, since M-C TCAD simulations are computationally expensive, and they require a large number of runs in each case to determine the statistical parameters with reasonable accuracy, we employed a methodology based on experimental design techniques to eliminate the need for full M-C simulations. First, we performed 2-D simulations for a pre-determined set of values for L_{fg}, ΔL, and δ at 0.5nm interval within 6nm range (-3nm to +3nm) around their respective means. The computational cost for the exploratory simulation of the three aforementioned parameters is $\mathcal{O}(n^3)$, where n is the number of steps in each of the three parameter dimensions (L_{fg}, ΔL, and δ). The choice of 0.5nm step size was based on a trade-off between the TCAD computational time and the investigative range of each parameter.

Figure 4.9 shows a three-dimensional plot of threshold voltage sensitivity to ΔL and δ for a device with nominal FG. Figure 4.10(a) shows that when the BG is smaller than FG, the

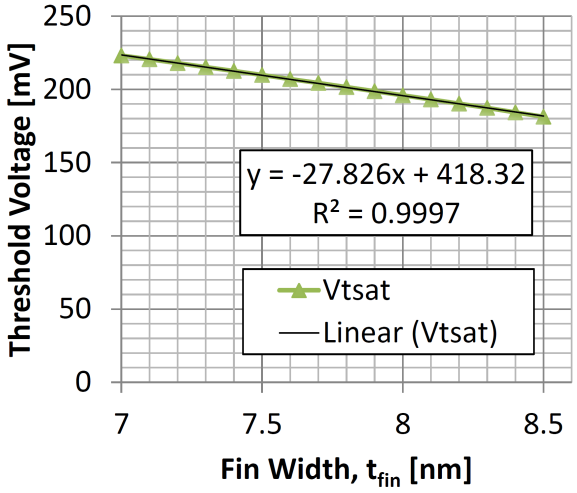


Figure 4.7: Fin width dependence of saturation threshold voltage. $V_{t,sat}$ is defined to be V_{gs} corresponding to 100 nA/mm I_{ds}, for $V_{ds} = 0.9$V.

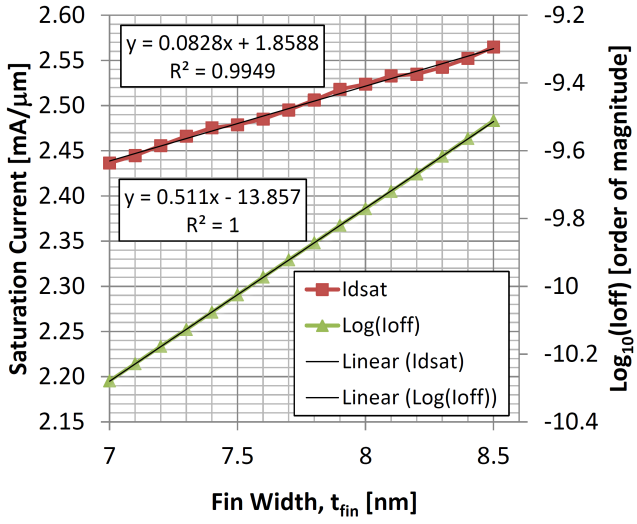


Figure 4.8: Fin width dependence of saturation current and sub-threshold leakage current.

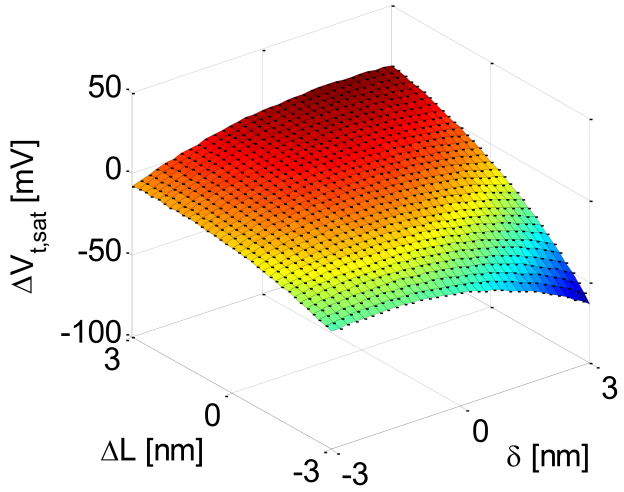


Figure 4.9: Threshold voltage variation over ΔL (CD difference) and δ (FG to BG offset) space. A 13nm FG gate length is assumed.

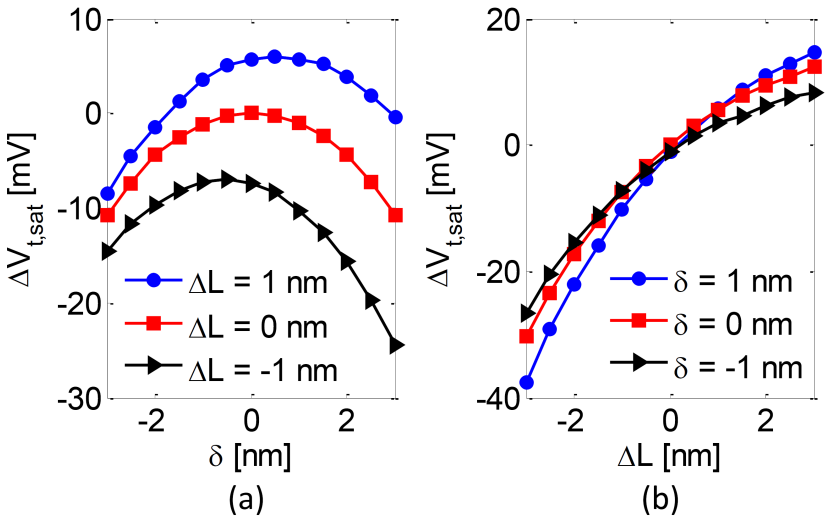


Figure 4.10: Threshold voltage dependence on δ and ΔL. FG gate length of 13nm is assumed. Positive values of DL correspond to larger BG compared to FG while positive values of δ correspond to BG shifted more towards the drain as compared to FG.

threshold voltage is lowered more for the BG shifted towards the drain vs. the BG shifted towards the source. This effect is reversed when BG is larger than FG. Another important observation from Figure 4.10(a) is that for a given CD mismatch between the FG and BG, FinFET threshold voltage is relatively invariant over some range of gate offset. In contrast, even for no gate offset, the threshold voltage is fairly sensitive to CD mismatch as shown in Figure 4.10(b). Thus, CD mismatch between the FG and BG is more critical than the gate offset.

By performing device simulation for this "grid", basically we mapped out the variability space for model parameters L_{fg}, ΔL, and δ. The computational efficiency of our approach is enabled by the structure of our model that parameterizes the FinFET structure in terms of L_{fg}, ΔL, and δ and relates them to LWR descriptors ξ, σ, and α. In other words, we have $L_{fg} \sim \mathcal{N}(L_{fg}, \sigma^2_{LWR})$, $\delta \sim \mathcal{N}(0, \sigma^2_{\delta})$, and $\Delta L \sim \mathcal{N}(0, \sigma^2_{\Delta L})$ where σ^2_{LWR}, σ^2_{δ}, and $\sigma^2_{\Delta L}$ are defined by (4.2.3), (4.3.9), and (4.3.12) or (4.3.13), respectively Any realization of gate LWR is translated into corresponding values of L_{fg}, ΔL, and δ, and subsequently, the device performance can then be estimated through straightforward interpolation using the pre-simulated grid. Thus, expensive TCAD simulations need to be performed only once at each of the grid values, and for any future set of LWR parameters (ξ, σ, and α), a M-C experiment can be performed *outside* of the TCAD environment (in any tool such as MATLAB [32]). A lithography process engineer may need to evaluate several scenarios of LWR descriptors before settling for a given process. An accurate assessment of each scenario would warrant a minimum of 200 M-C run. Thus, the initial "investment" of TCAD simulation is quickly paid off if many such scenarios need to be evaluated.

We generated 2000 M-C samples of L_{fg}, ΔL, and δ using (4.2.3), (4.3.9), and (4.3.12). We assumed $\alpha = 0.5$, $\xi = 10$, and $\sigma_{LWR} = 0.5$. This M-C set was directly simulated with Sentaurus; each run took approximately 200 seconds on 2GHz quad CPU running 64-bit Linux. The same M-C set was also approximated by interpolation using the pre-simulated grid values. Interpolation in MATLAB was completed in less than 5 seconds. The resulting probability density functions are compared in Figure 4.11, and we conclude that our "grid" approach produces reasonably accurate results with very good computational efficiency. For all subsequent analysis, the probability density function for each device parameter was approximated by interpolating 10,000 values of L_{fg}, ΔL, and δ using the pre-simulated basis set. $\alpha = 1$ was assumed for all roughness profiles. Figure 4.12 shows the impact of LWR parameters σ_{LWR} and ξ on the variability in saturation threshold voltage $V_{t,sat}$ for a resist-defined gate electrode. An increase in σ_{LWR} or a decrease in ξ results in greater variation in δ, and hence, in the effective channel length. Thus, the variation in the threshold voltage increases due to SCE. Among the gate resist requirements specified by the ITRS for the 32nm node, the allocated 3σ budget for low-frequency LWR is 1.7nm [60]. From Figure 4.12(a), it should be noted that for the ITRS roadmap stipulated value of 1.7nm for $3\sigma_{LWR}$, we observe 21-30mV variation (3σ) in $V_{t,sat}$ as compared to 16mV (1σ) variation reported due to WFV [57]. Additionally, we observe that $V_{t,sat}$ sensitivity to σ_{LWR} ranges from 14-17mV/nm. This is roughly half compared to 28mV/nm $V_{t,sat}$ sensitivity to t_{fin} variation observed in Fig-

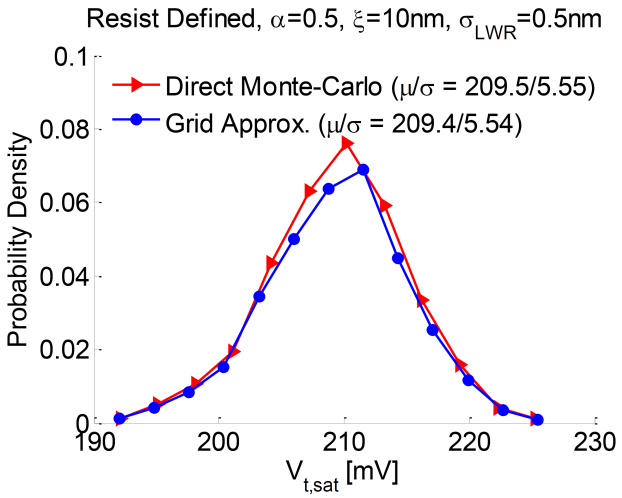


Figure 4.11: Comparison of threshold voltage distributions obtained via direct Monte-Carlo simulation and experimental grid for L_{fg}, ΔL, and δ. For the Monte-Carlo approach, 2000 random values of L_{fg}, ΔL, and δ were generated and directly simulated with Sentaurus. For the grid approach, pre-determined values of L_{fg}, ΔL, and δ at 0.5nm spacing were simulated, and then the same random values were interpolated to this grid.

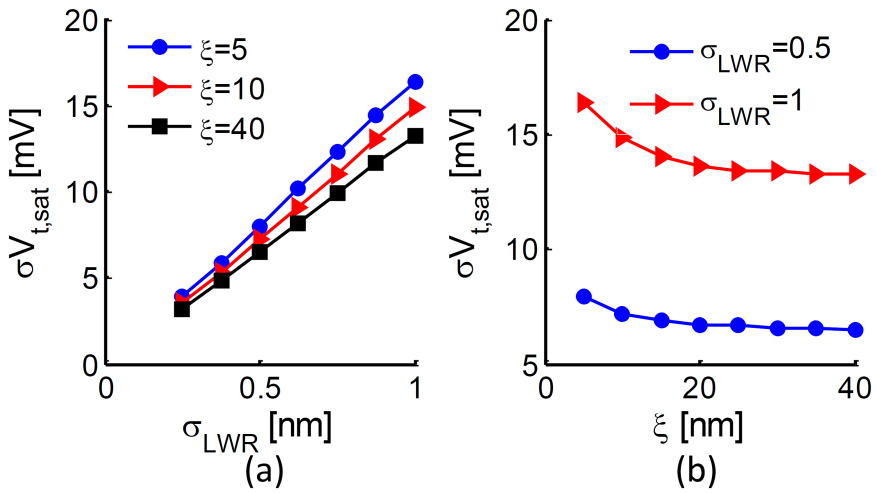


Figure 4.12: Variability in saturation threshold voltage $V_{t,sat}$ for resist-defined gate electrode, (a) as a function of LWR amplitude, and (b) as a function of correlation length. Note that variability in $V_{t,sat}$ is a much stronger function of LWR amplitude than it is of correlation length. The fin width in both plots is 7.5nm.

ure 4.7. Thus, the fin width variation is the more significant component than gate line width roughness. Variability trends for $I_{d,sat}$ and $\log_{10}(I_{off})$ are consistent with the trends observed for $V_{t,sat}$, as shown in Figure 4.13 and Figure 4.14. Figure 4.15 shows that the variability in $V_{t,sat}$ is further lowered when the gate electrode is spacer-defined. In the spacer-defined case, $\sigma^2_{LWR} = 2\sigma^2_{LER}$ is assumed. Consistent reductions in variability were also seen for $I_{d,sat}$ and $\log_{10}(I_{off})$ (not shown). It should be noted these trends observed for DG FinFETs contrast with those reported for planar bulk MOSFETs [80]. Constantoudis et al. observed that a larger correlation length increased threshold-voltage variability, and thus, lowered the yield (defined as 10% tolerance in threshold voltage) [80]. However, for FinFETs with either resist- or spacer-defined gate electrodes, an increase in correlation length reduces the variation in $V_{t,sat}$.

4.5 Summary

The impact of gate LWR on FinFET performance variability is studied in this work. Using a simple analytical model that relates LWR parameters to DG structure parameters, we were able to gain physical insight into LWR, and assess its impact on DG-FET performance. For any given LWR profile, we showed that the framework presented in this chapter can be used to assess device performance variability quickly without having the need to perform extensive

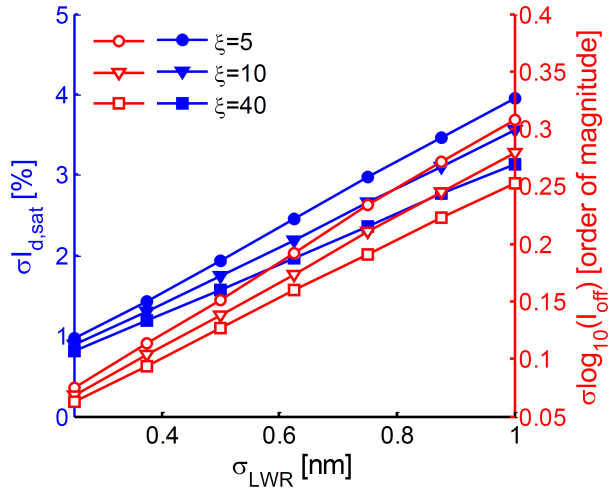


Figure 4.13: Variability in saturation drive current (filled symbols, left y-axis) and off-state leakage current (open symbols, right y-axis) for resist-defined gate electrode as a function of LWR amplitude. The fin width is 7.5nm.

