IIE Transactions

ISSN: 0740-817X (Print) 1545-8830 (Online) Journal homepage: https://www.tandfonline.com/loi/uiie20

Sequential measurement strategy for wafer geometric profile estimation

Ran Jin , Chia-Jung Chang & Jianjun Shi

To cite this article: Ran Jin , Chia-Jung Chang & Jianjun Shi (2012) Sequential measurement strategy for wafer geometric profile estimation, IIE Transactions, 44:1, 1-12, DOI: 10.1080/0740817X.2011.557030

To link to this article: https://doi.org/10.1080/0740817X.2011.557030

Accepted author version posted online: 24 May 2011.
Published online: 07 Nov 2011.

IIE Transactions (2012) **44**, 1–12

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2011.557030

Sequential measurement strategy for wafer geometric profile estimation

RAN JIN, CHIA-JUNG CHANG and JIANJUN SHI*

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive NW, Atlanta, GA 30332, USA
E-mail: Jianjun.shi@isye.gatech.edu

Received April 2010 and accepted December 2010

The geometric profile, factors such as thickness, flatness, and local warp, are important quality features for a wafer. Fast and accurate measurements of those features are crucial in multistage wafer manufacturing processes to ensure product quality, process monitoring, and quality improvement. The current wafer profile measurement schemes are time-consuming and are essentially an offline technology and hence are unable to provide a quick assessment of wafer quality. This article proposes a sequential measurement strategy to reduce the number of samples measured in wafers while still providing an adequate accuracy for quality feature estimation. In the proposed approach, initial samples are measured and then a Gaussian process model is used to fit the measured data and generate a true profile of the measured wafer. The profile prediction and its uncertainty serve as guidelines to determine the measurement locations for the next sampling iteration. The measurement stops when the prediction error of the testing sample set satisfies a pre-designated accuracy requirement. A case study is provided to illustrate the procedures and effectiveness of the proposed methods based on the wafer thickness profile measurement in slicing processes.

Keywords: Gaussian process model, inspection, sampling, semiconductor industry, sequential measurement

1. Introduction

In semiconductor manufacturing, the geometric shape of a wafer is an important index that can be used to evaluate its quality. For example, the profile can be used to estimate quality variables defined by the Semiconductor Equipment and Materials International as industrial standards, such as Total Thickness Variation (TTV), Bow, and Warp. These variables are not only used as measures of the quality of the final wafer but also to identify root the causes of surface imperfections created during a production (Pei *et al.*, 2003; Pei *et al.*, 2004; Zhu and Kao, 2005). Also, the geometric profiles of wafers are modeled to obtain the optimal values of process variables in a wafer manufacturing process (Zhao *et al.*, 2011). This requires online measurement of a wafer's geometric profiles in order to obtain relevant and timely information. This timely information is required if effective process control of a wafer manufacturing process is to be obtained. Unfortunately, current measuring procedures are time-consuming and are unable to provide wafer profile information in a timely manner. For example, the existing wafer measurement technologies, such as touching probe–type sensors, take more than 8 hours to measure a typical batch of wafers (e.g., 400 wafers in one production run). Time-consuming measurements prevent the implementation of advanced process monitoring and diagnosis technologies for quality improvement. Therefore, the objective of this article is to develop an efficient and systematic measurement strategy to reduce the measurement time using a sequential sampling and modeling approach. We propose to minimize a composite index that is based on the measured sample size and times for the model fittings as the efficiency improvement index:

$$Comp.Index = \tau \frac{n_{\text{total}}}{\max(n_{\text{total}})} + \frac{(1-\tau) I_{\text{total}}}{\max(I_{\text{total}})}, \qquad (1)$$

where n_{total} is the total sample size measured for a wafer, I_{total} is the total time for the model fitting in the measurement strategy, τ is a weighting coefficient that evaluates the measurement time for each point and the computation time, and $\max(n_{\text{total}})$ and $\max(I_{\text{total}})$ are the maximum of the total sample size and total time of model fittings for a batch of wafers, which are used to normalize the effects of sample size and number of model fittings. If the same accuracy is created with a smaller composite index, then we consider the measurement strategy to have a higher efficiency.

High-definition geometric profiles of each wafer are measured during the wafer manufacturing process. There are

*Corresponding author

numerous methods to model the geometric profiles presented in the literature. From an engineering perspective, physical analytical models, such as finite element analysis or partial differential equations, are adopted to model the geometric profiles (Zhang and Kapoor, 1990; Abburi and Dixit, 2006; Ozcelik and Bayramoglu, 2006; Huang and Gao, 2010). A major limitation of these methods is that they require a sophisticated understanding of the formation of the profile. Another limitation is that these methods are usually used to model a deterministic profile and therefore they have limited capabilities to model the randomness of the profile errors or the random field effects. Some other approaches, such as methods in computer graphics, use spline (Forsey and Bartels, 1988; Lee *et al*., 1997; Sederberg *et al*., 2004) or wavelet analysis (Schroder, 1996; Valette and Prost, 2004) to model the profile data. In most cases, potential factors that could influence the shape or characteristics are not considered in the profile modeling.

In this article, a Gaussian Process (GP) model is used to characterize the spatially correlated geometric shape of a wafer, including the profile mean, correlated variability, and measurement noise. One of the advantages of the proposed GP model is that the correlated variability can be further decomposed into global variability and local variability components that can be easily interpretated. Global variability represents the trend in variation over the whole wafer, whereas local variability captures the variation only within a small region around the measurement locations. An optimal sampling scheme is required to implement GP models. In the spatial statistics approach, the grid spacing determination approach has been applied to reduce the sample size. By maximizing the grid space, the sampling cost can be minimized in an optimal sampling scheme under the constraints of an allowed maximum error variance (Curran and Williamson, 1986; Curran, 1988). Another approach to determine the optimal grid spacing design for sampling multiple variables is to use the conditional kriging variance based on cross-correlations among variables (McBratney and Webster, 1983a, 1983b; Atkinson *et al.*, 1992, 1994). In addition the relationships between the estimation accuracy of the response variable and the required sample size have been explored and investigated (Wang *et al.*, 2005; Xiao *et al.*, 2005). However, a major limitation of these sampling strategies is that the local spatial variability in the response variable is neglected. Variable grid spacing approaches have been developed that address this deficiency (Anderson *et al.*, 2006). In this approach a smaller grid space is determined in regions with higher local variability, which is equivalent to measuring more samples in that region and vice versa. There are measurement strategies that are based on sequentially allocating the samples based on prior information. This type of strategy is widely used in optimal sensor selection and allocation problems. In the optimal sensor selection or allocation problem, posterior distributions based on prior measurements are used to determine the sensor location that maximizes the information gain. When it is difficult to evaluate an exact posterior distribution, the Sequential Monte Carlo (SMC) method is used for numerical approximation. The SMC method has shown a powerful ability to solve both sophisticated statistical problems and engineering applications (Liu and Chen, 1998; Doucet *et al.*, 2000; Doucet *et al.*, 2001). The Bayesian SMC method has been used to solve the optimal sensor selection problem and target tracking and localization problems (Guo and Wang, 2004). However, the performance of this method depends on the existance of an appropriate parametric form of the Bayesian model, and it is generally computationally intensive for posterior calculations.

