A Hierarchical Model for Characterizing Spatial Wafer Variations

Lulu Bao[a], Kaibo Wang[a*] and Ran Jin[b]

[a]Department of Industrial Engineering, Tsinghua University, Beijing, China;

[b]Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Silicon wafers are commonly used materials in semiconductor manufacturing industry. Their geometric quality directly affects the production cost and yield. Therefore, it is critical to improve the quality of wafers to meet today's competitive market needs. The conventional summary metrics, such as total thickness variation (TTV), bow and warp, cannot fully reflect the local variability within each wafer, nor provide useful insight for root cause diagnosis and quality improvement. The advancement of sensing technology enables two-dimensional (2-D) data map to characterize the geometric shape of wafers, which provides more information than the summary metrics. The objective of this research is to develop a statistical model to characterize the thickness variation of wafers based on the 2-D data maps. Specifically, the thickness variation of wafers is decomposed into the macro-scale variation and the micro-scale variation, which are modeled respectively as a cubic curve and a first-order intrinsic Gaussian Markov random field. The model can successfully capture both the macro-scale mean trend and the micro-scale local variation, and gives important engineering implications for process monitoring, fault diagnosis and run-to-run control. A real case study from a wafer manufacturing process is performed to show the effectiveness of the proposed methodology.

Keywords: Gaussian Markov Random Field; Hierarchical Model; Spatial Data

1. Introduction

Silicon wafers are widely used as the substrate in a variety of high-tech industries, including integrated circuits (ICs), sensor devices, micro-electro mechanical systems, optoelectronic components, and solar cells. Because small-scale devices and structures will be fabricated on the wafer substrate, it is essential for the wafer to be smooth and flat on both the macro and micro scales (O'Mara *et al.* (1990), Shen *et al.* (2006)). In other words, smaller surface variations result in better quality of wafers. Large surface variations are likely to result in defects, such as poor thickness uniformity, cracks, and breakage, which may further lead to defects in the final products. Additionally, poor surface quality may limit the enlargement of wafer diameters, which is expected to dramatically reduce the cost of IC fabrication by placing more chips on a wafer and thereby invoking the benefits from economies of scale (Quirk and Serda (2001), Orton (2009)). Therefore, the ability to improve the wafer quality is a critical objective in modern semiconductor manufacturing.

In general, the geometric quality metrics of a wafer are required to be smooth, uniform, and flat. The metrics that are commonly adopted by the semiconductor industry include center thickness, total thickness variation (TTV), bow, warp, total indicator reading (TIR), the maximum focal plane deviation (FPD) and etc.. Shen *et al.* (2006) used relative peak displacement and valley displacement to measure the waviness of a wafer. Fan (2000) used within-wafer non-uniformity to measure the quality of finished wafers of a chemical-mechanical planarization process. Essentially, all of these metrics are summary statistics; each of them is calculated based on multiple measures on a wafer. These metrics are then used to represent the quality of the wafer. For example, TTV is defined as the difference between the thickest and the thinnest thickness over several thousand measurement points on a wafer. A large TTV value indicates poor thickness uniformity. For another example, bow is

defined as the distance between the center and a fitted reference plane of wafers (see O'Mara *et al.* (2007) for the definition of these parameters), which measures the wafer flatness. However, such simple summary statistics are not adequate to characterize wafers because they overlook the rich spatial information of the wafers.

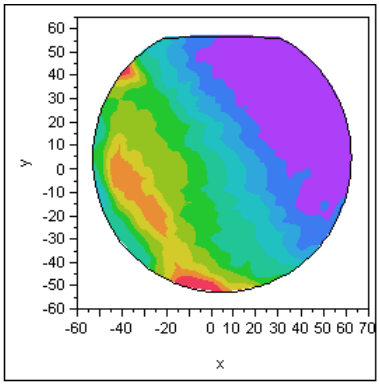
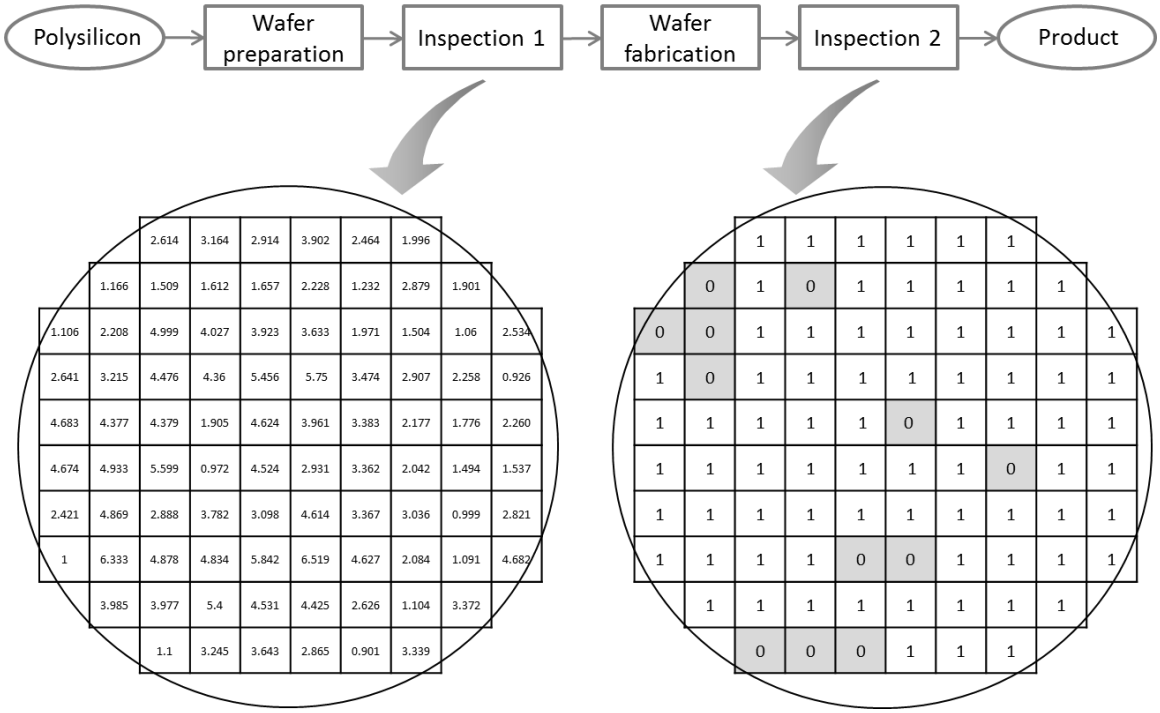


Figure 1: The heat map of a real wafer example

Figure 1 shows the thickness of a wafer in a heat map. Different colors represent the different thickness values of the wafer. From this example, we can see that the map contains rich information about the wafer: a macro-scale thickness trend and a micro-scale thickness similarity within adjacent areas are observed. A thorough analysis of the variation patterns will be given in Section 3. Based on these observations, we will develop a model that can capture the multi-scale variation patterns of the wafers.