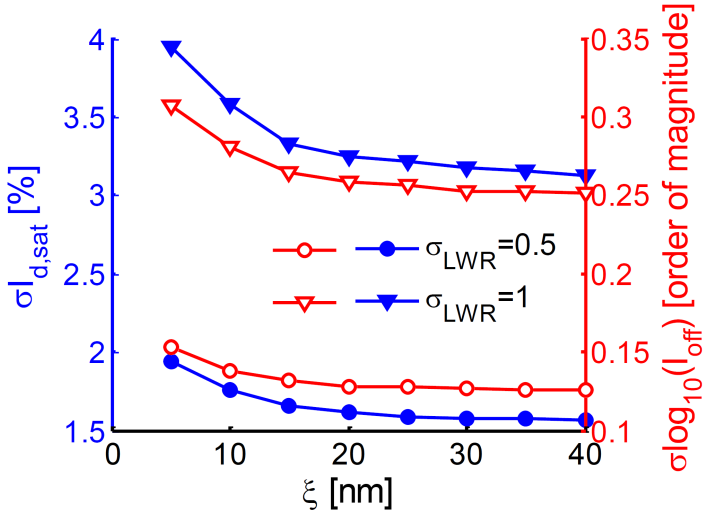


Figure 4.14: Variability in saturation drive current (filled symbols, left y-axis) and off-state leakage current (open symbols, right y-axis) for resist-defined gate electrode as a function of correlation length. The fin width is 7.5nm.

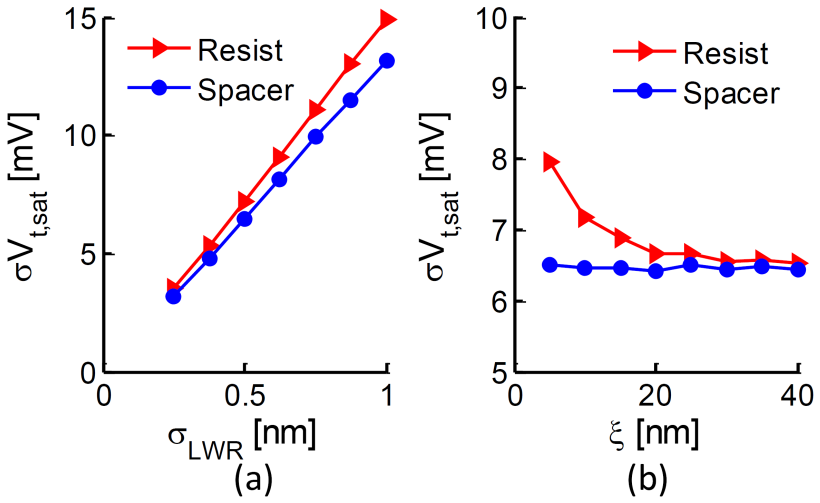


Figure 4.15: Comparison of variability in saturation threshold voltage $V_{t,sat}$ for resist-defined vs. spacer-defined gate electrode (a) as a function of LWR amplitude for $\xi = 10$nm, and (b) as a function of correlation length for $\sigma_{LWR} = 0.5$nm. The fin width in both plots is 7.5nm.

Monte-Carlo TCAD simulations each time a new LWR profile needs to be investigated. Furthermore, if a compact model for DG-FET were to be developed, and parameterized in terms of gate geometrical parameters, then our framework can be used to estimate the variability of any device parameter of interest.

Chapter 5

Decomposition of Semiconductor Process Variation

5.1 Introduction

Variability means lack of uniformity. Uniformity implies predictable performance while lack thereof results in yield loss or degraded circuit performance. Variability is experienced at all levels of semiconductor circuit fabrication hierarchy—lot, wafer, die, and devices or structures within a die. Variability arises from the multitude of processing steps it takes to fabricate a wafer (the unit of production), and it causes undesired variation in the performance of electrical circuits on a die or chip (unit of merit). Random (or intrinsic) and systematic (or deterministic) variation can be observed at all levels of the aforementioned hierarchy. Systematic spatial variation, at large or small scale, can cause spatial dependencies or correlation between collection of structures on a die. Devices or structures that are in close proximity behave much more similarly than those that are spaced farther apart. Several efforts have been made to incorporate *a priori* knowledge of variability (especially spatial correlation) in characterization of timing issues in digital logic circuits through statistical static timing analysis (SSTA) [8, 9, 10, 11, 12, 13]. However, the same level of attention has not been devoted to the extraction of model parameters from actual silicon data.

One of the early works in the area of variability modeling was due to Stine et al [81]. They provided methods for decomposing variability hierarchically, however, they did not address the statistics of across-wafer spatial variation, and also no method was provided to model spatial correlation. Reda and Nassif [82] presented a method of modeling the statistics of wafer-level variation through multivariate normal (MVN) approach, but their work did not address the intra-die variation. Xiong et al [83] emphasized the extraction of a *valid* correlation matrix, but the proposed model was isotropic, wafer-level systematic variation was ignored, and validation was not performed on actual silicon data. Sato et al [84] only addressed intra-die systematic and random variation, and they ignored the inter-die and

spatial correlation aspects of variation. Recently, there have been two publications that advocate variogram-based modeling of spatial correlation [85, 86]. Chopra et al [85] highlighted the anisotropic nature of spatial correlation through the use of variogram. However, they ignored intra-die systematic spatial variation, and they did not address the statistics of across-wafer spatial variation. Liu [86] used the universal kriging concept to capture the intra-die correlation. Kriging, however, is a prediction tool that uses the observed spatial correlation in data to predict values at unobserved locations. In case of semiconductor variability models, *new* simulation data needs to be generated for use in Monte-Carlo simulations with the express goal of emulating observed variance. Furthermore, the kriging variance is lower than the observed variance in the data, and so using it would result in simulated data with underestimated variance. As such, kriging is ill-suited for our application. However, Liu [86] did recognize the influence large-scale variation on intra-die spatial correlation; the median-polishing method was used to remove it. As expected, very weak correlation was observed as a result.

In this work, we use actual silicon data from a high-volume fabrication line to provide a holistic view of variability. Our proposed methodology uses multivariate normal (MVN) framework for modeling of the statistics of wafer-scale variation. Outliers and transient observations tend to corrupt the model estimates, and it can potentially have dire consequences on decision-making use of the model. In this work, the MVN framework is used to reject outlier wafers, and it is also used to enable cluster analysis as a wafer selection tool prior to model extraction. Variogram is used to estimate the spatial correlation structure of the residuals. The rest of the chapter is organized as follows: section 5.2 provides a background on process variation. Variability models are presented in section 5.4. Representation of wafer-level and die-level model terms are discussed in section 5.5 and section 5.7, respectively. The wafer-selection procedure is presented in section 5.6. Variogram-based method for residual analysis is presented in section 5.8. Lastly, results and concluding remarks are presented in section 5.9 and section 5.10, respectively.

5.2 Semiconductor Process Variation

In semiconductor device fabrication, there are three classes of parameters: *structural*, *material*, and *electrical* [87]. Film thickness and transistor gate length/width are examples of structural parameters. Compressive or tensile stress and doping variations are examples of material parameters. Threshold voltage (V_T) and effective electrical channel length (L_{eff}) are examples of electrical parameters. Any of the aforementioned class of parameters may be used for modeling. The choice of modeling parameter should be based on its ease of being incorporated in circuit simulation.

Figure 5.1 provides an overview of various components and attributes of variability and different factors affecting it. Semiconductor process variation can broadly be described by three attributes: (a) factors affecting the variation, (b) position or location of variation, and

(c) nature or type of variation. The factors affecting variation can be classified as *environmental*, *physical* and *temporal*. Environmental factors such as temperature, power supply voltage, noise coupling in networks, etc., impact the IC during its operation. Physical factors such as mask imperfections and manufacturing process variations also affect variability. Reliability degradation such as NBTI and oxide traps are examples of temporal factors that increase variability over time.

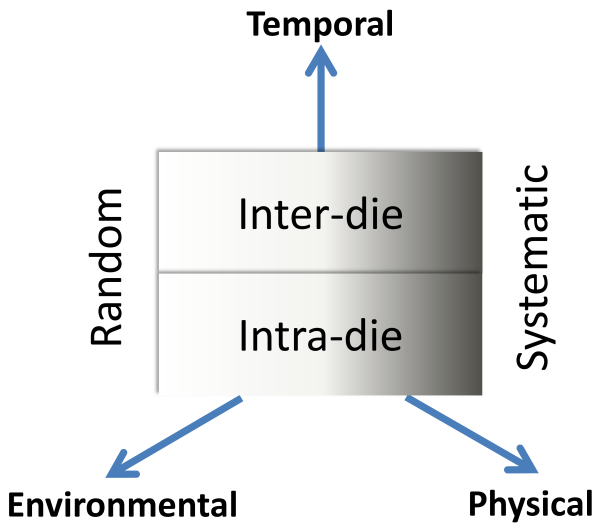


Figure 5.1: Overview of various components and attributes of variability and different factor affecting it. Any physical or electrical parameter is affected by enviromental (such as temperature, power supply voltage and noise coupling in networks), physical (such mask imperfections and manufacturing process variations, etc.), and temporal (such as NBTI and oxide traps) factors. Variation can be either intra-die (or local) or inter-die (or global). Variation may also have systematic (or deterministic) and random components.

Semiconductor circuit fabrication has a naturally occurring hierarchy—lot, wafer, die, and devices or structures within a die. Figure 5.2 partially depicts the semiconductor process variation hierarchy. Variation of any electrical or physical parameter is generally decomposed in a manner that respects this hierarchy. However, from circuit design perspective, the variation is simply either *inter-die* or *intra-die*. *Inter-die* variation is the deviation observed from one die to next. It equally affects all devices or structures on a given die. *Intra-die* variation is the deviation observed among the devices or structures within a given die. Inter-die and intra-die variation are also commonly referred to as *global* and *local* variation, respectively. Lot-to-lot, wafer-to-wafer, and across-wafer variation contribute to inter-die variation. On a larger scale, fab-to-fab variation also acts as an additional source of inter-die variation. Within-die variation is synonymous with intra-die variation.

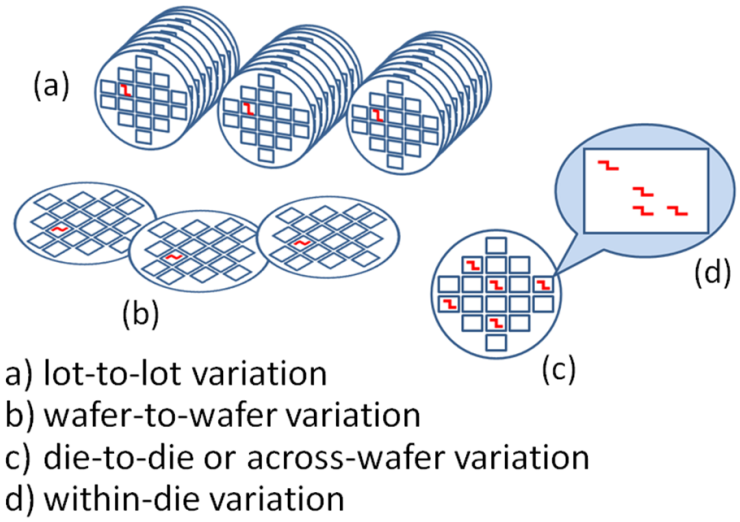


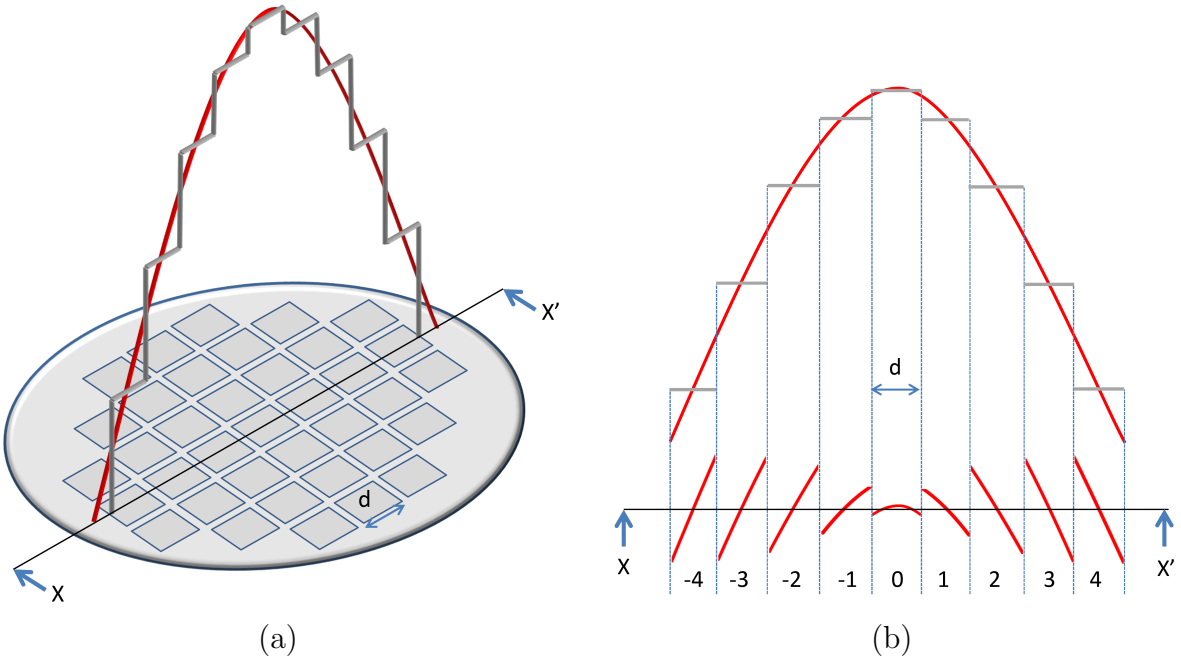
Figure 5.2: Illustration of semiconductor process variation hierarchy. Lot-to-lot, wafer-to-wafer, and across-wafer variation contribute to inter-die variation. Within-die variation is synonymous with intra-die variation. *Graphic courtesy of Kun Qian (UC Berkeley).*

We have thus far discussed the various hierarchical components of variation and the factors influencing these components. One final distinction of increased importance is the nature of variation. Variation may be *random* or *systematic* in nature. Random variation represents the uncertain component of variation. Random variation causes differentiation within a collection of devices on the same die. It is often assumed to be Gaussian. Random dopant, line edge roughness, metal gate granularity, high-k granularity, interface roughness, etc., are examples of random variation [4]. These sources of variation are *intrinsic* to semiconductor processing, and as such, the random variation is also interchangeably referred to a *intrinsic variation*.

Systematic spatial variation can occur at large (lot or wafer) or small (die) scale. Wafer-level systematic spatial variation is typically found to be a slowly varying, smooth function across the wafer. It is often due to thermal gradients in annealing, non-uniformity of film thickness, chemical-mechanical polishing (CMP), etc.. Systematic spatial variation across a die can be caused due to layout (lithography mask) variation or topography (pattern density for CMP). However, systematic variation need not be limited to die or wafer alone. It can also be observed at the lot-level. A temporal drift in a process chamber for single-wafer processing or non-uniformity of fluid flow in a chemical tank can cause systematic variation for individual wafers in lot.

5.3 Spatial Correlation

Spatial correlation helps us quantify spatial dependencies or correlation between collection of structures on a die. Devices or structures that are in close proximity behave much more similarly than those that are spaced farther apart. In this section, we demonstrate how the systematic wafer-level spatial variation is related to intra-die spatial correlation.