The sequential design approach found in computer experiments in CEs is another type sampling strategy that has been extensively used to optimize input variables (Schonlau *et al.*, 1998; Williams *et al.*, 2000; Park *et al.*, 2002; Kleijnen and Van Beers, 2004; Huang *et al.*, 2006). One of the objectives of the sequential design approach is to reduce the number of experimental runs required to reach the optimal solution, which refers to minimum or maximum of the response. A sequential measurement design strategy has been proposed that sequentially allocates more sampling points at the locations with a higher Expected Improvement (EI) to allow the minimum of an investigated surface to be reached quickly (Williams *et al.*, 2000). A larger EI improvement is defined as locations with a smaller predicted value or a larger predicted variance for the minimization of a problem.

Other than focusing on minimizing the required experimental runs to obtain the optimal solution, there are other GP-based sequential sampling approaches that focus on how to sample sequentially in order to obtain an improved model fitting, conditional on the new pair of sample points. These models are usually obtained from the posterior distributions via a Markov Chain Monte Carlo approach. Various thrifty criteria–based sequential sampling problems can be found in MacKay (1992), Cohn (1996), and Muller *et al.* (2004). Some sequential applications have been shown to approximate static optimal designs (Seo *et al.* 2000; Gramacy and Lee, 2009).

Optimal sampling schemes and sequential designs provide effective ways to reduce the sample size by solving a set of optimization problems. However, some of these approaches are computationally intensive and others are not applicable in online measurement situations. For grid spacing determination, the chosen sample locations are not directly associated with locations with higher local variability within each grid. In the sequential design approach, some methods are targeted at optimization objectives, which is not the same as for online measurements. Moreover, most of the sampling schemes and existing sequential measurements involve computationally intensive optimization procedures to determine additional samples.

This article continues the stream of sequential design in CEs to measure samples sequentially, called a

sequential measurement strategy, but differs in the following ways. First, the proposed sampling scheme aims to simultaneously consider both the global variation trend and the local variability pattern to achieve a more accurate prediction ability. Second, prior engineering knowledge on the input–output relationships is taken into account to determine the initial measurement samples. By combining these two aspects, the proposed sequential measurement strategy enhances the wafer quality profile prediction performance with a higher efficiency. Although the proposed framework is similar to the sequential design approach, the innovation of this article lies in two proposed empirical distributions for initial measurements and sequential measurements, which will be discussed in detail later in this article.

The rest of this article is organized as follows. The GP-based sequential measurement strategy is described in detail in Section 2. A case study on wafer thickness profile estimation is provided in Section 3 to evaluate the proposed measurement strategy. Finally, the conclusions and future work directions are summarized in Section 4.

2. GP-based sequential measurement strategy

2.1. *Overview of the sequential measurement strategy*

The framework of the proposed sequential measurement strategy is shown in Fig. 1.

In the proposed methodology, all measurement locations are determined by sampling empirical distributions. The empirical distributions are the estimated probability density functions evaluated at discrete potential measurement locations. The sequential measurement strategy starts to sample an empirical distribution based on engineering knowledge and then fits a GP model that is based on the initial measurements. In the estimation, the measured locations are partitioned into a training sample set and a testing sample set. A GP model is estimated based on the training sample set and the model accuracy is evaluated based on the testing sample set. If the stopping rule is satisfied, the iterative measurement procedure stops; otherwise, additional samples are measured to further improve the estimation performance. In this approach, the magnitude of the gradient and the predicted Mean Squared Error (MSE) from the previous GP model are used to determine the sequential measurements. By iteratively taking the samples and re-fitting the model, the GP models are expected to better approximate the true wafer profile.

2.2. *Measurement locations and data format*

The gauge used in this wafer measurement study is a touching probe sensor, which has a minimal distance of allowable movement and a maximum measuring range. The specification of the gauge defines a potential measurement zone, denoted as D for the set of potential measurement locations. In a continuous measuring scheme, there is a position calibration mechanism such that all wafers are measured at the same locations. For each potential measurement location, there are several quality features to be measured. In the wafer example, wafer thickness, flatness, and local warp are the quality features of interest. Each of these quality features forms a highly spatially correlated data profile, called a *wafer geometric profile*.

Figure 2 shows an example of the potential measurement locations on a wafer and the measured wafer thickness profile. In Fig. 2(a), the solid curve represents the edge of the wafer, and the inner rings formed by stars represent the potential measurement locations in D. The line segment at the top of the wafer is the reference edge. The total number of potential measurement locations is usually a large number in practice and is denoted as n_D. For example, n_D is larger than 5000 for a 6-inch wafer. In Fig. 2(b), the thickness readings are reported for the measurement locations of Fig. 2(a), which form the wafer thickness profile. The gray scale represents the thickness of the wafer. The thicknesses at unmeasured locations are estimated by interpolating the measured points. It is clear that there is a variation in the thickness of the wafer, with the left top corner being thinner. Similar data can be obtained using the measurement gauge for other geometric profiles.

2.3. *Determination of the initial measurement samples*

The determination of the measurement samples is equivalent to selecting a subset of all of the potential measurement locations. When there is no prior knowledge regarding the profile distribution, there are two typical ways to determine the initial samples. One is to use a random measurement strategy, i.e., to measure the samples at randomly selected

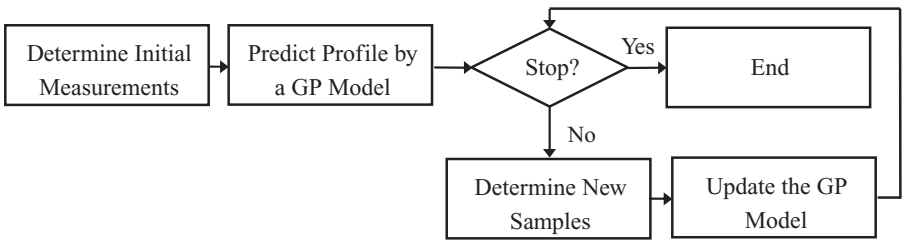

Fig. 1. The framework of the sequential measurement strategy.

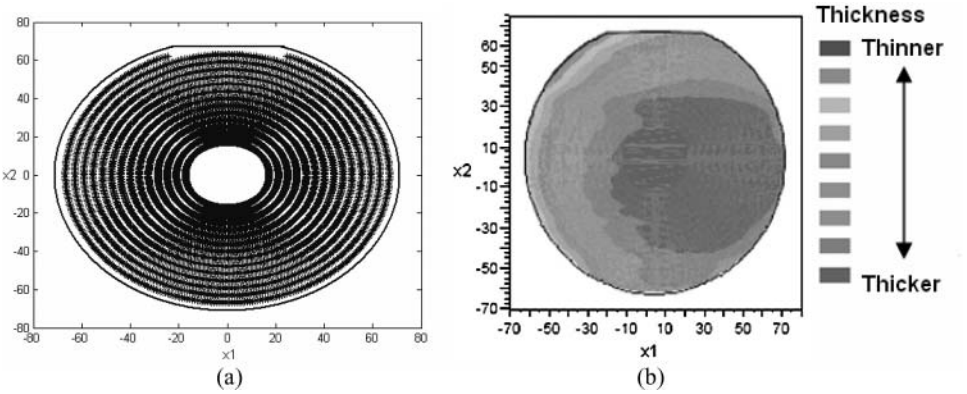

Fig. 2. Potential measurement points and measurement result (a) potential measurement locations and (b) thickness profile.

locations; the other way is to view the initial measurement samples as a design in a CE and then incorporate a Latin Hypercube Design (LHD) or uniform design with space filling criteria to define sample locations (Santner *et al.*, 2003).