The statistical analysis based on the spatial pattern of wafers has a major advantage over the analysis based on the summary metrics. This is because the spatial pattern gives much more rich information than the metrics. For example, the thickness of the wafer shown in Figure 1 changes sharply within a small edge area and is relatively flat throughout other large region. Such variation pattern cannot be explicitly reflected by TTV, since two wafers with the same TTV may exhibit totally different local thickness patterns. Instead, statistical analysis based on the rich spatial data can help

link the variation pattern with the engineering process, and thus the result can provide hints for quality improvement initiatives.



(a) 2-D continuous data (b) 2-D discrete data

Figure 2: Two types of data collected from a wafer manufacturing process

The purpose of the statistical analysis is to construct a quantitative model to characterize surface variation, which is important in many quality control applications. For example, unexpected process failures usually lead to changes of the surface quality. Therefore, the model for wafer variation patterns can be used to characterize wafers and develop control charts for process monitoring. For another example, the manufacturing process variation can be reduced via run-to-run (R2R) process control; a model that can characterize and predict wafer quality is also needed in designing the controller.

There are two types of quality inspection data that are available to represent the spatial patterns of wafers from a wafer manufacturing process: the 2-D continuous data, and the 2-D discrete data.

These two types of data are usually collected after the wafer preparation and the wafer fabrication processes, respectively. In a wafer manufacturing process as shown in Figure 2, single-crystal silicon material is used to produce the wafer substrate through the wafer preparation steps, such as slicing, lapping, chemical vapor deposition (CVD), polishing and testing. Then the wafer fabrication process continues to create devices and structures on the top of a wafer substrate in multiple steps, such as photolithography, etching, doping, chemical mechanical planarization (CMP), and die testing. At the end of the wafer preparation processes, continuous data is used to measure the spatial variability of wafers, where the data format is shown in Figure 2 (a). And at the end of the wafer fabrication process, binary data are used to indicate the functionality of devices on the wafer, where the data format is shown in Figure 2 (b).

Although geometrically similar, the continuous data contain more information than the binary data. Additionally, some of the defects shown on the binary map are caused by the spatial variation of wafer substrates. Therefore, we focus on the modeling and analysis of the continuous 2-D wafer data map in this paper.

The rest part of the paper is organized as follows. Due to the unique spatial data structure obtained for wafer thickness analysis, a review and comparison of existing methods for modeling surface variation is presented in Section 2. In Section 3, we analyze real wafer examples and provide insights for statistical modeling. A hierarchical model framework is then introduced for the wafer variation modeling in Section 4, followed by a case study based on real data in Section 5. Finally, we discuss the practical implications of the model in Section 6 and draw conclusions in Section 7.

2. The state-of-the-art for modeling surface variation

In the literature, process monitoring, defect modeling or wafer clustering based on 2-D maps have

received abundant attentions. Since both the binary data types share similar spatial structures, the literature on the modeling of both discrete and continuous data is reviewed in the following.

2.1. Modeling the 2-D discrete data

The discrete data acquired after the wafer fabrication process are mainly used for process monitoring and root causes diagnosis. Early research targeted on detecting the existence of spatial clustering defects on wafers. As Albin and Friedman (1989) noted, the Poisson distribution is generally assumed when defect data is encountered. However, defects tend to appear in clusters due to the spatial structure from the wafer manufacturing process. Albin and Friedman (1989) proposed the Neyman Type A distribution, which is more suitable for modeling the number of defects on individual wafers.

The modeling and analysis of the binary 2-D wafer map is also important to diagnosis the faults in the wafer fabrication process. Hansen *et al.* (1997) classified the defect patterns on wafers into two categories: random defects and clusters. A Markov random field is used to characterize the clustered defects. A similar approach was used by Hwang and Kuo (2007) in a model-based clustering strategy, which treats the observed defects as the composition of both global defects resulting from random causes, and local defects resulting from assignable causes. A spatial non-homogeneous Poisson process is used to model the global defects. A bivariate normal distribution or a principal curve is then used to model the local defects. A model-based clustering algorithm is further applied to identify the characteristics of local defect clusters. Liu *et al.* (2002) proposed to use a neural network to extract patterns from the 2-D wafer binary map. Wang *et al.* (2006) also focused on the identification of spatial defect patterns on wafer. Hsu and Chien (2007) proposed a data-mining framework to identify the failure patterns and classify wafers into different clusters based

on the 2-D binary wafer map.

Overall, the work on the monitoring, modeling and analysis of failure patterns based on the binary 2-D wafer map is an effective approach for studying the fabrication process. The knowledge learned from these analyses is valuable for quality improvement and yield enhancement in real practice. However, these approaches are used for discrete data rather than continuous data.

2.2. Modeling the 2-D continuous data

In the wafer manufacturing, the variations or defects of the wafer substrates will be transmitted to the fabrication process from the preparation process. The continuous 2-D wafer map contains rich information about the wafer quality and therefore should be further studied. There are a few methods for the modeling of the continuous data.

During image processing, the data usually exhibit a similar 2-D structure. One of the easiest ways to model 2-D continuous data is to build a polynomial response surface model, in which the surface is estimated as the sum of approximated polynomial functions and random errors. Taam (1998) used a second-order polynomial model to characterize a 2-D continuous surface. However, the spatial correlation was not considered in this approach. In computer graphics, approaches such as Spline (Lee *et al.* (1997)) and wavelet (Valette and Prost (2004)) are used as an interpolation method for image reconstruction. However, the characteristics of the wafer and the manufacturing process are ignored in these models (Jin *et al.* (2012)).

There is also a rich body of literature in geostatistics to model continuous data with spatial structures. These models can achieve better performance in the inference, prediction and estimation of surface variability by taking spatial dependence into consideration (Haran (2010)). The commonly used modeling techniques include the Kriging and random field models.

Kriging is an interpolation method to predict the value of at an unobserved location from its neighboring observations. It has been demonstrated that Kriging provides more accurate global approximations for the surfaces than the traditional response surface methods (Simpson *et al.* (2001)). Although the mean part of the Kriging could be a constant, it can perform as well as the second-order response surface model (Simpson *et al.* (1998)). Kriging also performs better than the cubic spline methods (Voltz and Webster (1990)).

Kriging has been widely used in many areas, such as hydrogeology (Tonkin and Larson (2002), Chiles and Delfiner (2012)) and mining (Journel and Huijbregts (1978), Richmond (2003)). Ordinary Kriging is usually used to model deterministic processes, such as the input-output relation in computer experiments (Qian and Wu (2008)). In a stochastic simulation, the nugget effect is usually added, which changes the model to a stochastic Kriging model to account for the random noises present in a heteroscedastic process (Yin *et al.* (2011)). Although Kriging has been used in many cases, the use of the Kriging technique still suffers from issues such as the selection of an appropriate model structure in the regression mean part (Martin and Simpson (2005)) or the specification of the correlation matrix structure.

Besides Kriging, Gaussian Markov Random Field (GMRF) models can also model the spatially correlated 2-D data. GMRF can be considered as a special case of either the Gaussian random field or Markov random field models. The Gaussian random field model assumes that the variables follow a multivariate Gaussian distribution, while the Markov random field model assumes the Markov property among the sites. The GMRF model further applies the Markov constraint to the correlation structure of a multivariate Gaussian distribution, i.e., the conditional probability of one site, given all its *immediate* neighbors, is equal to the conditional probability of the site, given *all* other sites.