(a) (b)

Figure 5.3: Source of spatial correlation due to wafer-level variation. (a) The *unknown* large-scale variation (assumed to be parabolic; denoted by the red curve) across the wafer is *estimated* as piecewise constant (flat gray horizontal lines) term ω_{ijk} in (5.4.3). (b) Removing ω_{ijk} does not remove the influence of the gradient of wafer-level variation, it merely zero centers it. As shown in (b), the amount of wafer-die interaction depends on the sampling dimension d. Lower d reduces the significance of wafer-die interaction, and it may even make this interaction negligible. Also, as shown in (b), lower gradient in wafer-level variation results in more effective removal of the influence of wafer-level variation.

Consider the across-wafer spatial variation in some observed variable as depicted in Figure 5.3. For simplicity, let us assume this variation is parabolic (as denoted by the solid line in Figure 5.3(a)), and that it is smooth and continuous over the entire wafer. This assumed parabolic variation is *unknown* to us. It is important to recognize it as being unknown, because we can only *estimate* it based on the observations made available to us. We have intra-die observations in each die, with many such die across the wafer and many wafers

across different lots. Suppose that we approximated this smooth across-wafer spatial variation as an average of intra-die observations at each die location. Subtracting the piecewise constant average values from intra-die observations simply zero-centers the variation over the individual die. As shown in Figure 5.3(b), the resulting residuals capture a segment of the across-wafer variation. If we were to estimate the auto-correlation at the die-level, the residuals would appear to be correlated. Note that one cannot estimate the auto-correlation using lags larger than the die dimensions, because doing so would result in erroneously high variation when the lag pair is across the die boundary. Using multiple wafers as observations is, however, perfectly valid and recommended.

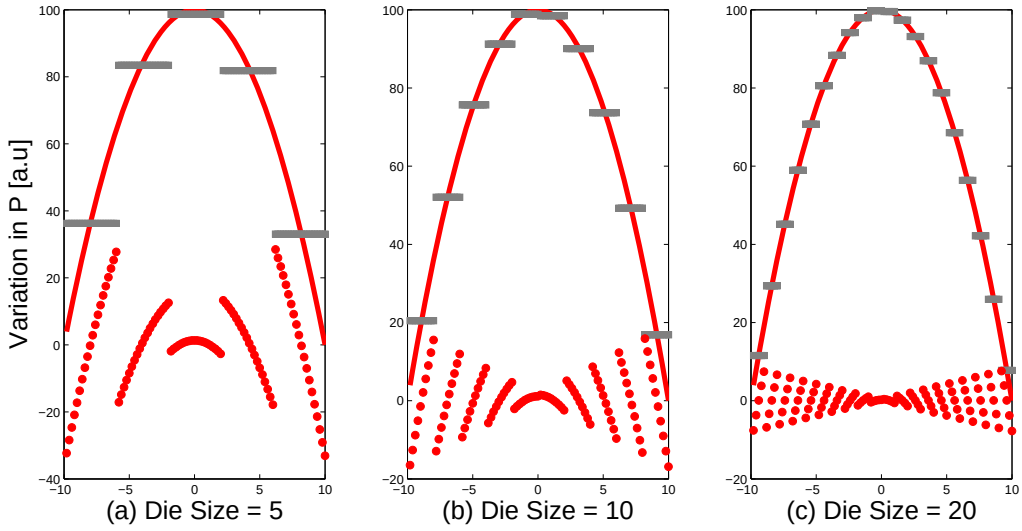


Figure 5.4: Example of the factors influencing our ability to estimate spatial correlation. Red curve is the unknown variation in the modeling parameter. Gray bars indicate the average of the red curve for a given die size. Red dots indicate the residual effect of curve on within-die locations after the die averages have been subtracted.

The amount of auto-correlation observed in the within-die residuals depends on the gradient of the wafer-level variation as well as the die dimension d. Figure 5.4 shows that as the die size is reduced, the wafer-level variation is better approximated. Note the reduction in variability of residuals as the die size is reduced. In practice, for a sufficiently small die size, the spatial correlation in the residuals could be undetectable due to "noise". The "noise" is due to the within-die independent variation, and also due to the variation in across-wafer spatial variation from wafer to wafer. Thus, the limit to which spatial correlation in the residuals is *observable* is determined by the granularity of the observations, the gradient of the wafer-level spatial variation and the within-die "noise".

5.4 Decomposition of Process Variation

In dealing with spatial data, the decomposition of a parameter p

$$p = \text{large-scale variation} + \text{small-scale variation}$$

cannot be specified uniquely and is largely operational in nature [28]. For instance, for the hierarchy described in Figure 5.2, a simple lumped model [87]

$$p = p_0 + p_{inter-die} + p_{intra-die} + p_\epsilon \tag{5.4.1}$$

can be refined further as [87]

$$p_{inter-die} = p_{lot-to-lot} + p_{wafer-to-wafer}(\text{lot}) + p_{die-to-die}(\text{wafer}). \tag{5.4.2}$$

Here p_0 is the nominal design value, p_X is the variation due to source "X" and p_ϵ is the unexplained or random variation. An important consideration in the decomposition of process variation comes from the proposed application of the model. The circuit performance often needs to be evaluated very early in the design cycle, when the physical design (chip layout) is not available. In such a context, an elaborate model that characterizes the intra-die spatial variation is useless and $p_{intra-die}$ and p_ϵ in (5.4.1) might as well be combined and treated as random [87]. In this section, we present a model for the benefit of the circuit designer. We include terms for global variation as well as systematic and random local variation. We implicitly assume that the location based within-die variation information can be utilized by the circuit designer. However, the model can also be collapsed to randomize the within-die location if it is used in the pre-layout context.

5.4.1 Nomenclature

In this work, the modeling parameter being modeled will be denoted as p, and as discussed previously, p can represent a structural, material, or electrical parameter. In this section, greek letters will be used to denote model terms, although, greek letters will also be used in subsequent sections for describing procedural details. Estimated model components will be accentuated by '$\wedge$'. Table 5.1 shows subscript notation used in this work. p_{ijkl} represents a specific observation of parameter p. $\bar{p}$ denotes an averaged observation with a $(\cdot)$ denoting the subscript over which the averaging was performed. For example, $\bar{p}_{\cdots}$ indicates the global average, $\bar{p}_{i\cdots}$ indicates a vector of lot averages (average of all observations for the i-th lot), and so forth.

5.4.2 Modeling Choices

One way of decomposing the process variation is

$$p_{ijkl} = \eta + \omega_{ijk} + \delta_l + \epsilon_{ijkl}. \tag{5.4.3}$$

Index	Attribute
i	Lot
j	Wafer within a lot
k	Die within a wafer
l	Location within a die

Table 5.1: Subscript notations used in this work.

S. No.	Model Term Estimator	Description
1.	$\widehat{\eta} = \bar{p}_{....}$	Global average
2.	$\widehat{\omega}_{ijk} = \bar{p}_{ijk.} - \bar{p}_{....}$	Die average; Global (inter-die) component
3.	$\widehat{\delta}_l = \bar{p}_{...l} - \bar{p}_{....}$	Average die; Local (intra-die) systematic component
4.	$\widehat{\epsilon}_{ijkl} = p_{ijkl} - \bar{p}_{ijk.} - \bar{p}_{...l} + \bar{p}_{....}$	Residual; Local (intra-die) random component

Table 5.2: Estimation of model terms in (5.4.3). The estimation of its terms is done in an hierarchical fashion—the residuals from one estimator become input to the next estimator.

Table 5.2 shows how each model term can be estimated from the observed data. The model represented by (5.4.3) is a *hierarchical* model. The estimation of its terms is done in an hierarchical fashion—the residuals from one estimator become input to the next estimator. For example, $\widehat{\omega}_{ijk}$ is estimated by subtracting the global average term from the observation p_{ijkl}, followed by averaging the result over index l. The model (5.4.3) decomposes the observation p_{ijkl} into a global average (η), die averages across all wafers and lots (ω_{ijk}), average die (δ_l), and a residual(ϵ_{ijkl}) term. The global average term is used as reference value for all other estimators. As indicated in Table 5.2, ω_{ijk} represents the global component of variation. It contains information regarding across-wafer spatial variation as well as temporal variation observed across individual wafers in different lots. If the purpose of decomposition was to include the temporal variation (beyond the scope of this work), one would further decompose ω_{ijk} into a separate lot term. Similarly, if the observed data had a systematic drift in processing of individual wafers in a lot, then the lot-wafer interaction term would be included in the decomposition. The δ_l term captures all systematic spatial variation observed across a die, and its relative contribution can be significant [11].

5.5 Wafer-Level Variation

In dealing with systematic spatial variation, only *single* wafer examples are typically addressed, i.e. some form of spatial variation on a specific wafer is modeled [88]. In practice, however, there are variations in the spatial signatures. Typically, each wafer exhibits statistically significant "individuality" in its across-wafer spatial signature. The *statistics* of wafer-level systematic spatial variation are seldom discussed [82]. In this section, we closely follow the methodology proposed in [82] and model $\widehat{\omega}_{ijk}$ using a multivariate normal framework.

5.5.1 Multivariate Normal Model

Let there be M wafers each comprised of n dice. Mathematically, we can represent each wafer as a vector $\boldsymbol{w}$ consisting of n measurements. Thus, $\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M}$ denote observations from M wafers. Let us further assume that the data is drawn from a multivariate normal distribution (MVN) with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The maximum likelihood estimators (MLE) of $\boldsymbol{\mu}$ and Σ can be given as

$$\widehat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{w_j} \tag{5.5.4}$$

and

$$\widehat{\Sigma} = \frac{1}{M}\sum_{j=1}^{M} (\boldsymbol{w_j} - \hat{\mu})(\boldsymbol{w_j} - \hat{\mu})', \tag{5.5.5}$$

where $\widehat{\boldsymbol{\mu}}$ is a vector of length n representing the average wafer, and $\widehat{\Sigma}$ is a $n \times n$ covariance matrix of all die locations on the wafer. In practice, the computation of $\widehat{\Sigma}$ is quite often complicated by the missing values (die) on wafers. Ignoring missing values decreases the accuracy of the estimates. The missing data can be estimated by a simple polynomial fit [88, 84], universal kriging [86], or by using an *expectation maximization* (EM) algorithm [82]. The detailed steps of these procedures are not repeated here, but they can be found in the cited literature.

5.5.2 Normality Test and Outlier Rejection

Before using $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$, the underlying MVN assumption must be ascertained. This can be accomplished by using the *Mahalanobis* distance. Given the MVN model parameters $\hat{\mu}$ and $\widehat{\Sigma}$, the *Mahalanobis* distance d_j of each wafer observation [82, 89] can be defined as

$$d_j^2 = (\boldsymbol{w_j} - \hat{\mu})'\widehat{\Sigma}^{-1}(\boldsymbol{w_j} - \hat{\mu}). \tag{5.5.6}$$

For multivariate normally distributed data, d_j^2 is approximately chi-square distributed, i.e. $d_j^2 \sim \chi_n^2$, where χ_n^2 represents the *chi-squared distribution* with n degrees of freedom [89]. As discussed previously in section 5.2, the wafer is the unit of production. In selecting the wafers for model extraction, one must ensure that the data set does not contain any outlier wafers. Outlier wafers can be detected by plotting d_j^2 against the quantiles of χ_n^2 [82]. Outlier points are identified by a cut-off value $\chi_{n,1-\alpha}^2$ for certain small α (e.g. $\alpha = 0.05$).

5.6 Wafer Selection

Rejecting the outlier wafers is an essential requirement to qualify a given data set for model estimation. Often large industrial data sets (from test chips or kerf structures) are used for modeling purposes. Such data sets may span significant period in time during which transient events such as process tool drifts, excursions, and experiments may occur. Within a given data set, there could be groups of lots with multiple systematic spatial patterns across-wafer such as "bull's eye" or "donut". One of the basic principles in statistics is that one must only aggregate *similar* observations (formally defined as being derived from the same distribution). Aggregating diverse spatial patterns should therefore be avoided. As such, we need some method to classify the wafers in groups based on their spatial similarities. Once wafers have been classified, one can choose which wafers should be allowed for model estimation. Such classification can also serve as a process diagnostic tool. In this work, we propose a method for classification based on *hierarchical clustering*. Wafers are clustered based on their across-wafer spatial similarities. Our proposed methodology can be summarized as follows: first we use principal component analysis (PCA) to generate some basis functions. Regression is then performed for each wafer using a technique called *least angle regression* (LAR), and parsimony in the regression coefficients is achieved by performing selection using C_p statistic [90]. Lastly, the parsimonious coefficient set is used to perform agglomerative hierarchical clustering.

5.6.1 Least Angle Regression Using PCA Basis Functions

The eigenvalue equation for $n \times n$ covariance matrix Σ (computed in section 5.5) can be expressed as

$$\Sigma U = U\Lambda, \tag{5.6.7}$$

were U is the eigenvector matrix of Σ of size $n \times n$, and Λ is a diagonal matrix of dimension $n \times n$ with non-negative diagonal elements (*eigenvalues*) in decreasing order. The matrix Λ describes the amount of variance explained by each eigenvector (orthonormal column of U), and it can easily be expressed as percentage. Since the columns of U are orthonormal, they can be used as basis functions for regression. However, it is usually preferred to reduce the dimension of the data. Here we choose $m < n$ eigenvectors such that they capture 95% of total variance. X will denote the $n \times m$ eigenvector matrix to be used as basis functions.

At this point, one can use ordinary least squares (OLS) to fit each wafer using the basis function set X. Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}$ be m covariate vectors each of length n (corresponding to die position on wafer). That is, $X_{n\times m} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}]$. Let $\boldsymbol{y}$ be the vector of responses for the n die locations. For a candidate vector of regression coefficients $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m)'$ and a prediction vector $\widehat{\boldsymbol{y}}$

$$\widehat{\boldsymbol{y}} = \sum_{j=1}^{m} \boldsymbol{x_j}\widehat{\beta}_j = X\widehat{\boldsymbol{\beta}}, \tag{5.6.8}$$

the OLS problem can be stated as

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \boldsymbol{x_j}\widehat{\beta}_j \right)^2 \right\}. \tag{5.6.9}$$

For purposes of clustering, we seek parsimony in a coefficient set, i.e. we seek to represent each wafer with the minimal needed number of coefficients and not the full least squares set $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m)'$. Lasso and LAR are a subclass of linear methods known as *shrinkage methods* that can be used to achieve parsimony [91]. Shrinkage methods shrink the regression coefficients by imposing a penalty on their size. They are in essence a constrained version of OLS. For example, the constrained Lasso problem is stated as

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \boldsymbol{x_j}\widehat{\beta}_j \right)^2 \right\}$$
$$\text{subject to } \sum_{j=1}^{m} |\widehat{\beta}_j| \leq t.$$

Due to the nature of the constraint, a sufficiently small t will cause some of the coefficients to be identically zero. The amount of shrinkage in the regression coefficients is usually standardized to the least squares value as $s = t/\sum_{j=1}^{m} |\widehat{\beta}_{LS,j}|$.

LAR was developed recently in the seminal work by Efron et al [90]. LAR and Lasso differ in the manner in which the covariates are chosen [90]. LAR is described as a democratic version of forward stepwise regression in that it only enters "as much" of a predictor as it deserves [91]. The detailed steps of the LAR procedure can be found in [90] and we describe only the high-level algorithmic steps below:

1. Standardize covariates to have zero mean and unit length and that the response to have zero mean

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1, 2, \ldots, m \tag{5.6.10}$$

2. Starting with $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m = 0$, compute the *current correlation* $\widehat{\boldsymbol{c}} = c(\widehat{\boldsymbol{y}}) = X'(\boldsymbol{y} - \widehat{\boldsymbol{y}})$ so that $\widehat{c}_j$ is proportional to the correlation between covariate $\boldsymbol{x}_j$ and the current residual $\boldsymbol{r} = (\boldsymbol{y} - \widehat{\boldsymbol{y}})$.