The random measurement approach is straightforward and easy to implement; however, a major limitation of this approach is that initial samples may be too close together. Since the wafer profile is spatially correlated, such samples within a region will not only have a high correlation and hence contribute little information toward better fitting but may also may create a singularity when correlation matrixes are inverted in the model estimation process (Santner *et al.*, 2003).

The space filling design approach has an advantage in CEs over the random sampling strategy in that it increases the distance between pairs of points in an initial design. However, this approach can be difficult to implement in regions with irregular shapes, as is the case for the potential measurement points on a wafer considered in this article. Other typical forms of space filling design choose the design points mainly based on the distances between the input variables, which ignores the relationship between response (profile) and input variables (locations). In this case, the design may be inefficient, especially when input–output relationships are available based on engineering knowledge. More advanced techniques may use these input–output relationships in a more complex manner (Santner *et al.*, 2003). Therefore, this article proposes a computationally efficient approach by sampling a weighted empirical distribution based on engineering knowledge.

To efficiently determine the measurement samples, the local variability of a wafer profile is defined as follows.

Definition 1. Local variability of a wafer profile. The local variability of a wafer profile at location $\mathbf{x}$ is the sample variance of the profile values at the measured locations within a region $\Re(\mathbf{x})$:

$$\sigma_Y^2(\mathbf{x}) = \frac{1}{n_{\mathbf{x}} - 1} \sum_{\mathbf{x}' \in \Re(\mathbf{x})} (Y(\mathbf{x}') - \overline{Y}(\mathbf{x}'))^2 \tag{2}$$

where $\mathbf{x}'$ is a location within the region around of $\mathbf{x}$, $\overline{Y}(\mathbf{x}')$ is the sample mean of the profile values in the region, and $n_{\mathbf{x}}$ is the sample size region. The region $\Re(\mathbf{x})$ is defined as the kth nearest neighborhood. Generally, k is chosen to have a value of at least 25 so that a good estimation of variance can be guaranteed.

To determine efficient initial measurement samples, regions that potentially have a higher local variability should have more samples taken from them, as in the grid spacing determination approach (Anderson *et al.*, 2006). In the wafer manufacturing process, the ideal wafer profile has a uniform thickness. Unfortunately, observed thickness values deviate from the ideal profile, which has a set of potential root cause factors, denoted as $U(\mathbf{x}) = \{u_1(\mathbf{x}), \ldots, u_q(\mathbf{x})\}$. Here, the $u_i(\mathbf{x})(i = 1, 2, \ldots, q)$ are the potential factors associated with the wafer locations, such as the contact span at $\mathbf{x}$. The local variability of the profile has a proportional relationship to its potential factors from the engineering perception and is verified by the data:

$$\sigma_Y^2(\mathbf{x}) \propto \frac{\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})}{\prod_{j=q_1+1}^{q} u_j^{t_j}(\mathbf{x})} \tag{3}$$

where $\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})$ is the product of the factors that are proportional to the local variability of the profile, $\prod_{j=q_1+1}^{q} u_j^{t_i}(\mathbf{x})$ is the product of factors that are inversely proportional to local variability of the profiles (there are q_1 proportional factors and $(q - q_1)$ inversely proportional factors), and t_i and t_j are the powers for each factor.

Based on the proportional relationship in Equation (3), the initial measurement points are determined by sampling an empirical distribution defined as

$$\Pr(\mathbf{x}) = \frac{1}{c_1} \frac{\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})}{\prod_{j=q_1+1}^{q} u_j^{t_j}(\mathbf{x})} \tag{4}$$

where c_1 is the corresponding normalizing constant.

By sampling the empirical distribution defined in Equation (4), the sample locations with larger local variability will have a higher probability to be selected as the initial measurements. The detailed procedure can be summarized as follows:

Step 1. Obtain the proportional relationship between the local variability of the wafer profile and its potential factors as in Equation (3).
Step 2. Estimate the empirical distribution for the initial measurement using Equation (4).
Step 3. Determine the sample size n_0 and allocate the sample sizes to circles, so that they are proportional to the summation of the probability of the points on that circle.
Step 4. Sample the points from the outermost circle to the innermost circle one by one. For each circle, the points are sampled G times, and the samples are selected to have max–min distances to the samples on the outer circles.
Step 5. Measure the wafer profile at the locations determined in Step 4.

In this procedure, we use stratified sampling and a max–min criterion to determine initial measurements from the outermost circle to the innermost circle. The max–min criterion is

$$\max \min_{\forall \mathbf{x}_{1\in \mathrm{Cir}_1}, \forall \mathbf{x}_{2\in \mathrm{Cir}_2}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{5}$$

where Cir_1 and Cir_2 are the sets of locations of the outer circle and the inner circle in Step 4.

2.4. *GP models for the geometric profile of a wafer*

Based on the measurements, a GP model is adopted to model the geometric profile of a wafer:

$$Y(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta} + Z(\mathbf{x}) + \varepsilon \tag{6}$$

where $\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$ represents the mean part of the wafer profile; in general, the basis functions $\mathbf{f}^{\mathrm{T}}(\mathbf{x}) = [f_1(\cdot), \ldots, f_p(\cdot)]$ are known; $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$ is the regression coefficient vector; $\mathrm{Z}(\mathbf{x})$ is a Gaussian process with mean zero and covariance function $\sigma_Z^2\psi$; σ_Z^2 is the variance of the covariance function, which represents the wafer profile fluctuation caused by manufacturing error; and ε is the uncorrelated noise term and follows a normal distribution $NID(0, \sigma_\epsilon^2)$, which represents the measurement noise. Note that the correlation function used is an anisotropic Gaussian correlation function that has the form

$$\psi(\mathbf{x}_j, \mathbf{x}_k) = \exp\left(-\sum_{i=1}^{p} \phi_i (x_{ij} - x_{ik})^2\right) \tag{7}$$

where ϕ_i is the scale parameter associated with the ith predictor; $\boldsymbol{\Phi} = [\phi_1, \ldots, \phi_p]$; and p is the dimension of the input variables. In the wafer profile estimation problem, $\mathbf{x}_j$ is the jth location on the wafer with coordinate (x_{1j}, x_{2j}), and $p = 2$. To be more specific, x_{1j} is the axis parallel to the reference edge of a wafer, and x_{2j} is the axis perpendicular to the reference edge of a wafer. The origin is at the geometric center of the wafer.