Because of the Markovian property, the GMRF models the surface with a comparable performance to a full Gaussian model, but significantly reduced computational costs (Hrafnkelsson and Cressie (2003), Hartman (2006)). Moreover, the GMRF model is also flexible in handling data. It can be used to investigate irregular 2-D map shapes (Lindgren and Rue (2008)) and spatial-temporal cases (Allcroft and Glasbey (2003)). If the precision matrix is not full rank, the model reduces to an intrinsic Gaussian Markov random field (IGMRF) model. Huang (2010) used an IGMRF model to characterize the local variability of the nanowires length.

Because the specification of the dependency of the Markov random field is realized through a conditional probability, the Markov Chain Monte Carlo (MCMC) method is usually used to estimate the GMRF models. Based on the 2-D continuous data of wafers that we obtained, the spatial correlation structure satisfies the assumptions of the IGMRF model. Therefore, we use an IGMRF model to characterize the wafer surface variation in this work.

3. Analysis of the variation patterns on the wafer surface

The data used in this paper are collected from the wafer preparation process. To construct an appropriate model for the wafers, it is essential to obtain the engineering knowledge for the variation patterns. Therefore, we first introduce the wafer manufacturing process from the engineering perspective, since the variations are directly introduced by the process. Then we analyze and interpret the macro and micro scale variation patterns from the real data.

3.1. The wafer preparation process

In the wafer preparation process, a lapping process is a critical step to improve wafer uniformity. Major quality metrics are mainly determined by this process. The continuous data are also collected right after the lapping process. Therefore, we give a brief introduction of this process first.

A schematic illustration of the lapping equipment is shown in Figure 3. To start the lapping, wafers are firstly placed on the lower plate, and the upper plate presses against the lower plate with the pressure applied. During the lapping process, the upper and lower plates rotate in opposite directions. The abrasive slurry is fed into the lapping process, where the abrasive particles in the slurry are used to remove the wafer material. The summary metrics introduced in Section 1 are used to measure the wafer quality. Due to the complexity of the lapping mechanism, different points on a wafer are lapped following different moving paths. Therefore, actual lapped wafers are not ideally flat or smooth. More information about the lapping process can be find in Marinescu *et al.* (2006).

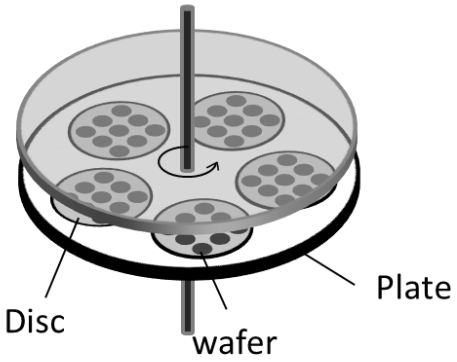


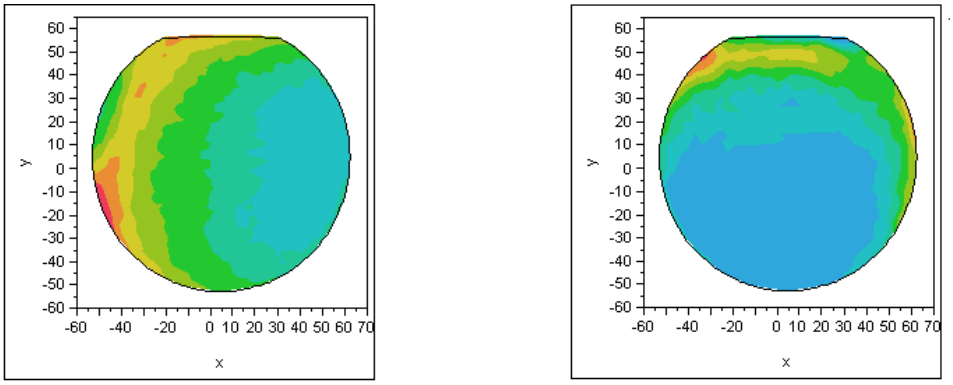
Figure 3: A schematic illustration of the lapping machine

From the engineering knowledge, the wafer quality is directly affected by the wear of the lapping plates, the density and efficiency of the slurry used in production, and the controllable process variables, such as time, pressure and rotation speed. Because of the similar manufacturing conditions apply to the same wafer, a strong spatial correlation of the thickness measures within the wafer is observed. The variation patterns will be analyzed by incorporating such engineering knowledge into the statistical models.

3.2. The macro-scale wafer variation pattern

If a wafer exhibits large surface variation, it may have a high defect ratio in the downstream wafer

fabrication step. Figure 4 shows the thickness heat maps of two wafers. Both wafers are flat and thick on one side, and rough and thin on the other side, which is the macro-scale variation pattern that we identified from the real samples. It is further learned from the engineering knowledge that such macro-scale variation patterns can be attributed to the uneven distribution of the slurry and forces in the slicing and the lapping process, where the removal rate is not uniform (Zhao *et al.* (2011)). Additionally, compared to the reference plane (the horizontal flat edge on the top), the variation patterns show different changing directions. The wafer in Figure 4 (a) changes in the horizontal direction, while the wafer in Figure 4 (b) changes in the vertical direction. The rotation effect of the wafer is one of the challenges in the modeling.



(a) (b)

Figure 4: Wafer thickness heat maps

3.3. The spatial dependency and micro-scale wafer variation

An important and unique feature of the 2-D wafer surface data is the spatial dependency among different sites. The quality measure of one site is correlated with others, and such correlation is higher when the sites are closer to each other. In this section, we investigate the spatial dependency based on a lattice of 86 sampling points taken from each wafer. The distribution and positional index of these sampling points is shown in Figure 5. The center area is missing from the measurement.

It should be noted that "neighbors" of a site are defined as the sites that surround it. Figure 5 also shows the neighbors (with the shaded color) of site #1, site #33, and site #75. Because the readings near the center area are missing, site #33 only has neighbors to its right. In building the statistical model in the next section, the neighborhood relationships are designated by a set, which shows the flexibility of the model in handling irregular data structures.