3. Find the covariate $\boldsymbol{x}_j$ most correlated with the residual $\boldsymbol{r}$ and include $\boldsymbol{x}_j$ into the active set $\mathcal{A}$
$$\widehat{C} = \max\{|\widehat{c}_j|\} \text{ and } \boldsymbol{\mathcal{A}} = \{j : |\widehat{c}_j| = \widehat{C}\}$$

4. Increase β_j from 0 towards its least squares value
$$\widehat{j} = \text{argmax}|\widehat{c}_j| \text{ and } \widehat{\boldsymbol{y}} \rightarrow \widehat{\boldsymbol{y}} + \varepsilon \cdot \text{sign}(\widehat{c}_{\widehat{j}}) \cdot \boldsymbol{x}_{\widehat{j}}$$
with some "small" constant ε. Continue until a competing covariate $\boldsymbol{x}_p$ is equally correlated to the current residual as $\boldsymbol{x}_j$.

5. Include $\boldsymbol{x}_p$ into the active set $\boldsymbol{\mathcal{A}}$, and move β_j and β_p in the direction defined by their joint least squares coefficient of the current residual. Continue until yet another competing covariate $\boldsymbol{x}_q$ is equally correlated to $\boldsymbol{r}$. Include $\boldsymbol{x}_q$ into the active set $\boldsymbol{\mathcal{A}}$.

6. At the k-th step, the active set $\boldsymbol{\mathcal{A}}$ will have k members. Letting
$$X_k = \{\boldsymbol{x}_j : j \in \boldsymbol{\mathcal{A}}_k\} \text{ and } \mathcal{G}_k = X_k' X_k, \tag{5.6.11}$$
the k-th step LARS estimator $\widehat{y}_k$ and the corresponding current residual are given as
$$\begin{aligned} \widehat{\boldsymbol{y}}_k &= X_k {\mathcal{G}_k}^{-1} X_k' \boldsymbol{y} \text{ and} \\ \boldsymbol{r}_k &= (\boldsymbol{y} - \widehat{\boldsymbol{y}}_k), \end{aligned} \tag{5.6.12}$$
respectively. At each step k, the regression coefficient set is given as
$$\boldsymbol{\beta}_k = \{\beta_j : j \in \boldsymbol{\mathcal{A}}_k\}. \tag{5.6.13}$$

7. Repeat until all m covariates have entered the active set, and a full least squares solution is reached.

A principled choice among the range of possible LARS estimates is made by choosing the optimal step. The optimal step is determined as the one that minimizes the expected prediction error. There are many methods available for model selection such as *AIC*, *C_p selection*, *k-fold cross-validation*, etc. [91]. Reference [91] provides an excellent discussion on this subject. In this work, we chose the C_p selection criteria to determine the optimal step as proposed in [90]. We can estimate the risk of a k-step LARS estimator $\widehat{y}_k$ using the statistic [90]
$$C_p(k) \doteq \frac{\| y - \widehat{y}_k \|^2}{\sigma} - n + 2k. \tag{5.6.14}$$
The optimal step k_{opt} is chosen as
$$k_{opt} = \underset{k}{\text{argmin}}\{C_p(k)\}. \tag{5.6.15}$$

5.6.2 Cluster Analysis

Cluster analysis relates to grouping objects into subsets or "clusters" based on some common attributes [92, 93]. Clustering is performed such that the objects within a given cluster are more closely related to each other than with the objects in other clusters. In the present context, the object is a wafer, and the attributes are the optimal set of regression coefficients.

Let $\widehat{\beta}_{ij}$ denote the regression coefficient of the j-th covariate for the i-th wafer. Since the degree of similarity (or dissimilarity) between the individual wafers is the central idea of cluster analysis, we must define a measure of dissimilarity. There are many available choices for the dissimilarity measure, but the most popular choice is the squared distance [91]. The dissimilarity or distance between wafers i and i' can be stated in terms of squared distance between the j-th attribute as

$$D_{ii'} = \sum_{j=1}^{m} (\widehat{\beta}_{ij} - \widehat{\beta}_{i'j})^2. \tag{5.6.16}$$

Given the dissimilarity measure between the wafers, we need an algorithm to perform the grouping. In *hierarchical clustering*, clusters are arranged into a natural hierarchy. There are two methods of hierarchical clustering: *agglomerative* (bottom-up) and *divisive* (top-down) [92, 91]. In this work, we employ agglomerative clustering in which new clusters are formed by pairing two clusters with the smallest dissimilarity. As wafers are clustered in hierarchical fashion, each cluster would ostensibly contain multiple wafers. Thus, in order to discern the dissimilarity between two clusters, each containing multiple wafers, we must define a measure of *intergroup dissimilarity*. Again, there are several methods available: *single linkage* (nearest-distance), *complete linkage* (farthest-distance), *group average* (un-weighted average distance), *centroid* (un-weighted center of mass distance), *median* (weighted center of mass distance) to name a few [91]. Reference [92] presents an excellent discussion on these methods. The choice of intergroup dissimilarity measure is heuristic in nature. If the wafers exhibit strong clustering tendency, then all methods would produce similar results.

The hierarchical structure derived from clustering is graphically displayed as a *dendrogram*, where the objects appear individually at the bottom, and gain membership into clusters as the *cophenetic distance* increases. The *cophenetic distance* $c_{ii'}$ is the intergroup distance at which two objects (i.e. wafers) i and i' are first joined together in the same cluster. Hierarchical clustering methods impose a hierarchical structure on the data irrespective of its actual existence [91, 92]. As such, the efficacy of clustering needs to be ascertained. The *cophenetic correlation coefficient* judges the extent to which the hierarchical structure "faithfully" represents the data itself. It is a kind of a measure of distortion in the classification structure. It indicates the amount of correlation between the $n(n-1)/2$ pairwise observation distances $d_{ii'}$ and their corresponding cophenetic distance $c_{ii'}$ derived from the dendrogram. Discussion on several other measures of distortion can be found in [92].

5.7 Die-Level Variation

In the preceding sections, we proposed methods to capture the statistics of wafer-level spatial variation, and we also showed how to perform wafer selection for model estimation. In this section, we focus on the modeling aspects of within-die systematic spatial variation. The residuals of the model described by (5.4.3) deserve a more involved discussion; the following section is dedicated to it.

The die-level systematic spatial variation is generally modeled using polynomial basis functions [84, 94]. Let $z(x, y)$ describe the systematic spatial variation at the within-die location (x, y). We estimate z using a k-th order polynomial

$$\hat{z}(x, y) = \boldsymbol{X}'\boldsymbol{c}$$

$$\boldsymbol{X} = (1, x, y, x^2, y^2, \ldots, x^k, y^k)' \text{ and } \boldsymbol{c} = (c_0, c_1, c_2, \ldots, c_{k-1}, c_k)'$$

The polynomial coefficients $\boldsymbol{c}$ are determined using the least-squares procedure that minimizes the residual sum of squares (RSS) error. The choice of polynomial order k here is critical: a large k would produce a model with low bias, but large variance in the estimate of p. The *Akaike information criterion* (AIC) or its corrected version (AICc) can be used to determine the appropriate polynomial order [95]. AICc balances the RSS error in the fit with the number of parameters in the model, and it is defined as [95]

$$AICc = \log\left(\frac{1}{n}\sum_{(x,y)}\left(z(x, y) - \hat{z}(x, y)\right)^2\right) + \frac{n + k}{n - k - 2}, \tag{5.7.17}$$

where the summation is performed over all available within-die locations n. Note that while a higher value of k reduces the RSS (first term in (5.7.17)) of the fit, the AICc is penalized by the increase in second term for higher values of k. Thus, the value of k that minimizes the AICc in (5.7.17) is the optimal order of the polynomial. As is customary, the residuals of the fit should be checked for normality, homoscedasticity, and independence. We should point out, however, that the k-th order polynomial is not parsimonious. The coefficients of many terms may not be significant. If a parsimonious within-die model is desired, then there are methods available for performing subset selection of polynomial terms [91].

5.8 Residual Analysis

The nature of the residuals greatly depends on how the variation model was formulated. As previously discussed in section 5.2, the extent to which such approach successfully results in *white* residuals greatly depends the granularity of observations, the gradient of the wafer-level spatial variation, and the within-die uncorrelated variation. In specific cases, it may be possible to realize uncorrelated residuals. In [88], it was accomplished by fitting

a paraboloid to wafer-level systematic variation. However, general applicability of such a parabolic assumption has not yet been demonstrated. Thus, we need a practical method of analyzing the residuals. In this study, we advocate a variogram-based numerical treatment of the residuals. Recently, the use of variogram for modeling spatial correlation has been proposed by two other studies [86, 85], but only [85] used the variogram in the proper context.

5.8.1 Variogram of Residuals

In this section, we shall briefly discuss the theoretical foundation of the variogram. A spatial sequence is considered *intrinsically stationary* if its finite dimensional joint distributions do not change when shifted in position. Consider a intrinsically stationary spatial process in region $\mathcal{R}$, i.e. let $\mathcal{X}_s = \{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}\}$ be a collection of random variables with an unknown mean $\mu \in I\!R$ observed at certain points $\{\boldsymbol{s}_i : i \in \mathbb{Z}\}$ such that

$$E(X(\boldsymbol{s}) - X(\boldsymbol{s}+\boldsymbol{h})) = 0 \tag{5.8.18}$$

and

$$\mathrm{Var}(X(\boldsymbol{s}) - X(\boldsymbol{s}+\boldsymbol{h})) = \mathrm{Var}(X(\boldsymbol{0}) - X(\boldsymbol{h})) = E\left[X(\boldsymbol{0}) - X(\boldsymbol{h})\right]^2 = 2\gamma(\boldsymbol{h}) \tag{5.8.19}$$

for all $\boldsymbol{s} \in \mathcal{R}$. (5.8.18) implies that the mean is constant everywhere in region $\mathcal{R}$. (5.8.19) implies that the variance of the *difference* is constant everywhere in region $\mathcal{R}$, and that it depends only on $\boldsymbol{h}$. In spatial statistics, 2γ is known as the *variogram* and γ is known as the *semi-variogram*. Additionally, if $\mathcal{X}_s$ is *second-order or weakly stationary*, then (5.8.18) holds, and $\mathcal{X}_s$ has a common *auto-covariance function* $C(\boldsymbol{h}) = \mathrm{Cov}(X(\boldsymbol{s}), X(\boldsymbol{s}+\boldsymbol{h}))$ such that

$$\mathrm{Var}(X(\boldsymbol{s}) - X(\boldsymbol{s}+\boldsymbol{h})) = \mathrm{Var}(X(\boldsymbol{s})) + \mathrm{Var}(\boldsymbol{s}+\boldsymbol{h}) - 2\mathrm{Cov}(X(\boldsymbol{s}), X(\boldsymbol{s}+\boldsymbol{h})) \tag{5.8.20}$$

$$= 2\left[C(\boldsymbol{0}) - C(\boldsymbol{h})\right]. \tag{5.8.21}$$

Thus, the second-order stationarity implies intrinsic stationarity, and we have

$$2\gamma(\boldsymbol{h}) = 2\left[C(\boldsymbol{0}) - C(\boldsymbol{h})\right]. \tag{5.8.22}$$

Usually, the spatial correlation is described using an *auto-correlation function* (ACF) $\rho(\boldsymbol{h})$ [88, 94, 83]. The auto-correlation function (or *correlogram*) can be defined in terms of auto-covariance function (or *covariogram*) as

$$\rho(\boldsymbol{h}) \equiv \frac{C(\boldsymbol{h})}{C(\boldsymbol{0})}, \tag{5.8.23}$$

where $C(\boldsymbol{0}) = \sigma^2$ is the finite sample bias-corrected variance of spatial sequence $\mathcal{X}_s$. Using (5.8.22) and (5.8.23), we can alternately define the variogram in terms of the ACF as

$$2\gamma(\boldsymbol{h}) = 2\sigma^2\left[1 - \rho(\boldsymbol{h})\right]. \tag{5.8.24}$$

5.8.2 Variogram Estimation

Generally speaking, there are four issues that are commonly encountered in the estimation of variogram. They are:

- Presence of outlier observations in the data.
- Recovering the covariogram (auto-covariance function) from the variogram.
- Anisotropy in the variogram.
- Insufficient observations to estimate the variogram accurately.

Each of the above issues will be addressed in this section.

A version of Matheron's *classical* variogram estimator is given as [28]

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{N(\boldsymbol{h})} \Big(X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j) \Big)^2, \ \boldsymbol{h} \in I\!R^d, \tag{5.8.25}$$

where

$$N(\boldsymbol{h}) \doteq \{(\boldsymbol{s}_i, \boldsymbol{s}_j) :\| \boldsymbol{s}_i - \boldsymbol{s}_j \| = h; i, j = 1, 2, \ldots, n\} \tag{5.8.26}$$

denotes the number of distinct pairs available at lag $\boldsymbol{h}$. Note that in the classical estimator, the square operation inside the summation greatly magnifies any outlier observation. A more robust estimator was proposed by Cressie and Hawkins [96]

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{0.457 + 0.494/N(\boldsymbol{h})} \left[\frac{1}{N(\boldsymbol{h})} \sum_{N(\boldsymbol{h})} \Big| X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j) \Big|^{1/2} \right]^4, \ \boldsymbol{h} \in I\!R^d, \tag{5.8.27}$$

where $N(h)$ is given by (5.8.26). The estimator (5.8.27) has been found to be robust against outliers [96, 29]. In our context, X represents the residuals of our model (5.4.3) and $\boldsymbol{s} = \{(x_i, y_i) : i \in \mathbb{Z}\}$ is the set of within-die observations at location (x, y).

In this work, our end goal is to use the estimated spatial correlation information to generate simulated data. For this purpose, we need to estimate the auto-covariance function or covariogram. Recall that the variogram and covariogram are related by (5.8.22). We prefer using variogram estimation over covariogram estimation, because it completely avoids the estimation of the mean (which is of little interest in this case). Additional statistical arguments in favor of the variogram have also been made [28]. The problem in recovering the covariogram from the variogram is that their respective estimators do not preserve the property (5.8.22) [28]. In other words, $2\widehat{\gamma}(\boldsymbol{h}) \neq 2\left[\widehat{C}(\boldsymbol{0}) - \widehat{C}(\boldsymbol{h})\right]$. To circumvent this problem, the estimated variogram can be fit to a valid parametric variogram expressed in terms of valid parametric ACF. Several valid analytic forms for ACF exist [28]. Adequacy of any

chosen form should be ascertained by examining its applicability to the given data. We use a specific functional form proposed for isotropic process by [7],

$$\rho(\boldsymbol{h};\xi,\phi) = \exp\left(-\left[\frac{||\boldsymbol{h}||}{\xi}\right]^{2\phi}\right),\ \boldsymbol{h}\in I\!R^d. \tag{5.8.28}$$

ξ is the characteristic dimension of ACF called the *correlation length* and it is defined as $\rho(\xi) \equiv 1/e$. For $h \sim \xi$, any two observations can be considered to be correlated, whereas for $h \gg \xi$, independence can be assumed. Using the Taylor expansion of the exponential function, it can be shown that the ACF exhibits power law behavior $\rho \sim h^{2\phi}$ for $h \ll \xi$. Thus, ϕ determines correlation at extremely short-range, and as such, we will refer to it as *short-range correlation factor*. It can be shown that $0 \leq \phi \leq 1$ [23]. The ACF described by (5.8.28) has the following additional important properties:

1. $\rho(\mathbf{0}) = 1$
2. $\rho(\boldsymbol{h}) = \rho(-\boldsymbol{h})$
3. $|\rho(\boldsymbol{h})| \leq \rho(\mathbf{0})$
4. $\lim_{\boldsymbol{h}\to\infty} \rho(\boldsymbol{h}) = 0$

The last relation holds for a wide class of stationary processes, including the spatial process considered here, but not in general. Using the definition (5.8.22) for the variogram, it is clear that $\gamma(\mathbf{0}) = 0$. However, in practice $\gamma(\boldsymbol{h}) \to c_0$ as $\boldsymbol{h} \to \mathbf{0}$. c_0 has been called the *nugget effect* by Matheron [28]. Although, mathematically the discontinuity at zero lag cannot occur, the existence of c_0 is attributed as *measurement or estimation error*. Thus, we must modify (5.8.24) as

$$2\gamma(\boldsymbol{h}) = 2c_0 + 2\sigma^2\left[1 - \rho(\boldsymbol{h})\right]. \tag{5.8.29}$$

The nugget can be interpreted as the independent (i.e. uncorrelated) component of variation, as it does not depend on the lag. The $2\sigma^2$ value is referred to as the *sill*, and the value of $\boldsymbol{h}$ at which $2\gamma(\boldsymbol{h})$ reaches the sill is known as the *range*.