In the wafer profile estimation problem, we use an ordinary kriging model to fit the wafer's geometric profile:

$$Y(\mathbf{x}) = \beta_0 + Z(\mathbf{x}), \tag{8}$$

where β_0 is the constant mean part. This simplification is based on the facts that (i) a GP model with a constant mean part can adequately model the wafer profile; and (ii) the measurement noise of the wafer profile ε is negligible compared with the profile accuracy requirement.

This model is obtained in the following way. We partition the measured samples into a training sample set $\{\mathbf{x}_i,\ Y_i\}_{i=1}^{n^{Tr}}$, and a testing sample set $\{\mathbf{x}_i,\ Y_i\}_{i=n^{Tr}+1}^{n^{Tr}+n^{Te}}$. Based on the training sample set, the predicted profile at an unobserved location $\mathbf{x}$ is obtained by the ordinary kriging predictor as

$$\widehat{\boldsymbol{Y}}(\mathbf{x}) = \widehat{\beta}_0 + \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\boldsymbol{Y} - \widehat{\beta}_0\mathbf{1}) \tag{9}$$

where $\mathbf{1}$ is a $n^{Tr} \times 1$ vector with all elements equal to 1; $\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} = [\psi(\mathbf{x} - \mathbf{x}_1)\psi(\mathbf{x} - \mathbf{x}_2)\ldots\psi(\mathbf{x} - \mathbf{x}_{n^{Tr}})]$; $\boldsymbol{\Psi}$ is a matrix with elements $\psi(\mathbf{x}_j - \mathbf{x}_k)$ in the row j and column k; $\mathbf{Y} = \lfloor Y_1, Y_2, \ldots, Y_{n^{Tr}} \rfloor^{\mathrm{T}}$; and $\widehat{\beta}_0 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{Y}/\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{1}$. The $\widehat{\boldsymbol{Y}}(\mathbf{x})$ is the best linear unbiased estimator that interpolates all the measured locations (Santner *et al.*, 2003).

In the parameter estimation, the scale parameter $\boldsymbol{\Phi}$ is estimated by Maximum Likelihood Estimation (MLE), denoted as $\widehat{\boldsymbol{\Phi}}$. Then $\widehat{\boldsymbol{\Phi}}$ is plugged into Equation (9) to calculate $\widehat{\beta}_0$ (Santner *et al.*, 2003). In this way, a predicted profile is obtained by changing $\mathbf{x}$ in Equation (9). When there are additional samples collected in new iterations, the unknown parameters are re-estimated, and a new profile for the locations of interest is predicted with the updated ordinary kriging model.

2.5. *Determination of sequential samples*

The samples are measured sequentially so that the samples collected at later measurement iterations can be appropriately selected based on prior information from the GP model. If the stopping rule is not satisfied, the measurement of the $(i+1)$th iteration is required based on the GP model in the ith iteration. We propose to sample an empirical distribution, weighted by the magnitude of the gradient and predicted MSE of the GP model.

Denote the predicted GP response as $\widehat{Y}^i(\mathbf{x})$ in the ith iteration, the gradient of the predicted GP as $\mathrm{d}\,\widehat{Y}^i(\mathbf{x})$ with the magnitude $|\mathrm{d}\,\hat{Y}^i(\mathbf{x})|$, and the MSE at any location $\mathbf{x}$ as $Err_i(\mathbf{x})$, then we have (Lophaven *et al.*, 2002)

$$d\,\widehat{Y}^i(\mathbf{x}) = \frac{\partial^p \widehat{Y}^i(\mathbf{x})}{\partial \mathrm{x}_1, \partial \mathrm{x}_2, \ldots, \partial \mathrm{x}_p},$$

and

$$Err_i(\mathbf{x}) = \widehat{\sigma}_Z^2(1 + (\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x}) - 1)^{\mathrm{T}}(\mathbf{1}\boldsymbol{\Psi}^{-1}\mathbf{1})^{-1} (\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x}) - 1) - \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x})).$$

Then the samples of the $(i+1)$th iteration are sampled from the following empirical distribution:

$$\Pr(\mathbf{x}) = \frac{1}{c_2}\left\{\lambda\left(\frac{|\mathrm{d}\widehat{Y}^i(\mathbf{x})|}{c_3}\right) + (1-\lambda)\left(\frac{Err_i(\mathbf{x})}{c_4}\right)\right\} \quad (10)$$

where λ is a weighting coefficient, which is a tuning parameter; c_2 is a normalizing constant for the distribution; and c_3 and c_4 are the maximum values of the magnitude of the gradient and prediction error, respectively, which are used to standardize the magnitude of the gradient and prediction error. In Equation (10), the first part represents the area with a large fluctuation, and the second part represents the area with a large prediction uncertainty. More measurements in these two types of local areas lead to a reduction in the prediction error since the ordinary kriging model interpolates these extra measurements. Recall that when two sampled locations are close to one another, their correlation may become high. In addition, for a large sample size the maximum distance between any two sampled locations is reduced. These points mean that a higher prediction accuracy can be achieved.

In practice, the distance between measurements should be larger than a minimal distance to avoid the singularity problem when computing the inverse of the correlation matrix. In other words, measurements will not be taken at sampling locations that are too close to previously sampled locations.

2.6. *Stopping rule*

In the sequential measurement strategy, the samples are measured sequentially until a stopping rule is satisfied. In most cases, the Root Mean Sum of the Prediction Error (RMSPE) of a profile is used to evaluate the overall profile prediction accuracy. It is defined as

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{x}_i) - \widehat{Y}(\mathbf{x}_i))^2} \quad (11)$$

where $\widehat{Y}(\mathbf{x}_i)$ is the predicted profile at the location $\mathbf{x}_i$ from the estimated GP model and n is the number of measurements that are compared.

It is ideal to evaluate the RMSPE of samples that are not used in modeling. In order to estimate the RMSPE of the overall wafer profile, we compute the testing error based on the testing sample set collected in each measurement iteration to determine if the measurement stops. The measurement will stop if

$$RMSPE_i^{\text{test}} \le \sqrt{Th_{\sigma^2}} \quad (12)$$

where Th_{σ^2} is a pre-determined estimation variance that represents the profile accuracy requirement. In this article. $RMSPE_i^{\text{test}}$ is the root mean sum of the testing sample set in the ith measurement iteration.

2.7. *Parameter estimation*

In the proposed method, there are several parameters to be determined: the initial sample size n_0, the sequential sample size n_i, and the weighting coefficient λ in Equation (10). In this article, these parameters are selected before the sequential measurement strategy is implemented online, based on a "golden" profile. The golden profile is regarded as a representative profile of a batch of profiles. In the wafer example, the wafer profiles from the *same batch* are assumed to follow the same distribution due to the similarity of the process conditions. A golden profile is selected from one of the representative wafers, where the measurements at all possible potential measurement locations are obtained. The parameters are determined when estimating the golden profile by the sequential measurement strategy.