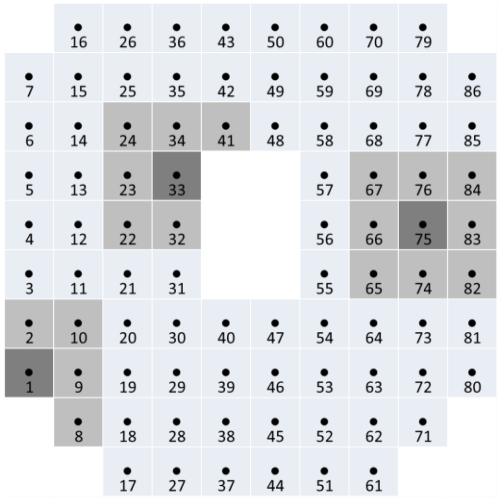
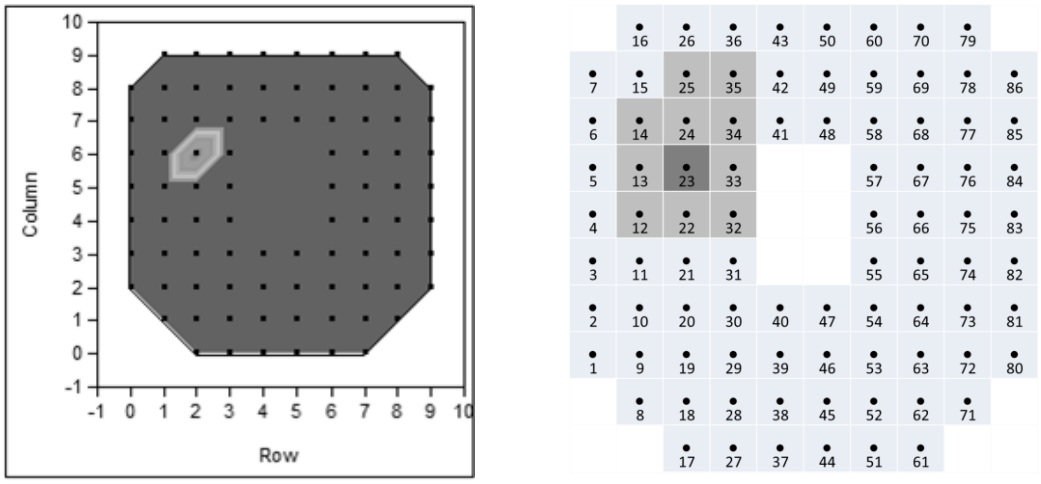


Figure 5: The definition of the neighbors of sites #1, #33 and #75

The dependency among the sites is indicated using their partial correlation. The partial correlation measures the degree of association between two random variables, with the effect of all other sites removed. Taking site #23 on the wafer as an example, the partial correlation coefficients are shown in Figure 6. Figure 6 (a) shows the partial correlation between site #23 and other sites. Figure 6 (b) shows the top ten sites with the largest partial correlation with site #23. It is learned from these two figures that the spatial dependency does exist. Additionally, the immediate neighbors of the site account for higher dependency, which indicates the Markovian property in spatial dependency, i.e., each site only depends on the values of its neighbors. This makes the GMRF model applicable for the wafer surface.



(a) A partial correlation between site #23 and the other sites. A brighter color represents a higher partial correlation.

(b) The top ten sites with the largest partial correlation with site #23. The sites are marked with dark shading.

Figure 6: The partial correlation analysis of site #23 on a wafer

The site dependency only shows the micro-scale variation of the wafer. The major geometric metrics, such as the TTV and TIR, are dominated by the combined effect of the macro-scale and micro-scale variations. Therefore, we propose to use a hierarchical IGMRF model to characterize the both scales of variation.

4. Modeling the wafer surface variation

As the wafer exhibits spatial dependency with the Markovian property, we model the wafer surface as a GMRF and adopt a hierarchical approach for the modeling. In general, hierarchical modeling can be used to imply a complicated distribution without losing the briefness of a conditional distribution (Ntzoufras (2011)). Additionally, the hierarchical structure makes the Bayesian analysis more robust, the interpretation and calculation simpler, and the approximation easier (Robert (2007)). In this section, we first introduce the general IGMRF model, and then extend the model to fit the continuous 2-D wafer map.

4.1. The hierarchical IGMRF model used to characterize the wafer surface variation

A hierarchical IGMRF model is built in multiple stages. In the first stage, a distribution assumption is specified for the observations. In the second stage, a prior model is assigned to the unknown parameters. In this stage, the GMRF is used to represent the unique spatial correlations illustrated above. In the third stage, a prior distribution is assigned to the unknown parameters of the GMRF (Rue and Held (2005)). Following this process, the form of a general hierarchical IGMRF model is constructed as follows:

$$\begin{aligned} &Stage\,1: && y(s) \sim N(\mu(s), \sigma_\varepsilon^2) \\ &Stage\,2: && \mu(s) = X^T\beta + w(s) \\ &Stage\,3: && w(s) \sim car(\kappa) \\ & && other\ priors \end{aligned} \tag{1}$$

We apply this three stage model for wafer surface variation characterization and explain the equations in greater detail.

4.1.1. Stage 1

Let $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ be the sampling sites and $y(s)$ be the value at site s. At stage 1, we assume that the wafer thickness at a site follows a normal distribution. When the mean pattern is removed, the residual varies randomly. The real data does not violate this assumption.

4.1.2. Stage 2

The mean pattern of the thickness distribution is assumed to follow the prior model $\mu(s) = X^T\beta + w(s)$, where $X^T\beta$ represents the macro-scale trend, and $w(s)$ represents the micro-scale variation. $w(s)$ is modeled as an IGMRF.

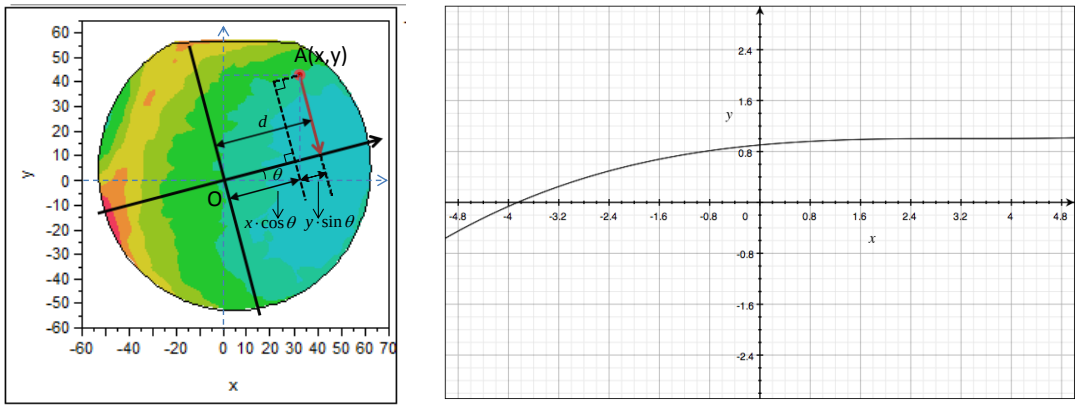
Although the thicknesses follow a generally decreasing trend from one side to the other side, each wafer has a distinct direction for this trend, as shown in Figure 4. The random rotation direction is

caused by both the initial placement and the random self-rotation of the wafer during lapping. Without the modeling of rotation effect, a data driven modeling approach will not be able to remove this effect and characterize the micro-scale variation by the model parameters.

To model the rotation effect, we firstly define a new reference line along the center of the wafer, as shown in Figure 7 (a). A site $A(x, y)$ is then projected onto the reference line. The distance between the projected point and the center of the wafer, O, is calculated as $d = (x \cdot \cos\theta + y \cdot \sin\theta)$, where θ is the angle of the reference line. Along the reference line, the thickness shows a trend similar to the curve drawn in Figure 7 (b). In the figure, the horizontal axis is the projected distance, and the vertical axis is the average thickness at that point. Further data analysis shows that the relationship between the thickness and the projected distance can be well estimated by a third order polynomial function. Therefore, we rewrite the macro-scale mean pattern $\mu(s)$ as

$$Stage\,2:\ \mu(s) = \alpha + \beta \cdot (x \cdot \cos\theta + y \cdot \sin\theta - \delta)^3 + w(s) \qquad (2)$$

where α is an intercept item, which is the baseline of the thicknesses; δ is the location shift of the curve; and β represents the magnitude of the curve. These parameters can show the flatness of the wafer and the decaying speed of the curve.