The ACF (5.8.28) is defined only for *isotropic* processes. However, the process $\mathcal{X}_s$ can also be *anisotropic* (i.e the dependence between $X(\boldsymbol{s})$ and $X(\boldsymbol{s}+\boldsymbol{h})$ is a function of both magnitude *and* direction of $\boldsymbol{h}$). The most common form of anisotropy is *geometric anisotropy*. In geometric anisotropy, the range changes with direction, but the sill remains constant [97]. Geometric anisotropy can be corrected by scale and rotation transformation of the directional vector $\boldsymbol{h}$ as [97]

$$\begin{gathered} h' = ||A\boldsymbol{h}||_2 \text{ with} \\ A = \begin{pmatrix} 1/a_x & 0 \\ 0 & 1/a_y \end{pmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} = \frac{1}{a}\begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}, \end{gathered} \tag{5.8.30}$$

where $||\cdot||_2$ denotes the Euclidean length of vector. Thus, the matrix A transforms the lags from anisotropic space (circle of radius a) to isotropic space (ellipse). The ratio a_x/a_y is known as *anisotropy ratio* λ. It is the ratio of minor range to major range, and $0 \leq \lambda \leq 1$. The rotation angle α is chosen as the direction of maximum range.

Using (5.8.28) and (5.8.30) in (5.8.29), we can restate the parametric form of variogram as

$$2\gamma(\boldsymbol{h};\boldsymbol{\theta}) = 2c_0 + 2\sigma^2\Big[1 - \exp\Big(-\Big[\frac{||A\boldsymbol{h}||_2}{\xi}\Big]^{2\phi}\Big)\Big],\ \boldsymbol{h} \in I\!R^d. \tag{5.8.31}$$

Figure 5.5 graphically illustrates various parameters used to describe the variogram. Let

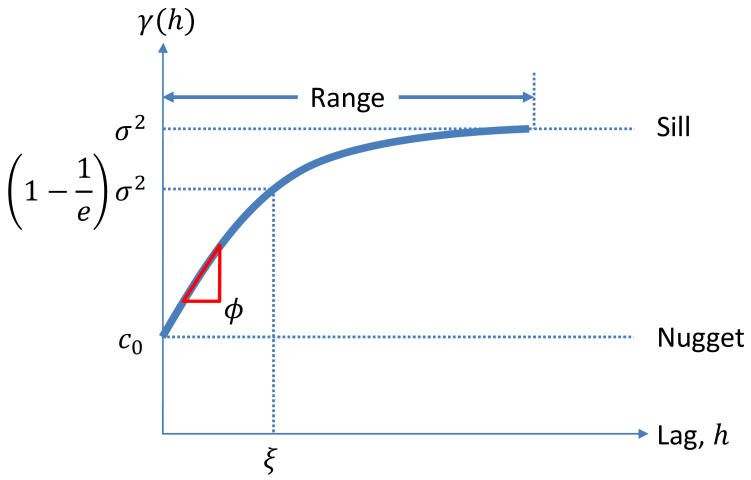


Figure 5.5: Graphically illustration of various parameters used to describe the variogram 2γ.

$\boldsymbol{\theta} \equiv (c_0,\ \sigma^2,\ \phi,\ \xi,\ \alpha,\ \lambda,\ a)'$ denote the vector of parameters that need to be estimated for characterizing the variogram. The subset of parameters $(\phi,\ \xi,\ \alpha,\ \lambda,\ a)$ are commonly referred to as the *structure parameters* as they define the structure or shape of the variogram. In this work, we will use the *weighted least squares* (WLS) method [27] to estimate (5.8.31). Several other methods have been proposed for fitting variogram models [28], but the robustness of WLS, and the absence of any distributional assumptions, makes it the most practical method for fitting the variogram [29]. Note that (5.8.26) indicates that fewer pairs are available at higher lag values. Thus, the variogram error increases at higher lags. The WLS method automatically provides heigher weights for the early lags and lower weights for the lags at which number of contributing pairs is low [27]. Assuming *heteroskedasticity*, the WLS criterion is to minimize

$$(2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta}))'\, V^{-1}\, (2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta}))\,,$$

where V is the diagonal matrix of variances of the variogram. Thus, we can define the WLS estimator of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{h=1}^{h_0} N(\boldsymbol{h}) \left[\frac{2\hat{\gamma}(\boldsymbol{h})}{2\gamma(\boldsymbol{h};\boldsymbol{\theta})} - 1 \right]^2. \tag{5.8.32}$$

where $2\hat{\gamma}(h)$ and $2\gamma(\boldsymbol{h};\boldsymbol{\theta})$ are given by (5.8.27) and (5.8.31) respectively, $N(h)$ is given by (5.8.26), and h_0 is a user-specified upper range of lag values. A practical choice of h_0 is [27]

$$h_0 = \operatorname{argmax}\{h : h \leq H/2 \text{ and } N(h) \geq 30\}, \tag{5.8.33}$$

where H denotes the largest possible lag. The WLS estimator for $\boldsymbol{\theta}$, defined in (5.8.32), can be solved by any non-linear optimization procedure. Most mathematical packages, such as MATLAB, provide built-in functions for constrained non-linear optimization [32].

Based on our earlier discussion in section 5.3, we concluded that if the die averages are removed from the intra-die observations, then the lag pairs used for the variogram estimation must be restricted to be within die. In other words, the lag pairs used for the variogram estimation cannot cross the die boundaries. The hierarchical model (5.4.3) proposed in this work does indeed remove the die averages (ω_{ijk}) in the estimation of residuals. We, therefore, need to estimate the variogram at the die-level. For reliable and accurate estimation of the variogram (5.8.27), we need many within-die observations to cover a wide range of lags. Depending on the source of the data, this may not always be possible. In order to circumvent this problem, we propose estimating the structure parameters (ϕ, ξ, α, λ, a) of the variogram at the wafer-level. However, in order to do so, we temporarily redefine our model as

$$p_{ijkl} = \eta + \delta_l + \varepsilon_{ijkl}. \tag{5.8.34}$$

In (5.8.34), we have dropped the ω_{ijk} term from (5.4.3). The terms $\widehat{\eta}$ and $\widehat{\delta}_l$ are estimated as previously defined in Table 5.2 and the residuals are estimated as $\widehat{\varepsilon}_{ijkl} = p_{ijkl} - \bar{p}_{\cdots l}$. Note that we can *only* estimate the structure parameters using the residuals $\widehat{\varepsilon}_{ijkl}$. The sill and nugget of $\widehat{\varepsilon}_{ijkl}$ (temporary model) and $\widehat{\epsilon}_{ijkl}$ (actual model) will be different. Thus, in the final parametric representation of the variogram (5.8.31), the structure parameters will be estimated using $\widehat{\varepsilon}_{ijkl}$ and the sill will be estimated using $\widehat{\epsilon}_{ijkl}$. Die-level estimation of variogram is more advantageous than wafer-level estimation because the latter method does not allow us to estimate the nugget (independent or uncorrelated component) of $\widehat{\epsilon}_{ijkl}$.

5.9 Results and Discussion

In order to model the process variation effectively, the data needs to be available at a level of abstraction that is conducive to be used in circuit simulation. For instance, if the modeled parameter is an electrical parameter like V_T, then it can be used much more easily in circuit simulation than a material parameter such as film stress. The latter choice would involve

more elaborate simulation setups, and in many cases it may require simplifying assumptions. Generally speaking, electrical parameters are a more desirable choice for modeling purposes, as they tend to incorporate all underlying physical and material phenomenona. In this work, we present ring oscillator frequency measurement data from 65nm technology process from a high-volume manufacturing line. In the early stages of a process development cycle, there is often scarcity of rich enough spatial and temporal data to accurately model process variation. Accuracy or reliability of any estimated variation model is directly proportional to the amount of data available during the extraction phase. Our data set consisted of 331 wafers spanning 23 lots with 104 die per wafer and 14 measurements per die. Due to the proprietary nature of the data, the measured frequency, die size, and the sampling distances are presented in arbitrary units with no loss in the generality of model or the proposed procedures. All numerical data analysis was performed using MATLAB [32].

5.9.1 Wafer Selection

We begin our analysis with wafer selection. Figure 5.6 shows the distribution of raw data. It is quite clear that our data set contains outlier wafers that need to be removed. We identified

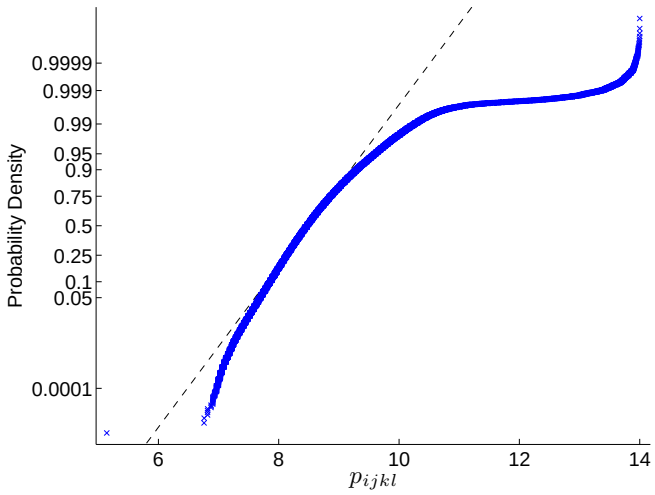


Figure 5.6: Probability density plot of raw data $\bar{p}_{ijk\cdot}$. Dotted line denotes the normal approximation of the data and it is evident that the data is not normally distributed.

outliers wafers using the Mahalanobis distance as described in section 5.5. Figure 5.7 shows the squared Mahalanobis distance plotted against the quantiles of the χ^2_n distribution. For our data, 56 outlier wafers were identified and rejected at the 95% quantile of χ^2_n distribution.

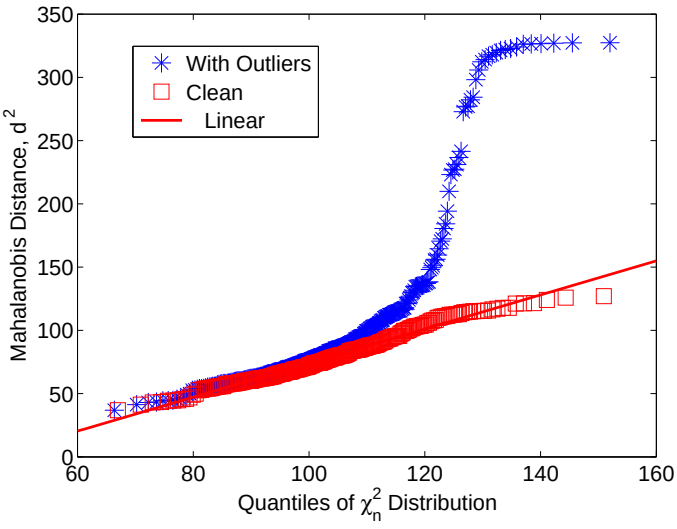


Figure 5.7: Mahalanobis distance before and after rejecting outlier wafers. Wafers were rejected at the 95% quantile of χ^2_n distribution.

Next we proceed to perform cluster analysis as described in section 5.6. In this work, we chose to perform PCA on model term ω_{ijk}. The first 10 principal components were chosen as the orthonormal basis for the LAR. LAR was performed as described in subsection 5.6.1. Figure 5.8 shows an example of results from the LAR. Figure 5.8(a) shows the C_p selection plot indicating that optimal step k_{opt} is 4. Figure 5.8(b) shows the evolution of the regression coefficients in the LAR. The x-axis represents the step number and the y-axis indicates the value of the regression coefficient (between -1 and 1). Recall that at each step in the LAR, a new predictor (coefficient) joins the active set, and at the end of the LAR algorithm, the coefficients arrive at the least-squares solution. In other words, the right most step in Figure 5.8(b) shows the least-squares solution with all 10 predictors. We also observe that the first predictor (first principal component) quickly saturates to its least-squares value. At step 4, there are 3 predictors (or coefficients) in the active set, and the 4th one about to join the active set. Similar LAR was performed on each of the 275 wafers.

Cluster analysis, as described in subsection 5.6.2, was then performed using the 275×10 regression coefficient matrix. Figure 5.9 shows a dendrogram for the 275 wafers. In our cluster analysis, we limited the maximum number of clusters to 20. Clustering algorithms discussed in subsection 5.6.2 always produce a hierarchical classification, even when it is inappropriate. One must therefore validate the results of any classification. As discussed previously, the cophenetic correlation coefficient is one measure of fidelity of the proposed hierarchical structure. For the dendrogram displayed in Figure 5.9, the cophenetic correlation coefficient was 0.84. The dendrogram shown in Figure 5.9 can be used to explore the number

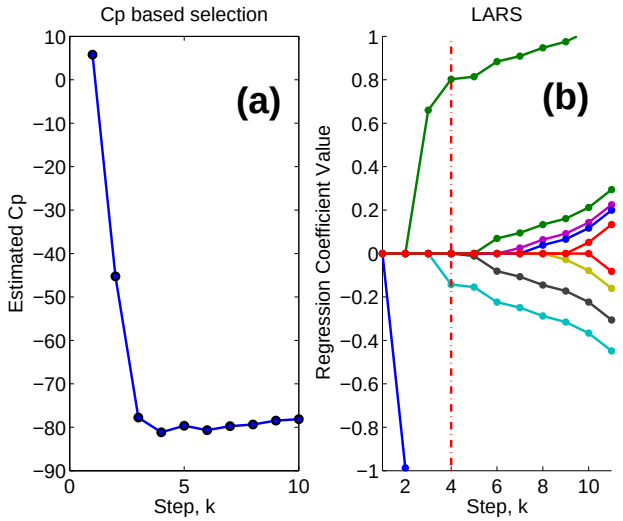


Figure 5.8: Example of Least Angle Regression: (a) C_p selection plot, and (b) evolution of regression coefficients at each LAR step. Optimal step k (which is 4 in this example) is determined by the minimum value of the C_p statistic (using (5.6.14) and (5.6.15)). The optimal value of k_{opt} determines the optimal regression coefficients. As shown in (b), at step 4 there are three coefficients selected by the C_p statistic.