The initial sample size n_0 is firstly determined by varying n_0 and comparing the RMSPE values in the golden profile. More specifically, we draw n_0 samples using Equation (4) from D in the golden profile N times, denoted as $[\mathbf{x}^{n_0}, Y^{n_0}]^1, [\mathbf{x}^{n_0}, Y^{n_0}]^2, \ldots, [\mathbf{x}^{n_0}, Y^{n_0}]^N$. Based on these samples, N GP models are estimated and their RMSPEs values for the unmeasured samples are calculated. We take the initial sample size as the minimal sample size with $M_{\text{RMSPE}} < T\sqrt{Th_{\sigma^2}}$, where M_{RMSPE} is the median of the RMSPE values of N GP models. T is an appropriately selected constant to have a reliable initial GP model for additional samples. If T is large, n_0 will be small. The estimated initial GP model may have large variation in estimation, and the additional samples may not be reliable for a quick approximation of the geometric profile. If T is small, n_0 will be large. It may take a considerable amount of time to measure a lot of samples, which is unnecessary.

After n_0 is determined, n_i and λ are determined to minimize the composite index defined in Equation (1). Here, we assume that the additional sample sizes n_i are the same in all measurement iterations. Following the sequential measurement strategy, we estimate the composite index for different combinations of n_i and λ based on the golden profile. In this strategy, we apply the same Th_{σ^2}. Therefore, the combination of n_i and λ with the smallest composite index yields the best measurement efficiency, and it will be selected as the parameter for the sequential measurement strategy.

3. Case study

A case study is now conducted on the prediction of the wafer thickness profiles created by a cutting process. Detailed procedures are provided in this section to illustrate

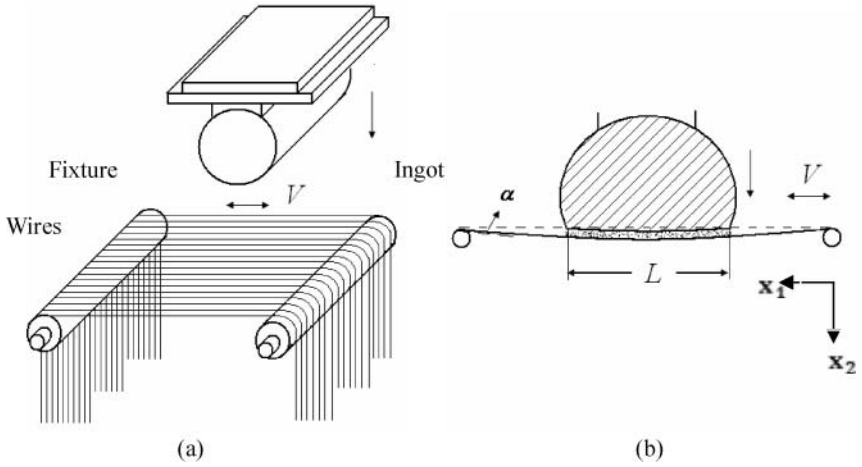

Fig. 3. The slicing processes: (a) the setup and (b) the cutting edge and contact span.

the effectiveness of the proposed sequential measurement strategy.

3.1. *Wafer cutting processes*

A cutting process is used to slice the silicon ingot into wafers that have a rough surface and non-uniform thickness. Figure 3(a) shows the setup used to slice the wafer. The ingot is attached to a holder and is then pressed against multiple equally spaced tensioned wires. The wires move back and forth at a given speed V, and a slurry is sprayed onto the cutting edge. Figure 3(b) is an illustration of the cutting edge and the contact span interaction. In the slicing process, the ingot is pressed against the wires such that there is a bow angle α formed between the wire and the horizontal axis. The contact length between the silicon ingot and the wire is called the *contact span*, defined as L. During the slicing process, a thin film of the applied slurry is formed between the wire and ingot, in which the abrasives remove the silicon material.

The sliced wafer determines the initial geometric quality of a wafer. The profile is significant for wafer monitoring and root cause diagnosis in slicing processes as well as downstream stages. One important wafer quality feature is the thickness profile, which represents the thickness over the wafer. The thickness profile dataset is further used to estimate the TTV of the wafers; i.e., the difference between the maximum and minimum of the thickness profile.

In this case study, a 6-inch ingot was sliced in an HCT wire saw system to create over 400 wafers. Several steps were taken to ensure that the collected wafer profiles were representative of the profile generated in the process.

1. The ingot was sliced using the same setup and conditions as in normal production.
2. The system was monitored to ensure that abnormal cutting conditions were not created.
3. The characteristics of the slurry were monitored to ensure a satisfactory slicing efficiency.

The thickness profiles were measured using a touching probe. First, the wafers were loaded onto a conveyer belt that passed through the measurement area, and the thickness at all possible locations in D were recorded. Since the measurement time for each wafer is over 60 seconds, it would take over 8 hours to measure all 400 wafers. Therefore, 71 wafers were selected to represent all of the ingots in this case study. For each wafer, the thickness profile was stored in a three-column matrix, with the first and second columns representing the coordinates of the measurement locations and the third column representing the wafer thickness.

3.2. *Parameter determination in the case study*

The profile data of the 71 wafers were used to evaluate the sequential measurement strategy. The measurement strategy was used to select a subset of the data on each profile to fit GP models, which mimics the actual measurement procedure. The thickness at each location in D and the TTV for each wafer was predicted by the final GP model when the stopping rule was satisfied. These predicted values were treated as the real measurements. The predicted thickness and TTV obtained from GP models were compared with the measured thickness and TTV to evaluate the measurement strategies. For these 71 wafers, one wafer was selected as the golden profile and the other 70 wafers were used in the evaluation procedure.

There are many factors that can potentially affect the variation in thickness, including slicing speed V, contact span L, and bow angle α, as marked in Fig. 3(b) (Zhu and Kao, 2005). Engineering knowledge was used to create a partial proportional relationship between the variables. The relationship suggests a larger variation in the thickness profile when the contact span is shorter or the location is

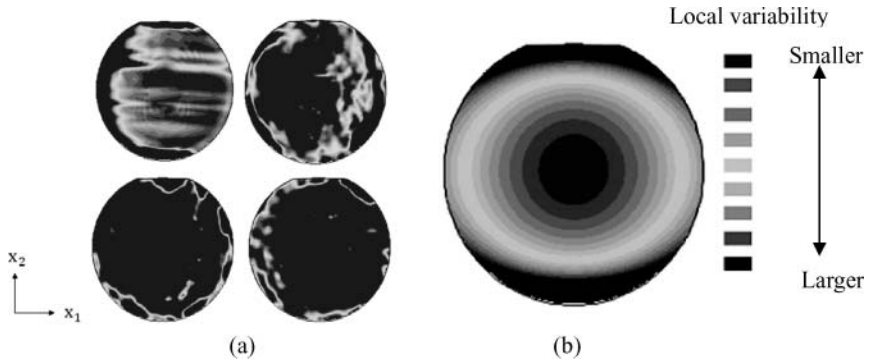

Fig. 4. (a) Local variability (nearest 25 points) and (b) fitted proportional relationship.

nearer to the edge of a wafer, denoted as

$$\sigma_Y^2(\mathbf{x}) \propto \frac{r^t(\mathbf{x})}{L^t(\mathbf{x})}, \tag{13}$$

where r is the radius from location $\mathbf{x}$ to the center of the wafer; L is the contact span of location $\mathbf{x}$; and this case study assumes $t_i = t_j = t$. The unknown parameter t was estimated from the local variability of the golden profile by MLE, leading to $\hat{t} = 1.97$.