(a) (b)

Figure 7: Modeling the rotation effect (a) The projection of points onto the reference line; (b) an illustration of the decreasing pattern along the reference line

In Equation **Error! Reference source not found.**, $w(s)$ is modeled as a first-order IGMRF model. Since the thickness of one site is mainly affected by its immediate neighbors, a first-order structure is sufficient to model the spatial variability caused by the dependence of the sites. We denote $w(s)$ at site i as w_i. Then the difference of $w(s)$ between any two adjacent sites i and j is assumed to follow a normal distribution, i.e.

$$w_i - w_j \sim N(0, \kappa^{-1}) \tag{3}$$

where κ is the precision parameter of the random field model.

Let $\mathbf{w(s)}$ be the vector of the thickness of all sites. Based on the above assumption, it is shown in the Appendix that the joint probability density of $\mathbf{w(s)}$ is given by

$$f(\mathbf{w(s)} \mid \kappa) \propto \kappa^{(n-1)/2} \exp[-\frac{\kappa}{2}\sum_{i\sim j}(w_i - w_j)^2] = \kappa^{(n-1)/2} \exp[-\frac{1}{2}\mathbf{w(s)}^T \mathbf{Q}\, \mathbf{w(s)}] \tag{4}$$

where $i \sim j$ means that site i and site j are neighbors; $\mathbf{Q}$ is the precision matrix with rank $n-1$ and $\mathbf{Q}\times\mathbf{1} = \mathbf{0}$. Note that $\mathbf{1}$ is a vector with all elements equal to 1. This means that the sum of the row elements of $\mathbf{Q}$ equals 0. $\mathbf{Q}$ has an element in the ith row and jth column as

$$Q_{ij} = \kappa \begin{cases} n_i, & if \quad i = j \\ -1, & if \quad i \sim j \\ 0, & otherwise \end{cases}$$

The precision matrix $\mathbf{Q}$ defines the neighborhood relationship among the sites, which is determined by the structure of the neighborhood relationship that we specified. It can be shown that

$$w_i \mid \mathbf{w}_{-i}, \kappa \sim N(\frac{1}{n_i}(\sum_{j:j\sim i} w_j), \frac{1}{n_i \kappa}), i = 1, 2, ..., n \tag{5}$$

where n_i denote the number of neighbors of site i; and $-i$ denotes all of the sites except site i. In this way, the conditional mean of w_i is simply a weighted average of its neighbors. The details of the derivation can be found in the Appendix.

In the wafer example, the precision matrix is a 86×86 matrix. This matrix is sparse due to the conditional independence between the corresponding site pairs, which can help to improve the computation speed.

4.1.3. Stage 3:

The above IGMRF model can be used to characterize any surface variation. To fit the model to the wafer examples, we need to specify all of the priors in the model. Such an assignment should be carried out based on the knowledge we learned from the analysis of the real wafer sample data.

The priors of all of the parameters in this model are specified as follows:

$$Stage\,3:\ w(s) \sim car(\kappa), \kappa \sim Gamma(0.05,\ 0.00005), \frac{1}{\sigma_\varepsilon^2} \sim Gamma\left(0.01, 0.01\right)$$
$$a \sim uniform(-\infty, \infty), \beta \sim N(0.005, 1), \theta \sim uniform(0, 6.2831852), \delta \sim N(5, 2)$$

In the model, *car* represents an approach to specify an IGMRF through full conditionals, and the model is also called conditional autoregressions (CAR). $w(s) \sim car(\kappa)$ in Stage 3 indicates that $w(s)$ is a IGMRF with the precision parameter κ. The CAR assumption makes the precision matrix

sparse, since each site only depends on its immediate neighbors. Based on the real data analysis in Figure 6, we confirm that such an assumption is suitable for the wafer example.

The prior of precision parameter κ is $gamma(0.5, 0.0005)$, which is the suggested value for the precision parameter of the spatial random effects in a CAR model, as given by Thomas *et al.* (2004). For the parameter $1/\sigma_\varepsilon^2$, the conjugate prior, the Gamma distribution (or more specifically, the $gamma(0.01, 0.01)$) was suggested by (Ntzoufras, 2011). The rotation angle θ is assumed to be uniformly distribution varying from 0 to 2π. A non-informative uniform distribution is assigned to α. After an initial trial based on the real data, we specify the normal distribution as priors for β and δ to make the algorithm converge faster. The mean of the distribution is learned from the historical data.

In this model, we use MCMC to inference the parameters. MCMC is widely used in Bayesian inference for highly complicated models. The implementation of the algorithm is developed using *WinBUGS*.

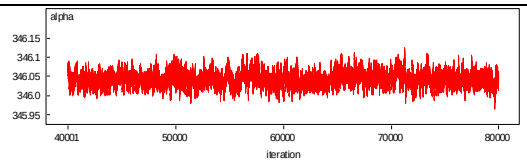
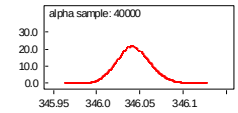
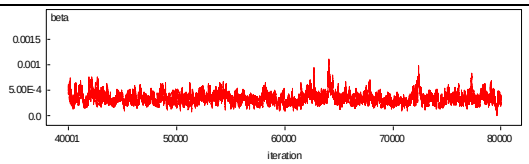
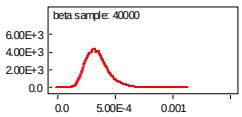
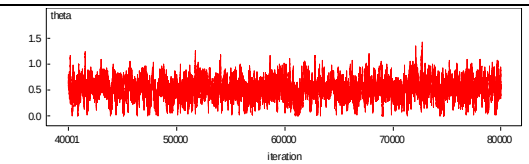
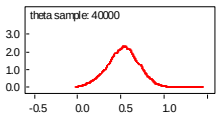
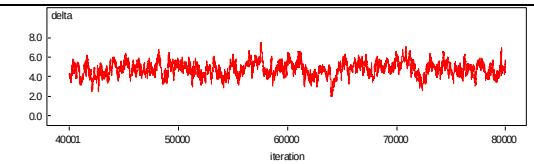
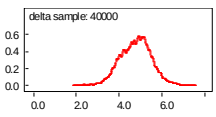
5. Case study

We demonstrate the proposed hierarchical Bayesian model for the real wafer data. For each wafer, 86 thickness readings are measured on locations as shown in Figure 5. We first show the fitted model parameters based on one sample wafer, and then compare the proposed model with other possible models for fitting the wafer data.

Table 1 shows the estimates of the model parameters based on one sample wafer. The mean and quantiles are calculated based on 40,000 iterations in the MCMC. The trace and posterior distributions of these parameters are shown in Figure 8 for evaluation purposes, which show that the MCMC estimation procedure converges.