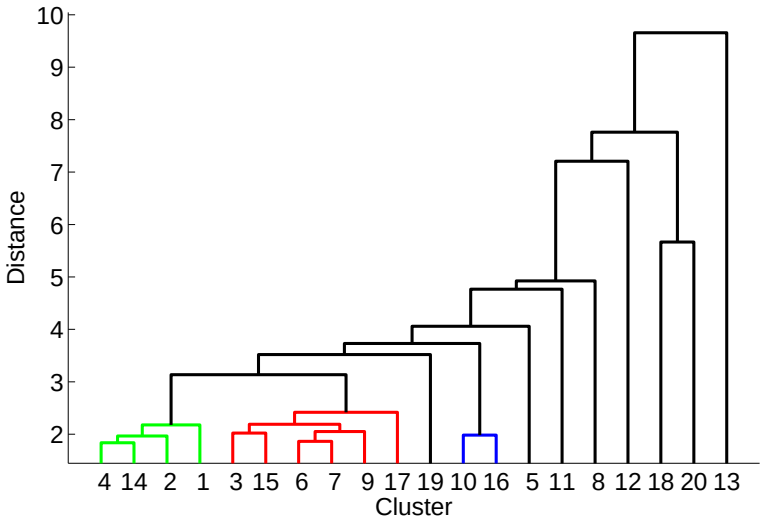
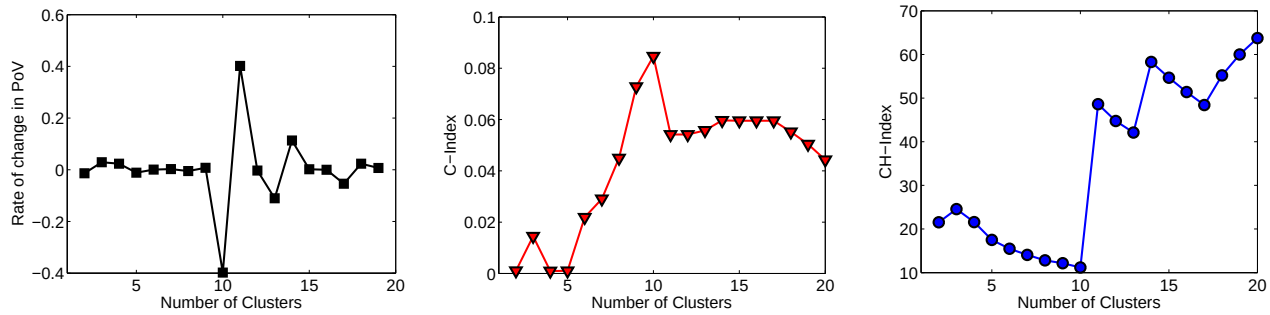


Figure 5.9: Dendrogram of hierarchy discovered in our data set. The cophenetic correlation coefficient is one measure of fidelity of the proposed hierarchical structure. The cophenetic correlation coefficient for this structure is 0.84, which indicates that this hierarchical structure is a fairly good representation of data.

of clusters present in the data. For example, if a distance of 9 was used to partition the clusters, then we would have two clusters. If a distance of 6.5 is used, then we would have four clusters, and so forth. In order to discover the optimal number of clusters (or "natural" clusters), several indices or methods can be used [93]. Here we explore three different indices: *proportion of variance*, *C-index*, and *CH-index* [93]. The *proportion of variance* (POV) is the ratio of norms of intra-cluster distance and original distances as defined by (5.6.16). The optimal number of clusters is determined when the rate of change in this ratio is maximum. It indicates the amount of order introduced by the particular cluster configuration. The *C-index* is computed as $[d_w - \min(d_w)]/[\max(d_w) - \min(SS_w)]$, where d_w is the sum of the within cluster distances. Here the minimum value indicates the optimal number of clusters. The *CH-index* is defined as $(SS_b/(k-1))/(SS_w/(M-k))$, where SS_b and SS_w are between and within the cluster sum of squares for k clusters, and M total number of objects (wafers). The optimal number of clusters is the value which maximizes this measure. However, we must examine the clusters to avoid dubious classification. Figure 5.10 shows the optimal cluster size identified by the three indices. The POV method indicates the presence of 11 clusters. Although, there was big dip in the C-index at 11 clusters, there was no global minima, and as such its results are inconclusive. Also, in case of the CH-index there was large jump in value at 11, but this was not a global maxima, and as such it too is deemed inconclusive. Indeed, the identification of the number of clusters present in a given data



(a) (b) (c)

Figure 5.10: Procedures to determine the number of clusters (a) Proportion of Variance (POV), (b) C-index and (c) CH-index. The POV method indicates the presence of 11 clusters, whereas the C-index (no minima) and the CH-index (no maxima) are inconclusive.

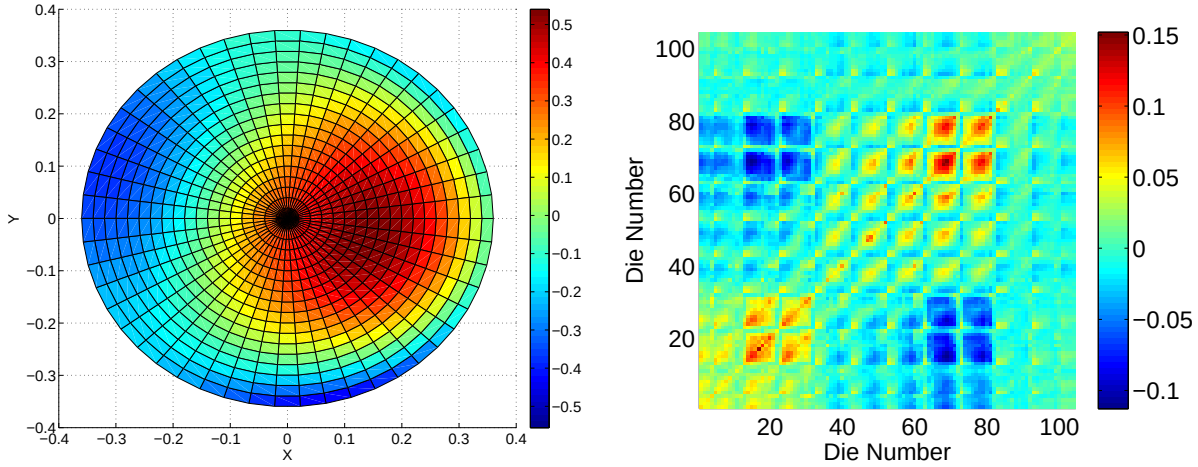
set is still largely heuristic. Milligan and Cooper [93] explored 30 different procedures to conclude that a universally prescribed method does not yet exist. For the most part, cluster analysis correctly grouped wafers with common lot number together based entirely on the similarity of spatial characteristics. It also helped correctly identify an experimental group of wafers which were previously unknown to us. Cluster analysis, such as the one described here, can be a powerful tool that can be used to explore the data set to further segregate and remove atypical wafers from model estimation process.

5.9.2 Model Estimators

For the results presented in this section, we restricted our data to a group of lots with interesting across-wafer spatial pattern. Although the selection of the group of wafers made here was somewhat arbitrarily, under other circumstances, it could have been made based on more legitimate reasons.

First, we examine the global variation component ω_{ijk}. Figure 5.11 displays the multivariate normal representation of ω_{ijk} as described in section 5.5. Figure 5.11(a) shows the spatial variation in the average wafer. The average wafer can also be represented by a polynomial fit as described in section 5.7. However, this polynomial would be intractable for an analytical treatment of spatial correlation as was proposed by [88]. Figure 5.11(b) shows the covariance of 104 dice on the wafer. Note the variation along the diagonal as well as the off-diagonal components.

Next, we estimate the die-level variation term δ_l. Figure 5.12 shows the spatial variation in the average die. Here the average die is represented by a polynomial of order 7 using procedure described in section 5.7. Note that the intra-die systematic spatial variation is significantly stronger than wafer-level systematic spatial variation observed in Figure 5.11(a).



(a) (b)

Figure 5.11: Multivariate normal representation of ω_{ijk}: (a) systematic spatial variation of average wafer ($\hat{\boldsymbol{\mu}}$), and (b) covariance matrix ($\hat{\Sigma}$)

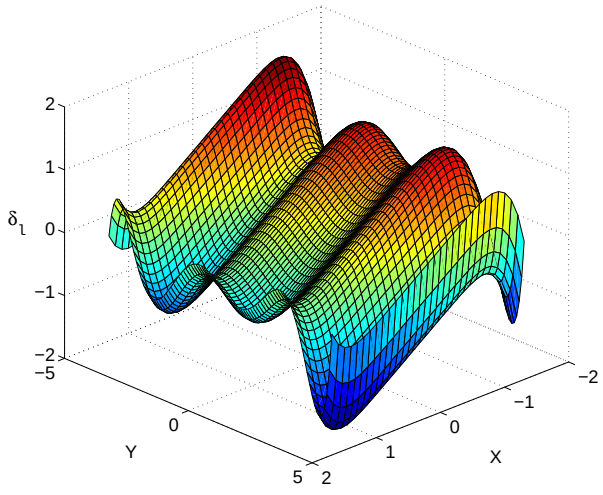


Figure 5.12: Systematic spatial variation average die corresponding to the model term δ_l in (5.4.3).

Lastly, we examine the residuals ϵ_{ijkl} from the model (5.4.3). Probability plot of ϵ_{ijkl} is shown in Figure 5.13. The distribution appears to be normal for the most part. Figure 5.14

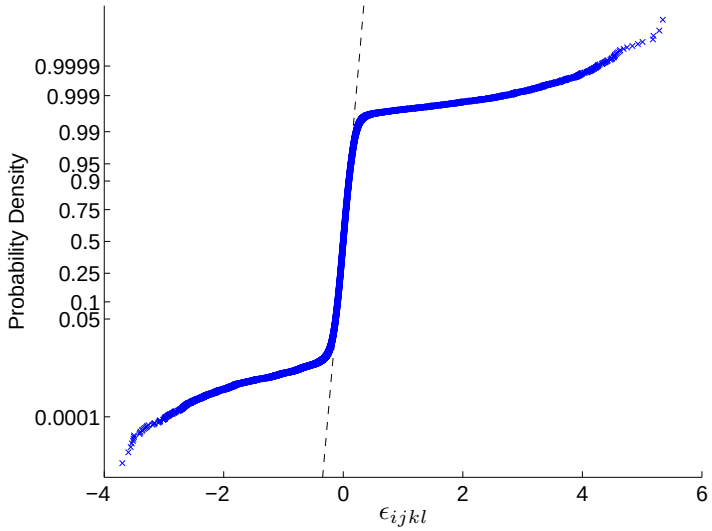


Figure 5.13: Probability plot of the residuals ϵ_{ijkl} from model (5.4.3). The distribution appears to be normal for the most part with some outlier observations.

shows the estimated variogram $\widehat{\gamma}(\boldsymbol{h})$ in X and Y directions. The variogram was estimated using the robust estimator defined by (2). Ideally, the variogram should be computed at the die level. However, in our case, since we had only 14 observations per die, the variogram was estimated at the wafer-level using technique described in subsection 5.8.2. Note the anisotropy in variogram in X and Y directions. One can observe noise at higher lags due to fewer lag pairs being available at long lag values. It is due to this noise that the parametric fit is performed at a maximum lag $h_0 = H/2$ as defined by (5.8.33). Figure 5.15 shows the parametric variogram $\gamma(\boldsymbol{h}, \theta)$ fitted using weighted-least squares technique. The fitting residuals appear to be normally distributed. Note that only the structure parameters are to be used from this fit as it was computed at the wafer level. The parametric form used in this work (see (5.8.31)) is characterized by the structure parameters ξ (correlation length), ϕ (short-range correlation factor), and the anisotropy matrix A (characterized by θ, λ, and a). The estimated values of structure parameters were $(\phi,\ \xi,\ \alpha,\ \lambda,\ a) = (0.24, 3.48, 1.56, 0.08, 38.34)$. As a reminder, correlations among transistors need to be established in two ranges: $h \ll \xi$ and $h \sim \mathcal{O}(\xi)$. For $h \gg \xi$, independence is naturally assumed. Also, recall that value of ϕ is important at very short-lags as it implies higher correlation ($\rho \sim h^{2\phi}$ for $h \ll \xi$). Note that $\theta \simeq \pi/2$ and $\lambda \ll 1$ are in line with the observed $\widehat{\gamma}(h)$ in Figure 5.14.

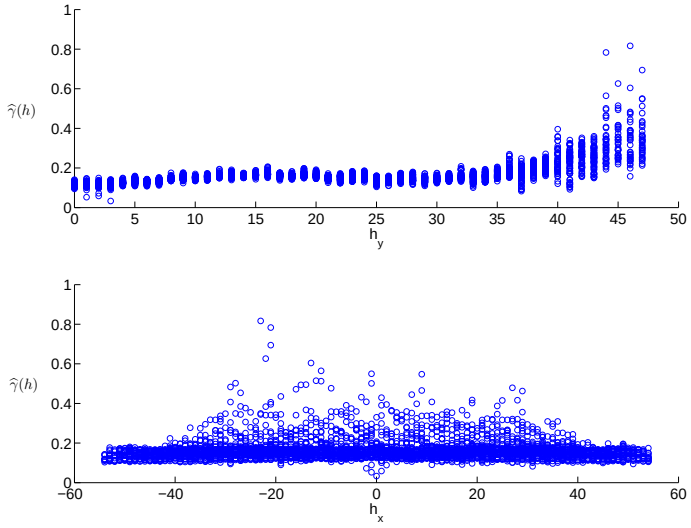
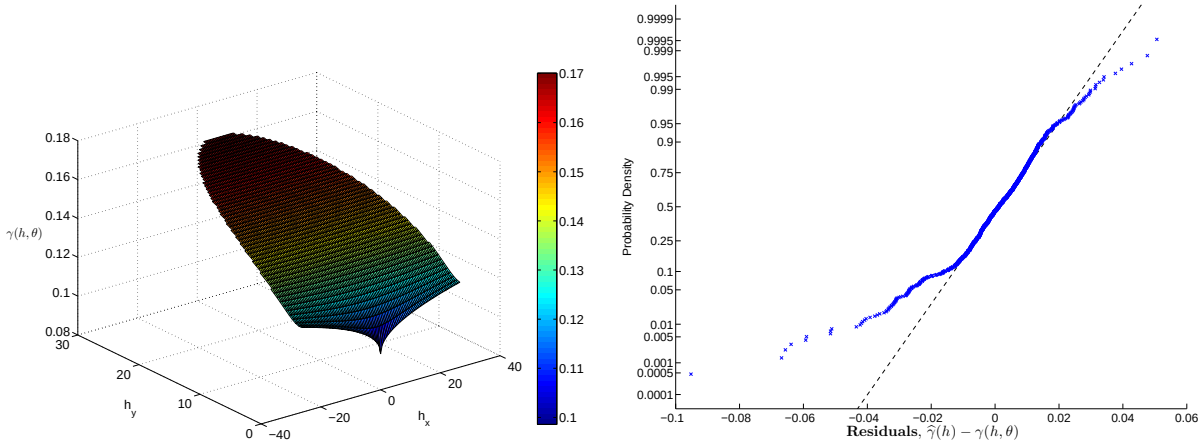


Figure 5.14: Estimated variogram $\widehat{\gamma}(\boldsymbol{h})$ in X and Y directions. The variogram was estimated using the robust estimator defined by (2). Note the anisotropy in variogram in X and Y directions. One can observe noise at higher lags due to fewer lag pairs being available at long lag values. It is due to this noise that the parametric fit is performed at a maximum lag $h_0 = H/2$ as defined by (5.8.33).