Figure 4(a) shows the local variability patterns estimated from the wafer profiles using Equation (2). It is clear that there is a large variation at locations where the contact span is short or the radius is large. Based on these common characteristics, we used the proportional relationship in Equation (13) to capture these variation patterns and the golden profile to calibrate the parameter t, as shown in Fig. 4(b).

Following the procedure outlined in Section 2.7, we determined the parameters based on the golden profile. In this case study, we chose $T = 3$ to determine the initial sample size n_0. In this case, $n_0 = 100$. Then we obtained the composite index as a function of λ and n_i by analyzing the golden profile. In this case study, we chose $\tau = 0.5$ by assuming the measurement time for a batch of samples and model fitting time for these samples to be comparable. Thus, when $n_i = 70$ and $\lambda = 0.8$, the composite index was minimized. These values were used in the simulation studies.

3.3. *Performance analysis and comparison*

After the parameters were determined, a series of GP models was estimated and the samples were determined. Figure 5 shows the results obtained using the sequential measurement procedure for thickness profile prediction. The initial sampling distribution was weighted by radius and contact span, shown in Fig. 5(a). The initial samples were measured to create an initial sample distribution of $n_0 = 100$, and these locations are shown in Fig. 5(b), marked as stars. The measured data were then partioned into a training sample set (75%) and a testing sample set (25%). A GP model was estimated based on the training sample set (75 samples for the initial model fitting). In this model, the mean of the thickness profile was removed before the modeling. Figure 5(c) shows a comparison of the measured thickness profile (solid lines) and the estimated profile obtained using the GP model ("+" lines). The GP model was

$$\hat{Y}(\mathbf{x}) = -0.3011 + Z(\mathbf{x}), \tag{14}$$

where $\widehat{\mathbf{\Phi}} = [5.95, 2.50]$, $\sigma_Z^2 = 2.79$, and $RMSPE_0^{\text{test}}$ is 0.8662 μm. From Fig. 5(c), it can be seen that the GP model provides a close approximation of the overall wafer profile. However, a large prediction error can be observed at some locations (marked by a dashed circle), and at these points the prediction is not accurate enough. Additional samples may be needed to further reduce these errors.

An additional 70 points were measured and the magnitude of the gradient and predicted MSE for the additional measurement could be estimated using the GP model in Equation (14), and the results are shown in Figs 5(d) and 5(e), respectively. Finally, the empirical distribution for sequential samples is shown in Fig. 5(f). A distance threshold of 3 mm was used during the sampling of the data.

Once the additional samples were determined, all 170 measured samples were again randomly partitioned into a training sample set (75%) and a testing sample set (25%). The GP model was updated based on the training sample set as

$$\hat{Y}(\mathbf{x}) = -0.4808 + Z(\mathbf{x}), \tag{15}$$

where $\widehat{\mathbf{\Phi}} = [4.20, 5.00]$, $\sigma_Z^2 = 4.96$, and $RMSPE_0^{\text{test}}$ is 0.8389 μm. In this way, the GP model improves as the sample size increases.

The sequential samples were measured iteratively until the accuracy of the estimated quality variables satisfied the stopping rule. In this article, $Th_{\sigma^2} = 0.04$; i.e., the requirement is a standard deviation of 0.2 μm. This accuracy requirement is an engineering specification of the design in wafer manufacturing processes.

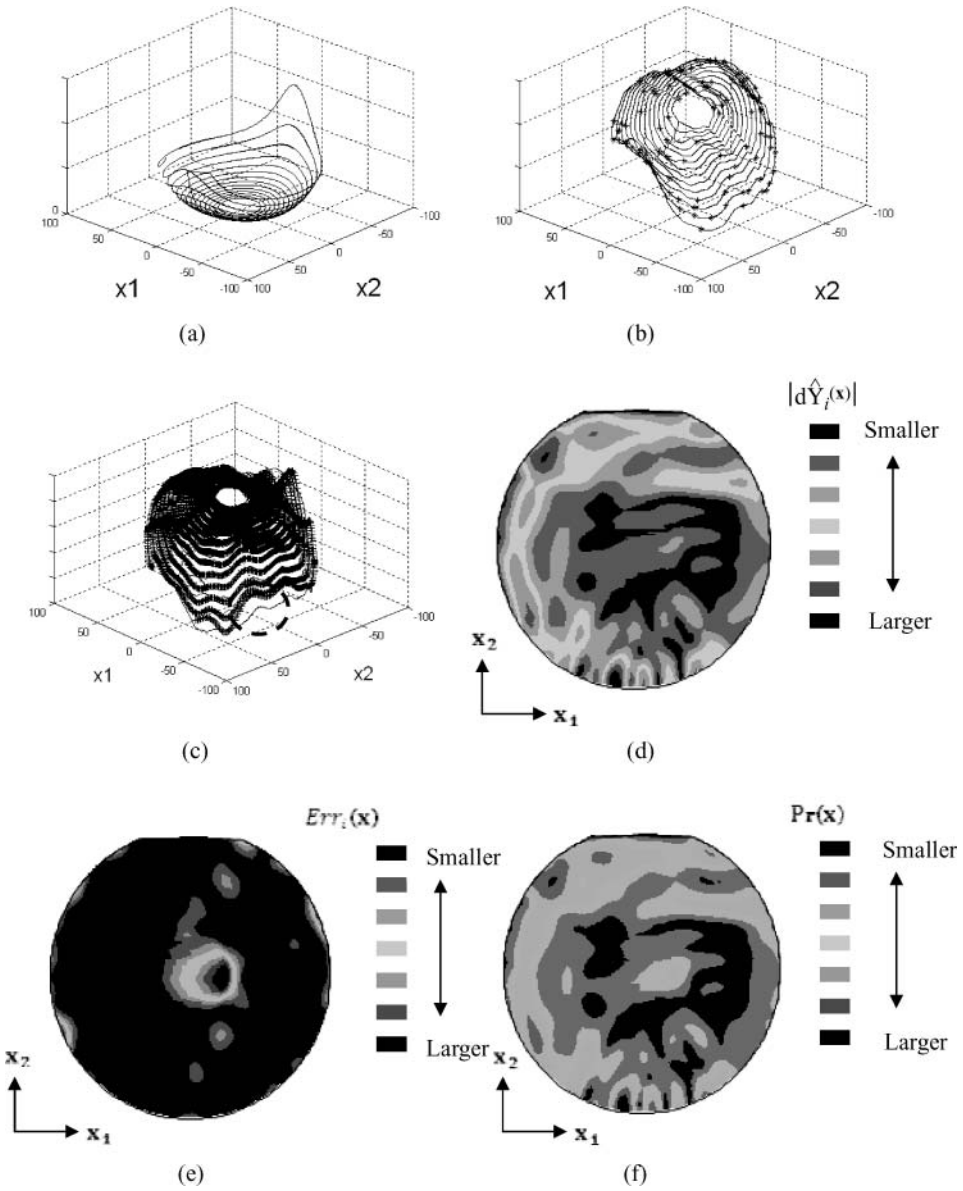

Fig. 5. Intermediate results of sequential measurement strategy: (a) initial empirical distribution, (b) initial samples, (c) GP model and true profile, (d) magnitude of the gradient, (e) predicted MSE, and (f) empirical distribution of sequential samples.