Table 1: The parameter estimates based on the real data

Parameter	mean	sd	2.50%	median	97.50%
α	346.0423	0.0189	346.0068	346.0417	346.0812
β	0.0003	0.0001	0.0002	0.0003	0.0006
θ	0.5225	0.1837	0.1522	0.5254	0.8752
δ	4.7536	0.7300	3.3236	4.7867	6.1636
κ	74.3374	15.5785	48.2565	72.7339	109.5906
$\sigma = \sqrt{1/\kappa}$	0.1179	0.0123	0.0955	0.1173	0.1440
$1/\sigma_\varepsilon^2$	762.4495	209.3075	426.1014	736.8439	1239.3650
σ_ε	0.0372	0.0051	0.0284	0.0368	0.0484

Parameter	Iteration history	Posterior distribution
α		
β		
θ		
δ		

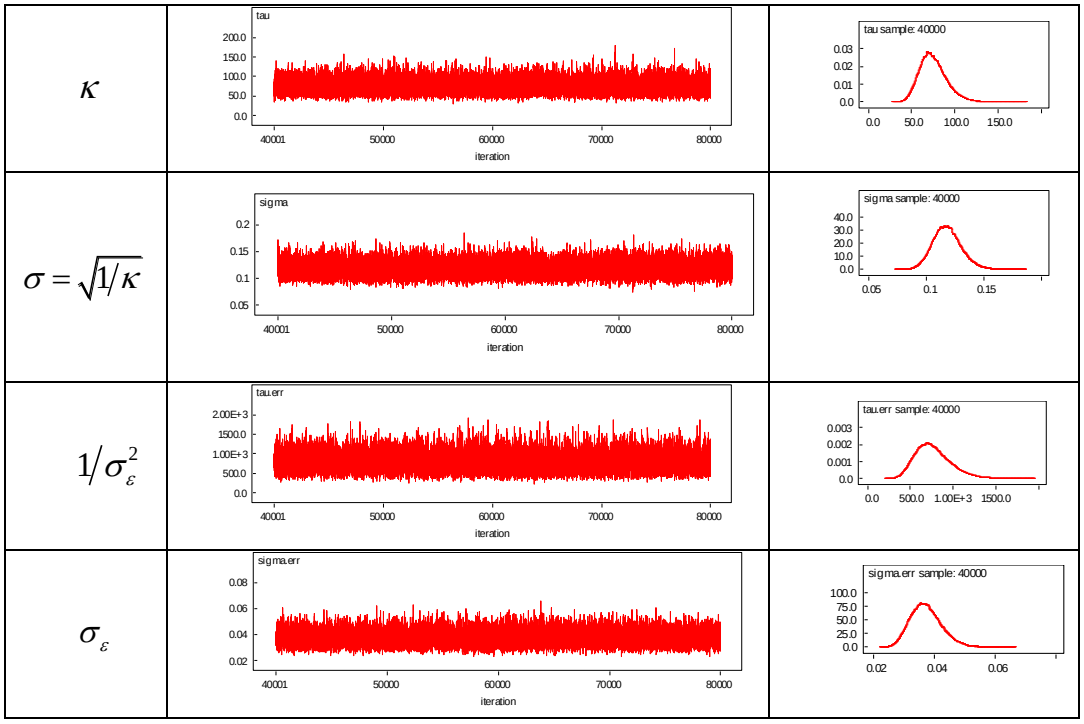
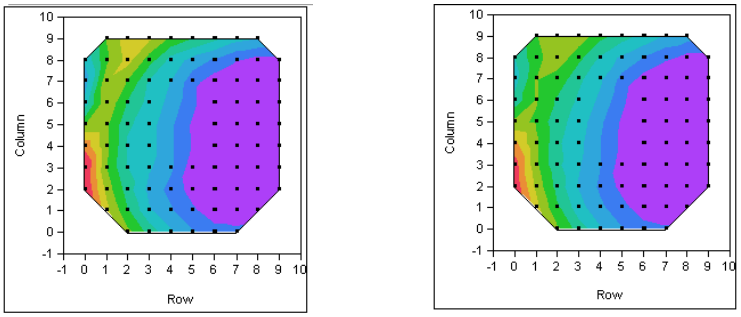


Figure 8: MCMC trajectories and distributions of the estimated parameters



(a) Heat map of actual thickness (b) Heat map of predicted thickness

Figure 9: A comparison between the real and estimated thicknesses

Figure 9 (a) shows the heat map of the actual thickness values, and Figure 9 (b) shows that of the predicted values. It can be seen that the model successfully captures the macro-scale variation. The macro thickness trend is characterized by the cubic term $\alpha + \beta(x \cdot \cos(\theta) + y \cdot \sin(\theta) - \delta)^3$. To validate that the cubic curve we use fits the data well, we now compare two curves. One is a predictive curve, which comes from our parameter estimation result

$$\mu(\mathrm{s}) = 346.0423 + 0.0003(x \cdot \cos(0.5225) + y \cdot \sin(0.5225) - 4.7536)^3$$

The other curve is the real thickness values after removing the micro-scale variation.

Figure 10 shows the predicted and real macro-scale thickness values. We use the projected distance, $x \cdot \cos(\theta) + y \cdot \sin(\theta)$, as the horizontal axis, so that we can see the mean pattern clearer. It is shown that the macro thickness variation trend fits the cubic curve quite well after projection. Therefore, our proposed model is appropriate.

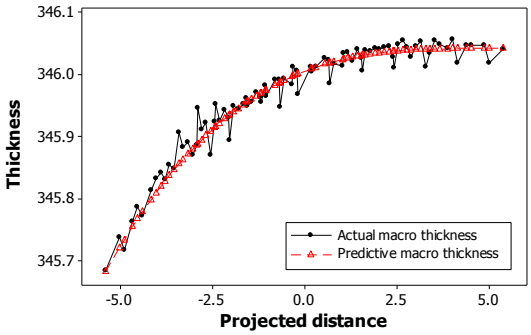


Figure 10: The predicted and real macro-scale variation patterns

Next, we compare the model with other candidate models when fitted to the wafer data. The model performance is measured using the sum of the squared errors (SSE), which is calculated by:

$$SSE = \sum_{s} (y(s_i) - y(s_i))^2 \qquad (6)$$

We fitted four model alternatives and compare their performance.

Model 1: $\mu + \varepsilon$, which uses the mean of the site thicknesses to characterize the whole surface. This model is used because the mean thickness is considered as one of the major quality variables in current industrial practice. A comparison with this model can show the loss of information incurred if the thickness pattern and spatial correlation are not considered.

Model 2: $\mu + a \cdot x + b \cdot y + \varepsilon$, which uses a linear regression model to discover the relationship of thickness to location. In Kriging analysis, it is common to consider the linear effect of location. Therefore, a comparison with this model can review the importance of considering the thickness trend

pattern, which practitioners have tended to ignore.

Model 3: $\alpha + \beta\left(x \cdot \cos(\theta) + y \cdot \sin(\theta)\right)^3 + \varepsilon$, which uses a polynomial function to characterize the surface, without considering the spatial correlations among the sites. This is similar to the model used by Taam (1998), which proposed a second-order polynomial model to fit the wafer surface data. However, the spatial correlations are ignored in that model.

Model 4: $\alpha + \beta\left(x \cdot \cos(\theta) + y \cdot \sin(\theta)\right)^3 + w(s) + \varepsilon$, which uses a combination of the polynomial function and the IGMRF. This is the proposed model.