(a) (b)

Figure 5.15: (a) Parametric variogram $\gamma(\boldsymbol{h}, \theta)$ fitted using weighted-least squares and (b) Residuals of variogram fitting, $\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}, \theta)$. The fit was performed at a maximum lag h_0 defined by (5.8.33) using WLS method as described in subsection 5.8.2. The estimated values of structure parameters were $(\phi,\ \xi,\ \alpha,\ \lambda,\ a) = (0.24, 3.48, 1.56, 0.08, 38.34)$.

5.10 Summary

In this chapter, we performed a hierarchical decomposition of semiconductor process variation. We examined the across-wafer and across-die components of variation in depth. We also provided a robust method for estimating spatial correlation in the intra-die observations. With the increasing complexity of advanced high speed circuits, sophisticated simulations can utilize the information from variability models to achieve better performance. Our proposed modeling approach provides circuit designer with on-demand views of global and local variation. The multivariate normal framework proposed for wafer-level variation is capable of estimating an unknown distribution of across-wafer spatial variation. Wafer selection procedure was demonstrated using the least angle regression procedure and agglomerative hierarchical clustering. The intended use of our proposed model is as an input to Monte-Carlo circuit simulation. The extracted model enables the circuit designer to use it in either of the two desired contexts—pre-layout or post-layout. It should be noted that the extracted model may not be viable if layout of the structures used for estimation is different than the one intended in the final circuit. For instance, in terms of CMP, similar local and global pattern densities must exist in the test chip which generated the data and the final circuit. Similar argument can be made in terms of local stress. In other words, layout style and environment are of critical importance.

Chapter 6

Conclusion

In depth understanding of the sources of variability, and their influence on design parameters, is critical for design robustness, manufacturing, and yield. Process variations are inevitable and with the ever increasing process complexity, accurate models are required to identify critical areas of improvement. *The robust estimation of existing model parameters is more important than increasing the level of sophistication of the model.* Process variability is not stationary in time. Model parameters should be re-estimated periodically to ensure the soundness of assumptions and adequacy of coverage. Implicit assumptions, such as equivalency of final product design layout and the test structure design layout that generated data for estimation, must be ascertained. Other commonly made assumptions, such as *normality of data*, must also be scrutinized and validated.

6.1 Summary of Contributions

The key contributions of this work can be summarized as follows:

- In chapter 2, we developed a method for robustly estimating line width roughness (LWR) parameters. Specifically, our proposed method provides a *better* unbiased estimate of roughness amplitude σ than existing method. Our rigorous treatment also provides estimated error in LWR parameter estimates—which has sorely been missing from the existing method [21, 5, 6]. We reported a critical finding that the value of σ can be *overestimated* by a factor greater than 2, if the local variation in CD is not accounted for and removed in the estimation process. Lastly, our proposed method allows for more flexibility in capturing SEM images in that we do not need a special test structure with all lines with same designed CD; any IC layout region with straight lines and arbitrary CDs would suffice.

- It is now generally accepted that LWR is completely characterized by three parameters: σ is root-mean squared amplitude of roughness, α is the roughness exponent, and ξ is

the correlation length [5, 6]. However, generally only σ is reported in literature; ξ is reported in extremely rare studies. The least appreciated of all the parameters is α. $\alpha = 0.5$ or $\alpha = 1$ is often assumed either due to *lack of data* or *for sake of convenience*. The latter reason is due to closed form of power spectral density function for $\alpha = 0.5$ and $\alpha = 1$. In chapter 3, we utilized the robust method of estimation developed in chapter 2 to *uniformly* and *comprehensively* compare many next-generation lithography (NGL) processes in terms of their LWR characteristics. We studied the evolution of LWR parameters from resist to final substrate for each of the NGL processes. In this study, it was discovered that directed self-assembly is the most promising technology purely in terms of LWR characteristics.

- In chapter 4, we examined the impact of LWR on the device performance of double-gate FinFET. We incorporated the aforementioned LWR parameters into the FinFET framework. Our study provided useful physical insights into how the gate line edge roughness impacts the FinFET performance. It was found that the spacer-defined gate electrode (vs. a resist-defined gate electrode) provides for reduced variability in performance, indicating that gate-length mismatch has more impact than lateral offset between the front and the back gates.

- In chapter 5, we performed a hierarchical decomposition of semiconductor process variation. We provided a holistic view of variability that incorporated all aspects of variability that are important to the circuit designer, namely, global and local variation, and spatial correlation. In doing so, we discovered anisotropy in the intra-die correlation structure and highlighted the need it in future improvement of SSTA. We also proposed employing cluster analysis to avoid using atypical wafers in model estimation.

6.2 Suggested Future Work

In the robust method of estimation proposed in chapter 2, there is room for improvement in estimating the optimal block length. We used computationally simple method for block length selection as proposed by Politis and White [42]. A better method to assess the optimality of block length would be to simultaneously minimize the MSE of α, ξ, and σ.

In chapter 3, we studied self-aligned double patterning (SADP) based on the use of spacer technology. We extracted the LWR parameters at each intermediate steps by examining the *lines*. The inner and outer edges of spacers could potentially have different roughness characteristics. As such, it would be useful to study line *edge* roughness than line *width* roughness.

In chapter 4, we only investigated the impact of *gate* line edge roughness. During the formation of the fin (body) of FinFET, line edge roughness of the fin uniquely *creates* roughness in the vertical sidewall surfaces. Front and back transistors are formed on these surfaces,

and as such, it would be meaningful to study variation in the device performance to this aspect.

In chapter 5, we discovered anisotropy in the intra-die variogram. It would be interesting to perform a Monte-Carlo simulation of ISCAS'85 benchmark circuit to estimate the impact of anisotropy on timing.

Appendix A

Bias in Finite Length Variance

Consider $\mathcal{X}_n$ as defined in (2.3.1). We can rewrite (biased) sample variance (2.3.12) as

$$\widehat{\sigma}^2_{LWR} = \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}_i^2(L), \tag{A.0.1}$$

where the sample variance of i-th line is given by

$$\hat{\sigma}_i^2(L) = \frac{1}{L}\sum_{s=1}^{L}\left(X_{is} - \frac{1}{L}\sum_{t=1}^{L}X_{it}\right)^2 \tag{A.0.2}$$

$\bar{X}_i = L^{-1}\sum_{t=1}^{L}X_{it}$ is the sample average of i-th line. Note that we write $\hat{\sigma}_i^2$ as $\hat{\sigma}_i^2(L)$ to explicitly highlight its dependence on L. If this sample variance is averaged over many length L sequences, then this estimate will approach its mean. We can find an expression for the mean of $\hat{\sigma}^2$ as a function of L from (A.0.2). We begin by adding and subtracting the population mean inside the square,

$$\begin{aligned}\hat{\sigma}_i^2(L) &= \frac{1}{L}\sum_{s=1}^{L}\left[(X_{is}-\mu_i) - \frac{1}{L}\sum_{t=1}^{L}(X_{it}-\mu_i)\right]^2 \\ &= \frac{1}{L}\sum_{s=1}^{L}(X_{is}-\mu_i)^2 - \left(\bar{X}_i - \mu_i\right)^2.\end{aligned}$$

Now using the identity

$$E\left[\left(\sum_{i=1}^{N}(Y_i - EY_i)\right)^2\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}E[(Y_i-EY_i)(Y_j-EY_j)], \tag{A.0.3}$$

and the stationarity property of $\mathcal{X}_n$ (2.3.2), we get

$$\begin{aligned} E\left[\hat{\sigma}_i^2(L)\right] &= \frac{1}{L}\sum_{s=1}^{L} E\left[(X_{is}-\mu_i)^2\right] - \frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L} E\left[(X_{is}-\mu_i)(X_{it}-\mu_i)\right] \\ &= \sigma^2\left(1-\frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L}\rho(s-t)\right). \end{aligned} \tag{A.0.4}$$

Finally, taking the expectation of (A.0.1) and substituting (A.0.4) in it, we get

$$E[\widehat{\sigma}_{LWR}^2(L)] = \sigma^2\left(1-\frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L}\rho(s-t)\right). \tag{A.0.5}$$

We can capture the length dependence in (A.0.5) by defining

$$f(L) \equiv \left(1-\frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L}\rho(s-t)\right). \tag{A.0.6}$$

Thus, we have shown that $E[\widehat{\sigma}_{LWR}^2(L)] = \sigma^2 f(L)$.

For the assumed form of the auto-correlation function given by (2.2.9), it is possible to derive a closed form expression for $f(L)$ for $\alpha = 0.5$ and $\alpha = 1$. We will simply state the result below:

$$f(L) = \begin{cases} 1-\frac{2\xi}{L^2}\left[L+\xi\left(\exp(-\frac{L}{\xi})-1\right)\right] & \text{for } \alpha = 0.5 \\ 1-\frac{\xi}{L^2}\left[L\sqrt{\pi}\,\mathrm{erf}(\frac{L}{\xi})+\xi\left(\exp(-\frac{L^2}{\xi^2})-1\right)\right] & \text{for } \alpha = 1 \end{cases}. \tag{A.0.7}$$

In A.0.7, $\mathrm{erf}(\cdot)$ indicates the error function [98]. Note that in (A.0.7), the length of line L is used as a physical quantity and not as number of grid points at regular spacing.

Appendix B

Validity of LLE Approach

Consider $\mathcal{X}_n$ as defined in (2.3.1) and let the variance in CD be estimated by (2.3.11). Let $Z_{is} = X_{is} - \mu_i$, $\bar{Z}_i = L^{-1}\sum_{s=1}^{L} Z_{is}$, and $\bar{Z} = (ML)^{-1}\sum_{i=1}^{M}\sum_{s=1}^{L} Z_{is}$. Note that $\bar{Z} = M^{-1}\sum_{i=1}^{M}\bar{Z}_i$ and that the $\bar{Z}_i$'s are independent. Hence, assuming $\mu_1 = \ldots = \mu_M$, we have,

$$\begin{aligned}
E\left[\hat{\sigma}^2_{CD}(L)\right] &= \frac{1}{M-1}\sum_{i=1}^{M} E\left[\bar{Z}_i - \bar{Z}\right]^2 \\
&= \frac{1}{M-1}\sum_{i=1}^{M} E\bar{Z}_i^2 - E(\bar{Z}^2) \\
&= \frac{M}{M-1}E\bar{Z}_1^2 - \frac{1}{M-1}E\bar{Z}_1^2 \quad \text{(by independence)} \\
&= E\bar{Z}_1^2 \\
&= \frac{1}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L} E[(X_{1s}-\mu_1)(X_{1t}-\mu_1)].
\end{aligned}$$

Substituting (2.3.2) in the above expression, we get,

$$E\left[\hat{\sigma}^2_{CD}(L)\right] = \frac{\sigma^2}{L^2}\sum_{s=1}^{L}\sum_{t=1}^{L}\rho(s-t). \tag{B.0.1}$$

Finally, adding (A.0.5) and (B.0.1), we get

$$E\left[\widehat{\sigma}^2_{LWR}(L) + \hat{\sigma}^2_{CD}(L)\right] = \sigma^2. \tag{B.0.2}$$

Note that the above result is valid only as long as $\hat{\sigma}^2_{CD}$ term is not influenced by non-LER sources of variation. In reality, some residual local CD (systematic or local) variation always exists and as such $\hat{\sigma}^2_{CD}$ term represents the sum of bias correction term and non-LER variation term.

Bibliography

[1] "International Technology Roadmap for Semiconductors," http://www.itrs.net, 2009.

[2] G. Moore, "Progress in digital integrated electronics," in *Electron Devices Meeting, 1975 International*, 1975, pp. 11–13.

[3] K. Kuhn, "Moore's Law Past 32nm: Future Challenges in Device Scaling," in *Proc. of Intl. Workshop on Computational Electronics*, 2009, pp. 1–6.

[4] A. Asenov, "Statistical nano cmos variability and its impact on sram," in *Extreme Statistics in Nanoscale Memory Design*, ser. Integrated Circuits and Systems, A. Singhee and R. Rutenbar, Eds. New York: Springer, 2010, ch. 3, pp. 17–49.

[5] V. Constantoudis, G. P. Patsis, A. Tserepi, and E. Gogolides, "Quantification of line-edge roughness of photoresists. ii. scaling and fractal analysis and the best roughness descriptors," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 21, no. 3, pp. 1019–1026, 2003.

[6] V. Constantoudis, G. P. Patsis, L. H. A. Leunissen, and E. Gogolides, "Line edge roughness and critical dimension variation: Fractal characterization and comparison using model functions," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 22, no. 4, pp. 1974–1981, 2004.

[7] S. K. Sinha, E. B. Sirota, S. Garoff, and H. B. Stanley, "X-ray and neutron scattering from rough surfaces," *Phys. Rev. B*, vol. 38, no. 4, pp. 2297–2311, Aug 1988.

[8] K. Bowman, S. Duvall, and J. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, feb. 2002.

[9] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *DAC '04: Proceedings of the 41st annual Design Automation Conference*. New York, NY, USA: ACM, 2004, pp. 331–336.

[10] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single pert-like traversal," in *ICCAD '03: Proceedings of the 2003 IEEE/ACM international conference on Computer-Aided design*. Washington, DC, USA: IEEE Computer Society, 2003, p. 621.

[11] M. Orshansky, L. Milor, P. Chen, K. Keutzer, and C. Hu, "Impact of systematic spatial intra-chip gate length variability on performance of high-speed digital circuits," in *ICCAD '00: Proceedings of the 2000 IEEE/ACM international conference on Computer-aided design*. Piscataway, NJ, USA: IEEE Press, 2000, pp. 62–67.

[12] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *ICCAD '03: Proceedings of the 2003 IEEE/ACM international conference on Computer-aided design*. Washington, DC, USA: IEEE Computer Society, 2003, p. 900.

[13] F. N. Najm, N. Menezes, and I. A. Ferzli, "A yield model for integrated circuits and its application to statistical timing analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 3, pp. 574–591, mar. 2007.

[14] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sep 2003.

[15] "International Technology Roadmap for Semiconductors," http://www.itrs.net, 2008.

[16] T. Yoshimura, H. Shiraishi, J. Yamamoto, and S. Okazaki, "Nano edge roughness in polymer resist patterns," *Applied Physics Letters*, vol. 63, no. 6, pp. 764–766, Aug. 1993.

[17] T. Yamaguchi, K. Yamazaki, M. Nagase, and H. Namatsu, "Line-edge roughness: Characterization and material origin," *Japanese Journal of Applied Physics*, vol. 42, no. Part 1, No. 6B, pp. 3755–3762, 2003.

[18] K. Patel, T.-J. K. Liu, and C. Spanos, "Gate line edge roughness model for estimation of finfet performance variability," *Electron Devices, IEEE Transactions on*, vol. 56, no. 12, pp. 3055–3063, Dec 2009.

[19] S.-D. Kim, H. Wada, and J. Woo, "TCAD-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale MOS transistor performance and scaling," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 2, pp. 192–200, 2004.

[20] C. Diaz, H.-J. Tao, Y.-C. Ku, A. Yen, and K. Young, "An experimentally validated analytical model for gate line-edge roughness (LER) effects on technology scaling," *IEEE Electron Device Letters*, vol. 22, no. 6, pp. 287–289, 2001.