Once the stopping rule is satisfied, the TTV can be estimated as

$$\widehat{TTV} = \max(\hat{Y}(\mathbf{x})) - \min(\hat{Y}(\mathbf{x})). \tag{16}$$

To evaluate the performance of the sequential measurement approach, the thickness profiles of 70 wafers were measured based on three different measurement strategies: a random measurement strategy (denoted as "Rand."), a sequential measurement strategy with space filling of initial measurements (denoted as "Space-seq."), and a sequential measurement strategy with initial measurements from engineering knowledge (denoted as "Eng.-seq."). In the random measurement strategy, the measurement locations were randomly selected following a discrete uniform distribution. The Euclidean distances of samples were calculated to reject samples that were too close to one another. The sampling process for the random measurement was also completed in a sequential way using the same sample size and stopping rule.

The Space-seq. measurement strategy is different from the Eng.-seq. measurement strategy in initial measurements. In the Eng.-seq. measurement strategy, an engineering-driven initial empirical distribution was used to determine the initial samples. In the Space-seq. measurement strategy, n_0 space filling initial measurements were measured. Then additional samples were measured using the Eng.-seq. measurement strategy. Since some typical design approaches, such as LHD, cannot be directly used in

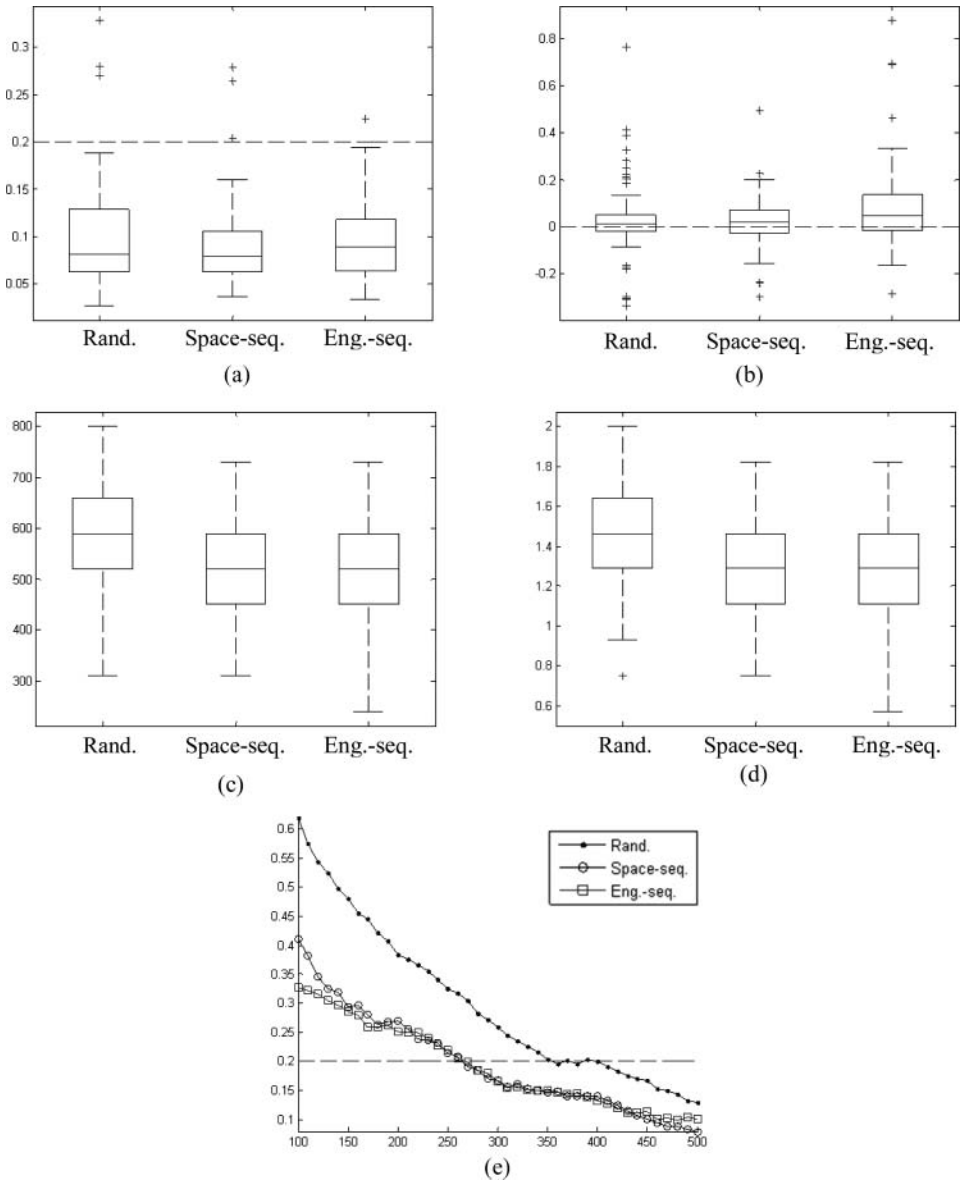

Fig. 6. Performance measure for three measurement strategies: (a) RMSPE, (b) $\widehat{TTV}$ deviation, (c) sample size, (d) composite index, and (e) RMSPE as a function of sample size.

an irregular region, the initial measurements of the space filling design were determined in the following way.

Step 1. The sample size is allocated to circles, which is proportional to the radius of the circles.
Step 2. The samples on the outermost circle are selected as equally spaced samples.
Step 3. The samples on the second outermost circle are also equally spaced, but a max–min criterion is applied to maximize the minimal distance to the samples on outer circles.
Step 4. Step 3 is repeated for the circles with smaller radii, until all samples are selected.

The results obtained using these three measurement strategies are summarized in Fig. 6. Here the RMSPE refers to the RMSPE of all unmeasured locations; i.e., the locations not selected by the measurement strategies. The RMSPEs of the unmeasured locations were used to quantify how well the GP models approximated the profile. Figures 6(a) and 6(b) represent box-plots of RMSPE and $\widehat{TTV}$ deviation for 70 sliced wafers using the three considered strategies. The $\widehat{TTV}$ deviation refers to the deviation between the calculated $\widehat{TTV}$ based on the final GP model and the measured TTV of the wafers. Since the stopping rule sets the same standards in the estimation accuracy of the profile, we have a comparable accuracy performance for the three strategies in terms of the RMSPE and $\widehat{TTV}$ deviation. However, to

achieve a comparable estimation accuracy of the profile, both the Space-seq. measurement strategy and Eng.-seq. measurement strategy use fewer samples, as shown in Fig. 6(c), and they have smaller composite indexes, as shown in Fig. 6(d).

The RMSPE values at different sample sizes are compared in Fig. 6(e). It is clear that the RMSPE values of the Space-seq. and Eng.-seq. measurement strategies have a better estimation performance than the random measurement strategy. The Eng.-seq. measurement strategy has a better estimation performance when the sample size is small, but it quickly converges to a similar performance as that of the Space-seq. measurement strategy. This result indicates that the initial empirical distribution provides useful information to obtain a reliable initial GP model for sequential measurements. The sequential measurement strategy performs well, even if engineering knowledge is not available and the space filling initial measurements are used instead.