The models are applied to eight real wafer samples. Table 2 shows the *SSE* of each model. It can be seen that Models 1 and 2 are generally worse than Models 3 and 4, because Models 1 and 2 overlook the rotation effect. Therefore, it is important to estimate the rotation angle in the macro-scale variation analysis. Comparing Model 3 and Model 4, the proposed Model 4 shows much better prediction performance. This indicates that considering the spatial correlation among sites further improves the prediction accuracy. The proposed hierarchical Bayesian IGMRF performs the best among different model alternatives.

Table 2: SSE comparison of models fitting to the real data

Wafer ID	**Model 1**	**Model 2**	**Model 3**	**Model 4 (Proposed)**
1	1.9175	0.4960	0.6332	0.0222
2	3.3729	0.4940	0.3142	0.0408
3	3.4492	0.5985	0.4971	0.0313
4	0.4520	0.2556	0.1758	0.0240
5	0.5500	0.3824	0.3655	0.0433
6	0.3666	0.3017	0.2901	0.0272
7	3.5348	1.5008	0.8378	0.0193
8	0.7534	0.2777	0.2590	0.0214

6. The practical implications of the model

Modeling the wafer surface variation is an essential step in analyzing and improving the wafer quality during the preparation process. The model not only helps to characterize the quality variation but also provides insights to the variation patterns. Such variation patterns contain valuable information for the engineering discipline regarding the root cause diagnosis and process improvement. The proposed model in Equation **Error! Reference source not found.** has two important engineering implications.

First, the rotated cubic curve in the mean part shows the uneven surface variation pattern. From the pattern analysis shown above, we identified this trend in the thickness, which has not received sufficient attention in the past. Based on the fitted model and the significance of the parameters, we characterize such rotation effects. This non-uniformity is possibly attributed to the flatness of the lapping plate. In the lapping process, the upper lapping plate is physically held by a corner joint. Therefore, the upper plate has the freedom to tilt in rotation. Such a tilt could introduce an uneven lapping force to the wafer, thus this process could produce uneven thickness. Additionally, monitoring the fitted parameters by using a control chart is helpful for the identification of sudden or gradual failure in the lapping process, which is important to the prevention of serious quality losses.

Second, the random field and the noise terms in the model show the spatial correlation and the micro-scale variation. In the lapping process, the roughness of the surface is mainly dominated by the lapping slurry and some controller factors, such as pressure and rotation speed. In practice, the changes in the lapping slurry, including its density and particle diameter distributions, are hard to measure. However, such changes would lead to the deterioration of the wafer surface quality. Therefore, we can identify such process changes by checking the fitted parameters, and further

develop condition based replacement strategy for the slurry with the hierarchical model.

7. Conclusions

Wafers are important substrates in semiconductor manufacturing. However, current industrial practice mainly uses summary metrics to characterize wafer quality and variation reduction. In this paper, we proposed a three-stage hierarchical model to characterize the 2-D continuous wafer data. In this model, the wafer thickness variation is decomposed into the macro-scale variation and the micro-scale variation. The macro-scale variation is characterized by a third order polynomial function, which represents the rotation effects of wafers in the preparation process. The micro-scale variation is modeled as a first-order IGMRF. It characterizes the spatial correlation between the sites with simple neighborhood relations. A comparison with other potential model formulations is performed in a case study based on real wafer data, which indicates that the proposed model not only models the wafer surface variation better, but also provides physically meaningful interpretations for macro and micro-scale variation.

The developed statistical model has potential for many quality control applications. By linking the model with the engineering process, we can better understand the variation source and develop advanced methods to improve product quality. Control charts and R2R control strategy based on the spatial data and model will be developed for future research. The use of the proposed model, which is a combination of a regression model and IGMRF model, is not limited to the wafer example. This model could also be extended to similar data format with spatial dependency.

Appendix

In this appendix, the construction of intrinsic Gaussian Markov random filed (IGMRF) and its

precision matrix will be introduced in details.

The thicknesses at different locations on the wafer are correlated, which was validated by the real data samples that we collected. There are two assumptions needed for constructing a Gaussian Markov random field (GMRF) model (Rue and Held (2005)). One is that the random vector follows a multivariate normal distribution, and the other is the conditional independence between the non-adjacent sites, which means two non-adjacent sites are independent when the values of other sites are given. Based on these two assumptions, the fields we constructed must be proper. This implies that a normal distribution with a symmetrical positive definite covariance matrix is equivalent to a GMRF.

The GMRF model has two disadvantages (Besag and Kooperberg (1995)). One is that the marginal variances of the variables in the vector often vary in scale, which is not desirable for modeling a smooth surface. The other is that to get appreciable correlations between the neighboring sites, the parameters may be required to be very close to the boundary of the parameter space. One solution to these disadvantages is to use an improper version, i.e., the Intrinsic GMRF (IGMRF). The precision matrix of IGMRF is not full rank. Therefore, the above assumptions cannot be used to construct the IGMRF. Besag and Kooperberg (1995) found that a Gaussian vector which has a symmetrical positive semi-definite precision matrix satisfies $\mathbf{Q}\cdot\mathbf{1}=\mathbf{0}$, where $\mathbf{Q}$ is the precision matrix and $\mathbf{1}$ is a vector with all the elements equal to 1. Any linearly independent differences among the variables will have a proper Gaussian distribution. And it provides a way to construct the IGMRF by using the independent increments, which follow a proper Gaussian distribution.

For two sites i and j in neighbors, we define a normal increment as

$$w_i - w_j \sim N(0, \kappa^{-1})$$

$$f(w_i - w_j) = \frac{1}{\sqrt{2\pi}} \kappa^{1/2} \exp\left(-\frac{\kappa}{2} (w_i - w_j)^2 \right)$$

For all of the unordered neighbor pairs, the joint probability density is

$$\prod f(w_i - w_j) = \left(\frac{1}{\sqrt{2\pi}} \right)^{n-1} \kappa^{n-1/2} \exp\left(\sum_{i \sim j} -\frac{\kappa}{2} (w_i - w_j)^2 \right)$$

The power of κ equals $n-1$, due to the presence of hidden linear constraints that exist among the variables. Taking the example of a three-element vector $\mathbf{w} = (w_1, w_2, w_3)'$, every two sites are neighbors. Hence, there are three unordered adjacent differences, $w_2 - w_1, w_3 - w_1, w_3 - w_2$, and we can see that $w_3 - w_2 = (w_3 - w_1) - (w_2 - w_1)$.