[21] L. H. A. Leunissen, W. G. Lawrence, and M. Ercken, "Line edge roughness: experimental results related to a two-parameter model," *Microelectron. Eng.*, vol. 73-74, no. 1, pp. 265–270, 2004.

[22] G. Wakamatsu, Y. Anno, M. Hori, T. Kakizawa, M. Mita, K. Hoshiko, T. Shioya, K. Fujiwara, S. Kusumoto, Y. Yamaguchi, and T. Shimokawa, "Double patterning process with freezing technique," in *Advances in Resist Materials and Processing Technology XXVI*, C. L. Henderson, Ed. SPIE, 2009, p. 72730B.

[23] M. Pelliccione and T.-M. Lu, *Evolution of Thin Film Morphology: Modeling and Simulations*, ser. Springer Series in Materials Science. New York: Springer, 2007, vol. 108.

[24] Y. Zhao, G.-C. Wang, and T.-M. Lu, *Characterization of Amorphous and Crystalline Rough Surface: Principles and Applications*, ser. Experimental Methods in the Physical Sciences. San Diego: Academic Press, October 2000, vol. 37.

[25] G. Palasantzas, "Roughness spectrum and surface width of self-affine fractal surfaces via the k-correlation model," *Phys. Rev. B*, vol. 48, no. 19, pp. 14 472–14 478, Nov 1993.

[26] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[27] N. Cressie, "Fitting variogram models by weighted least squares," *Mathematical Geology*, vol. 17, no. 5, pp. 563–586, Jul. 1985.

[28] N. A. C. Cressie, *Statistics for Spatial Data*, 1st ed. New York: Wiley-Interscience, January 1993.

[29] D. L. Zimmerman and M. B. Zimmerman, "A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors," *Technometrics*, vol. 33, pp. 77–91, 1991.

[30] S. N. Lahiri, *Resampling Methods for Dependent Data*, 1st ed. Springer, August 2003.

[31] Y. D. Lee and S. N. Lahiri, "Least squares variogram fitting by spatial subsampling," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 837–854, 2002.

[32] "The MathWorks Inc." http://www.mathworks.com.

[33] C. R. Rao, *Linear Statistical Inference and its Applications*, 2nd ed., ser. Wiley Series in Probability and Mathematical Statistics. New York: Wiley-Interscience, December 2001.

[34] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, January 1979.

[35] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, October 1997.

[36] K. Singh, "On the asymptotic accuracy of efron's bootstrap," *The Annals of Statistics*, vol. 9, no. 6, pp. 1187–1195, 1981.

[37] E. Carlstein, "The use of subseries values for estimating the variance of a general statistic from a stationary sequence," *The Annals of Statistics*, vol. 14, no. 3, pp. 1171–1179, 1986.

[38] H. R. Kunsch, "The jackknife and the bootstrap for general stationary observations," *The Annals of Statistics*, vol. 17, no. 3, pp. 1217–1241, 1989.

[39] R. Y. Liu and K. Singh, "Moving blocks jackknife and bootstrap capture weak dependence," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds. New York: John Wiley, 1992, pp. 225–248.

[40] D. N. Politis and J. P. Romano, "A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation," *The Annals of Statistics*, vol. 20, no. 4, pp. 1985–2007, 1992.

[41] P. Hall, J. L. Horowitz, and B. Y. Jing, "On blocking rules for the bootstrap with dependent data," *Biometrika*, vol. 82, no. 3, pp. 561–574, 1995.

[42] D. N. Politis and H. White, "Automatic block-length selection for the dependent bootstrap," *Econometric Reviews*, vol. 23, no. 1, pp. 53–70, 2004.

[43] A. R. Pawloski, A. Acheta, S. Bell, B. L. Fontaine, T. Wallow, and H. J. Levinson, "The transfer of photoresist ler through etch," in *Advances in Resist Technology and Processing XXIII*, Q. Lin, Ed. SPIE, 2006, p. 615318.

[44] T. Wallow, A. Acheta, Y. Ma, A. Pawloski, S. Bell, B. Ward, C. Tabery, B. L. Fontaine, R. han Kim, S. McGowan, and H. J. Levinson, "Line-edge roughness in 193-nm resists: lithographic aspects and etch transfer," in *Advances in Resist Materials and Processing Technology XXIV*, Q. Lin, Ed. SPIE, 2007, p. 651919.

[45] G. P. Patsis, V. Constantoudis, A. Tserepi, E. Gogolides, and G. Grozev, "Quantification of line-edge roughness of photoresists. i. a comparison between off-line and on-line analysis of top-down scanning electron microscopy images," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 21, no. 3, pp. 1008–1018, 2003.

[46] P. P. Naulleau and J. P. Cain, "Experimental and model-based study of the robustness of line-edge roughness metric extraction in the presence of noise," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 25, pp. 1647–1657, 2007.

[47] B. D. Bunday, M. Bishop, J. S. Villarrubia, and A. E. Vladar, "Cd-sem measurement line-edge roughness test patterns for 193-nm lithography," in *Metrology, Inspection, and Process Control for Microlithography XVII*, D. J. Herr, Ed. SPIE, 2003, pp. 674–688.

[48] A. Yamaguchi, R. Steffen, J. Yamamoto, H. Kawada, and T. Iizumi, "Single-shot method for bias-free ler/lwr evaluation with little damage," *Microelectron. Eng.*, vol. 84, no. 5-8, pp. 1779–1782, 2007.

[49] "EUV Technology," http://www.euvl.com/summit/index.html.

[50] G. M. Gallatin, "Resist blur and line edge roughness," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, B. W. Smith, Ed., vol. 5754, May 2005, pp. 38–52.

[51] A. H. Lettington and Q. H. Hong, "Image restoration using a lorentzian probability model," *Journal of Modern Optics*, vol. 42, no. 7, pp. 1367–1376, Jul. 1995.

[52] H. Namatsu, M. Nagase, T. Yamaguchi, K. Yamazaki, and K. Kurihara, "Influence of edge roughness in resist patterns on etched patterns," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 16, no. 6, pp. 3315–3321, 1998.

[53] Y. C. Bae, Y. Liu, T. Cardolaccia, J. C. McDermott, P. Trefonas, K. Spizuoco, M. Reilly, A. Pikon, L. Joesten, G. G. Zhang, G. G. Barclay, J. Simon, and S. Gaurigan, "Materials for single-etch double patterning process: surface curing agent and thermal cure resist," in *Advances in Resist Materials and Processing Technology XXVI*, C. L. Henderson, Ed. SPIE, 2009, p. 727306.

[54] C. Bencher, Y. Chen, H. Dai, W. Montgomery, and L. Huli, "22nm half-pitch patterning by cvd spacer self alignment double patterning (sadp)," in *Optical Microlithography XXI*, H. J. Levinson and M. V. Dusa, Eds. SPIE, 2008, p. 69244E.

[55] M. P. Stoykovich and P. F. Nealey, "Block copolymers and conventional lithography," *Materials Today*, vol. 9, no. 9, pp. 20–29, 2006.

[56] C. B. Brooks, D. L. LaBrake, and N. Khusnatdinov, "Etching of 42-nm and 32-nm half-pitch features patterned using step and flash imprint lithography," in *Emerging Lithographic Technologies XII*, F. M. Schellenberg, Ed. SPIE, 2008, p. 69211K.

[57] S. O'uchi, T. Matsukawa, T. Nakagawa, K. Endo, Y. Liu, T. Sekigawa, J. Tsukada, Y. Ishikawa, H. Yamauchi, K. Ishii, E. Suzuki, H. Koike, K. Sakamoto, and M. Masahara, "Characterization of metal-gate finfet variability based on measurements and compact model analyses," in *Electron Devices Meeting, 2008. IEDM 2008. Digest International*, dec 2008, pp. 1–4.

[58] N. Lindert, L. Chang, Y.-K. Choi, E. Anderson, W.-C. Lee, T.-J. King, J. Bokor, and C. Hu, "Sub-60-nm quasi-planar finfets fabricated using a simplified process," *Electron Device Letters, IEEE*, vol. 22, no. 10, pp. 487–489, Oct. 2001.

[59] C. Gustin, L. Leunissen, A. Mercha, S. Decoutere, and G. Lorusso, "Impact of line width roughness on the matching performances of next-generation devices," *Thin Solid Films*, vol. 516, no. 11, pp. 3690–3696, 2008.

[60] "International Technology Roadmap for Semiconductors," http://www.itrs.net, 2007.

[61] A. Asenov, S. Kaya, and A. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1254–1260, May 2003.

[62] E. Baravelli, M. Jurczak, N. Speciale, K. De Meyer, and A. Dixit, "Impact of ler and random dopant fluctuations on finfet matching performance," *Nanotechnology, IEEE Transactions on*, vol. 7, no. 3, pp. 291–298, May 2008.

[63] V. Constantoudis, E. Gogolides, J. Roberts, and J. K. Stowers, "Characterization and modeling of line width roughness (LWR)," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, R. M. Silver, Ed., vol. 5752, May 2005, pp. 1227–1236.

[64] J. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, and H. Maes, "Line edge roughness: characterization, modeling and impact on device behavior," in *Electron Devices Meeting, 2002. IEDM 2002. Digest. International*, 2002, pp. 307–310.

[65] L. Leunissen, M. Ercken, M. Goethals, S. Locorotondo, and K. Ronse, "Transfer of line edge roughness during gate patterning processes," in *Proc. Int. Symp. Dry Process*, Sept. 2004, pp. 1–6.

[66] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," in *Simulation of Semiconductor Processes and Devices, 2000. SISPAD 2000. 2000 International Conference on*, 2000, pp. 131–134.

[67] L. Sponton, L. Bomholt, D. Pramanik, and W. Fichtner, "A full 3d tcad simulation study of line-width roughness effects in 65 nm technology," in *Simulation of Semiconductor Processes and Devices, 2006 International Conference on*, sep 2006, pp. 377–380.

[68] Y.-K. Choi, D. Ha, E. Snow, J. Bokor, and T.-J. King, "Reliability study of cmos finfets," in *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International*, dec 2003, pp. 7.6.1–7.6.4.

[69] W. Xiong, G. Gebara, J. Zaman, M. Gostkowski, B. Nguyen, G. Smith, D. Lewis, C. Cleavelin, R. Wise, S. Yu, M. Pas, T.-J. King, and J. Colinge, "Improvement of finfet electrical characteristics by hydrogen annealing," *Electron Device Letters, IEEE*, vol. 25, no. 8, pp. 541 – 543, Aug. 2004.

[70] L. Ge and J. Fossum, "Analytical modeling of quantization and volume inversion in thin si-film dg mosfets," *Electron Devices, IEEE Transactions on*, vol. 49, no. 2, pp. 287–294, Feb. 2002.

[71] J. Fossum, L. Wang, J. Yang, S. Kim, and V. Trivedi, "Pragmatic design of nanoscale multi-gate cmos," in *Electron Devices Meeting, 2004. IEDM '04 Technical Digest. IEEE International*, Dec. 2004, pp. 613–616.

[72] X. Wu, P. Chan, and M. Chan, "Impact of non-vertical sidewall on sub-50 nm finfet," in *SOI Conference, 2003. IEEE International*, Sep. 2003, pp. 151–152.

[73] Y. Li and W.-H. Chen, "Simulation of nanoscale round-top-gate bulk finfets with optimal geometry aspect ratio," in *Nanotechnology, 2006. IEEE-NANO 2006. Sixth IEEE Conference on*, vol. 2, Jun. 2006, pp. 569–572.

[74] J.-W. Yang, P. Zeitzoff, and H.-H. Tseng, "Highly manufacturable double-gate finfet with gate-source/drain underlap," *Electron Devices, IEEE Transactions on*, vol. 54, no. 6, pp. 1464–1470, Jun. 2007.

[75] S. Balasubramanian, B. Nikolic, and T. jae King, "Circuit-performance implications for double-gate mosfet scaling below 25nm," in *Proceedings of the 2003 Silicon Nanoelectronics Workshop*, 2003, pp. 16–17.

[76] "Sentaurus TCAD User Manual (Version Z-2007.03)," http://www.synopsys.com.

[77] R. Granzner, V. Polyakov, F. Schwierz, M. Kittler, R. Luyken, W. Rsner, and M. Stdele, "Simulation of nanoscale mosfets using modified drift-diffusion and hydrodynamic models and comparison with monte carlo results," *Microelectronic Engineering*, vol. 83, no. 2, pp. 241–246, 2006.

[78] O. M. Nayfeh and D. A. Antoniadis, "Calibrated hydrodynamic simulation of deeply-scaled well-tempered nanowire field effect transistors," in *Simulation of Semiconductor Processes and Devices 2007*, T. Grasser and S. Selberherr, Eds. Springer Vienna, 2007, pp. 305–308.

[79] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics For Experimenters*, 1st ed. New York, NY, USA: John Wiley and Sons, 1978.

[80] V. Constantoudis, G. P. Patsis, and E. Gogolides, "Correlation length and the problem of line width roughness," in *Metrology, Inspection, and Process Control for Microlithography XXI*. SPIE, 2007, p. 65181N.

[81] B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, no. 1, pp. 24–41, Feb. 1997.

[82] S. Reda and S. Nassif, "Analyzing the impact of process variations on parametric measurements: Novel models and applications," in *Design, Automation Test in Europe Conference Exhibition, 2009. DATE '09.*, Apr. 2009, pp. 375–380.

[83] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 4, pp. 619–631, Apr. 2007.

[84] T. Sato, H. Ueyama, N. Nakayama, and K. Masu, "Determination of optimal polynomial regression function to decompose on-die systematic and random variations," in *ASP-DAC '08: Proceedings of the 2008 Asia and South Pacific Design Automation Conference*. Los Alamitos, CA, USA: IEEE Computer Society Press, 2008, pp. 518–523.

[85] K. Chopra, N. Shenoy, and D. Blaauw, "Variogram based robust extraction of process variation," in *Proceedings of ACM/IEEE International Workshop on Timing Issues*, 2007, pp. 112–117.

[86] F. Liu, "A general framework for spatial correlation modeling in vlsi design," in *DAC'07: Proceedings of the 44th Annual Design Automation Conference*. New York, NY, USA: ACM, 2007, pp. 817–822.

[87] D. Boning and S. Nassif, "Models of process variations in device and interconnect," in *Design of High Performance Microprocessor Circuits*. IEEE Press, 2000.

[88] L. Cheng, P. Gupta, C. Spanos, K. Qian, and L. He, "Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability," in *DAC '09: Proceedings of the 46th Annual Design Automation Conference*. New York, NY, USA: ACM, 2009, pp. 104–109.

[89] D. F. Morrison, *Multivariate Statistical Methods*, 4th ed. New York, NY, USA: Duxbury Press, 2004.

[90] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[91] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2003.

[92] A. D. Gordon, "A review of hierarchical classification," *Journal of the Royal Statistical Society. Series A*, vol. 150, no. 2, pp. 119–137, 1987.

[93] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, jun 1985.

[94] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," in *ISQED '05: Proceedings of the 6th International Symposium on Quality of Electronic Design*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 516–521.

[95] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer, 2006.

[96] N. Cressie and D. M. Hawkins, "Robust estimation of the variogram," *Mathematical Geology*, vol. 12, no. 2, pp. 115–125, 1980.

[97] E. H. Isaaks and R. M. Srivastava, *Introduction to Geostatistics*, 1st ed. New York: Oxford University Press, January 1990.

[98] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae, 31st Edition*, 31st ed. Chapman and Hall/CRC, November 2002.