From the above joint probability density function, we therefore have

$$f(\mathbf{w}) \propto \kappa^{n-1/2} \exp\left(\sum_{i \sim j} -\frac{\kappa}{2} (w_i - w_j)^2 \right)$$

$$\sum_{i \sim j} -\frac{\kappa}{2} (w_i - w_j)^2 = -\frac{\kappa}{2} \sum_{i \sim j} (w_i - w_j)^2 = -\frac{\kappa}{2} \sum_{i \sim j} (w_i^2 - 2 w_i w_j + w_j^2)$$

Because the number of neighbors for site i is n_i, we can write the sum in a matrix form, giving

$$-\frac{\kappa}{2} \sum_{i \sim j} (w_i^2 - 2 w_j w_j + w_j^2) = -\frac{\kappa}{2} \mathbf{w}^T \mathbf{R} \mathbf{w}$$

where $\mathbf{R}$ is called the structure matrix, which is symmetric, and

$$R_{ij} = \begin{cases} n_i, & \text{if } i = j \\ -1, & \text{if } i \sim j \\ 0, & \text{otherwise} \end{cases}$$

Let $\mathbf{Q} = \kappa \mathbf{R}$, and then we have

$$f(\mathbf{w}) \propto \kappa^{n-1/2} \exp\left(-\frac{1}{2} \mathbf{w}^T \mathbf{Q} \mathbf{w} \right)$$

$$Q_{ij} = \kappa \begin{cases} n_i, & if \quad i = j \\ -1, & if \quad i \sim j \\ 0, & otherwise \end{cases}$$

The conditional distribution of w_i can be obtained as

$$w_i \mid \mathbf{w}_{-i}, \kappa \sim N(\frac{1}{n_i}(\sum_{j:j\sim i} w_j), \frac{1}{n_i \kappa}), i = 1, 2, ..., n$$

References

Albin, S., and Friedman, D. (1989). "The impact of clustered defect distributions in ic fabrication." *Management Science*, 35(9), 1066-1078.

Allcroft, D. J., and Glasbey, C. A. (2003). "A latent gaussian markov random field model for spatiotemporal rainfall disaggregation." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4), 487-498.

Besag, J., and Kooperberg, C. (1995). "On conditional and intrinsic autoregressions." *Biometrika*, 82(4), 733-746.

Chiles, J. P., and Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*, Wiley.

Fan, S. K. S. (2000). "Quality improvement of chemical-mechanical wafer planarization process in semiconductor manufacturing using a combined generalized linear modelling-non-linear programming approach." *International Journal of Production Research*, 38(13), 3011-3029.

Hansen, M. H., Nair, V. N., and Friedman, D. J. (1997). "Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects." *Technometrics*, 39(3), 241-253.

Haran, M. (2010). "Gaussian random field models for spatial data." *Handbook of Markov Chain Monte Carlo: Methods and Applications*, 449.

Hartman, L. W. (2006). "Bayesian modelling of spatial data using markov random fields, with application to elemental composition of forest soil." *Mathematical Geology*, 38(2), 113-133.

Hrafnkelsson, B., and Cressie, N. (2003). "Hierarchical modeling of count data with application to nuclear fall-out." *Environmental and ecological statistics*, 10(2), 179-200.

Hsu, S. C., and Chien, C. F. (2007). "Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing." *International Journal of Production Economics*, 107(1), 88-103.

Huang, Q. (2010). "Physics-driven bayesian hierarchical modeling of the nanowire growth process at each scale." *IIE Transactions*, 43(1), 1-11.

Hwang, J. Y., and Kuo, W. (2007). "Model-based clustering for integrated circuit yield enhancement." *European journal of operational research*, 178(1), 143-153.

Jin, R., Chang, C. J., and Shi, J. (2012). "Sequential measurement strategy for wafer geometric profile estimation." *IIE Transactions*, 44(1), 1-12.

Journel, A. G., and Huijbregts, C. J. (1978). *Mining geostatistics*, Academic press.

Lee, S., Wolberg, G., and Shin, S. Y. (1997). "Scattered data interpolation with multilevel b-splines." *IEEE Transactions on Visualization and Computer Graphics*, 3(3), 228-244.

Lindgren, F., and Rue, H. (2008). "On the second‐order random walk model for irregular locations."

Scandinavian journal of statistics, 35(4), 691-700.

Liu, S., Chen, F., and Lu, W. (2002). "Wafer bin map recognition using a neural network approach." *International Journal of Production Research*, 40(10), 2207-2223.

Marinescu, I. D., Uhlmann, E., and Doi, T. (2006). *Handbook of lapping and polishing*, Taylor & Francis.

Martin, J. D., and Simpson, T. W. (2005). "Use of kriging models to approximate deterministic computer models." *AIAA journal*, 43(4), 853-863.

Ntzoufras, I. (2011). *Bayesian modeling using winbugs*, Wiley.

O'Mara, W., Herring, R., and Hunt, L. (2007). *Handbook of semiconductor silicon technology*, Crest Publishing House, South Africa.

O'Mara, W. C., Herring, R. B., and Hunt, L. P. (1990). *Handbook of semiconductor silicon technology*, William Andrew.

Orton, J. W. (2009). *Semiconductors and the information revolution: Magic crystals that made it happen*, Academic Press.

Qian, P. Z. G., and Wu, C. F. J. (2008). "Bayesian hierarchica modeling for integrating low-accuracy and high-accuracy experiments." *Technometrics*, 50(2), 192-204.

Quirk, M., and Serda, J. (2001). *Semiconductor manufacturing technology*, Prentice Hall New Jersey.

Richmond, A. (2003). "Financially efficient ore selections incorporating grade uncertainty." *Mathematical Geology*, 35(2), 195-215.

Robert, C. P. (2007). *The bayesian choice: From decision-theoretic foundations to computational implementation*, Springer Verlag.

Rue, H., and Held, L. (2005). "Gaussian markov random fields theory and applications." Chapman and Hall/CRC Press.

Shen, J., Pei, Z., Fisher, G., and Lee, E. (2006). "Modelling and analysis of waviness reduction in soft-pad grinding of wire-sawn silicon wafers by support vector regression." *International Journal of Production Research*, 44(13), 2605-2623.

Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F. (1998). "Comparison of response surface and kriging models for multidisciplinary design optimization." *AIAA paper 98*, 4758(7).

Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F. (2001). "Kriging models for global approximation in simulation-based multidisciplinary design optimization." *AIAA journal*, 39(12), 2233-2241.

Taam, W. (1998). "A case study on process monitoring for surface features." *Quality and reliability engineering international*, 14(4), 219-226.

Thomas, A., Best, N., Lunn, D., Arnold, R., and Spiegelhalter, D. (2004). "Geobugs user manual." *Accessed online: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs12manual.pdf*, 2012.

Tonkin, M. J., and Larson, S. P. (2002). "Kriging water levels with a regional‐linear and point‐logarithmic drift." *Ground Water*, 40(2), 185-193.

Valette, S., and Prost, P. (2004). "Wavelet-based multiresolution analysis of irregular surface meshes." *Visualization and Computer Graphics, IEEE Transactions on*, 10(2), 113-122.

Voltz, M., and Webster, R. (1990). "A comparison of kriging, cubic splines and classification for predicting soil properties from sample information." *Journal of Soil Science*, 41(3), 473-490.

Wang, C. H., Wang, S. J., and Lee, W. D. (2006). "Automatic identification of spatial defect patterns for semiconductor manufacturing." *International Journal of Production Research*, 44(23),

5169-5185.

Yin, J., Ng, S., and Ng, K. (2011). "Kriging metamodel with modified nugget-effect: The heteroscedastic variance case." *Computers & Industrial Engineering*.

Zhao, H., Jin, R., Wu, S., and Shi, J. (2011). "Pde-constrained gaussian process model on material removal rate of wire saw slicing process." *Journal of Manufacturing Science and Engineering*, 133, 021012.