4. Conclusions

The geometric profiles of wafers are an important quality feature in semiconductor manufacturing. In most cases, the measurements of the wafer profile are not available during production, since it is time-consuming to measure profiles for a large batch of wafers.

This article proposes an efficient sequential measurement strategy to approximate the thickness profile by estimated GP models. New empirical distributions are proposed to determine measurement locations, including both the initial distribution from engineering knowledge and the sequential measurement distribution from the estimated GP models. In this article, the presented case study indicates that the proposed sequential measurement strategy requires a smaller sample size to achieve a comparable estimation accuracy to that of the random measurement strategy. Moreover, the initial empirical distribution allows a reliable initial GP model to be obtained, when compared to the space filling measurement strategy.

In the GP model estimation, the computation complexity is high when the training sample size becomes large, and the inversion of the covariance matrix may easily become ill-conditioned. In future research, computationally more efficient metamodels will be studied to develop new measurement strategies.

Acknowledgement

The authors gratefully acknowledge the financial support of the National Science Foundation under grant NSF CMMI-1030125.

References

Abburi, N.R. and Dixit, U.S. (2006) A knowledge-based system for the prediction of surface roughness in turning process. *Robotics and Computer-Integrated Manufacturing*, **22**, 363–372.

Anderson, A.B., Wang, G. and Gertner, G. (2006) Local variability based sampling for mapping a soil erosion cover factor by cosimulation with Landsat TM images. *International Journal of Remote Sensing*, **27**, 2423–2447.

Atkinson, P.M., Webster, R. and Curran, P.J. (1992) Cokriging with ground-based radiometry. *Remote Sensing of Environment*, **41**, 45–60.

Atkinson, P.M., Webster, R. and Curran, P.J. (1994) Cokriging with airborne MASS imagery. *Remote Sensing of Environment*, **50**, 335–345.

Cohn, D.A. (1996) Neural network exploration using optimal experimental design. *Neural Networks*, **9**, 1071–1083.

Curran, P.J. (1988) The semivariogram in remote sensing: an introduction. *Remote Sensing of Environment*, **24**, 493–507.

Curran, P.J. and Williamson, H.D. (1986) Sample size for ground and remotely sensed data. *Remote Sensing of Environment*, **20**, 31–41.

Doucet, A., De Freitas, J.F.G. and Gordon, N. (2001) *Sequential Monte Carlo in Practice*, Cambridge University Press, Cambridge, UK.

Doucet, A., Godsill, S.J. and Andrieu, C. (2000) On sequential simulation based methods for Bayesian filtering. *Statistics and Computing*, **10**, 197–208.

Forsey, D. and Bartels, R. (1988) Hierarchical B-spline refinement. *Computer Graphics*, **22**, 205–212.

Gramacy, R.B. and Lee, H.K.H. (2009) Adaptive design and analysis of supercomputer experiment. *Technometrics*, **51**, 130–145.

Guo, D. and Wang, X. (2004) Dynamic sensor collaboration via sequential Monte Carlo. *IEEE Journal on Selected Areas in Communications*, **22**, 1037–1047.

Huang, D., Allen, T.T., Notz, W.I. and Miller, R.A. (2006) Sequential kriging optimization using multiple-fidelity evaluations. *Structure and Multidisciplinary Optimization*, **32**, 369–382.

Huang, X. and Gao, Y. (2010) A discrete system model for form error control in surface grinding. *International Journal of Machine Tools and Manufacture*, **50**, 219–230.

Kleijnen, J.P.C. and Van Beers, W.C.M. (2004) Application-driven sequential designs for simulation experiments: kriging metamodelling. *Journal of the Operational Research Society*, **55**, 876–883.

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

Liu, J.S. and Chen, R. (1998) Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, **93**, 1032–1044.

Lophaven, S.N., Nielsen, H.B. and Søndergaard, J. (2002) *Manual of DACE*. Available at http://www2.imm.dtu.dk/~hbn/dace/.

Mackay, D.J.C. (1992) Information-based objective functions for active data selection. *Neural Computation*, **4**, 589-603.

McBratney, A.B. and Webster, R. (1983a) How many observations are needed for regional estimation of soil properties? *Journal of Soil Science*, **135**, 177–183.

McBratney, A.B. and Webster, R. (1983b) Optimal interpolation and isarithmic mapping of soil properties V. Co-regionalization and multiple sampling strategy. *Journal of Soil Science*, **34**, 137–162.

Muller, P., Sanso, B. and De Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, **99**, 788–798.

Ozcelik, B. and Bayramoglu, M. (2006) The statistical modeling of surface roughness in high-speed flat end milling. *International Journal of Machine Tools & Manufacture*, **46**, 1395–1402.

Park, S., Fowler, J.W., Mackulak, G.T., Keats, J.B. and Carlyle, W.M.M. (2002) D-optimal sequential experiments for generating a

simulation-based cycle time-throughput curve. *Operations Research*, **50**, 981–990.

Pei, Z.J., Kassir, S., Bhagavat, M. and Fisher, G.R. (2004) An experimental investigation into soft-pad grinding of wire-sawn silicon wafers. *International Journal of Machine Tools & Manufacture*, **44**, 299–306.

Pei, Z.J., Xin, X.J. and Liu, W. (2003) Finite element analysis for grinding of wire-sawn silicon wafers: a designed experiment. *International Journal of Machine Tools & Manufacture*, **43**, 7–16.

Santner, T.J., Williams, B.J. and Notz, W.I. (2003) *The Design and Analysis of Computer Experiments*, Springer, New York, NY.

Schonlau, M., Welch, W.J. and Jones, D.R. (1998) Global versus local search in constrained optimization of computer models. *New Developments and Applications in Experimental Design*, **34**, 11–25.

Schroder, P. (1996) Wavelets in computer graphics. *Proceedings of the IEEE*, **84**, 615–625.

Sederberg, T.W., Cardon, D.L., Finnigan, G.T., North, N.S., Zheng, J. and Lyche, T. (2004) T-spline simplification and local refinement. *ACM Transactions on Graphics*, **3**, 276–283.

Seo, S., Wallat, M., Graepel, T. and Obermayer, K. (2000) Gaussian process regression: active data selection and test point rejection, in *Proceedings of the International Joint Conference on Neural Networks*, pp. 241–246.

Valette, S. and Prost, R. (2004) Wavelet-based multiresolution analysis of irregular surface meshes. *IEEE Transaction on Visualization and Computer Graphics*, **10**, 113–122.

Wang, G., Gertner, Z.G. and Anderson, A.B. (2005) Sampling design and uncertainty based on spatial variability of spectral reflectance for mapping vegetation cover. *International Journal of Remote Sensing*, **26**, 3255–3274.

Williams, B.J., Santner, T.J. and Notz, W.I. (2000) Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, **10**, 1133–1152.

Xiao, X., Gertner, G.Z., Wang, G. and Anderson, A.B. (2005) Optimal sampling scheme for estimation and landscape mapping of vegetation cover. *Landscape Ecology*, **20**, 375–387.

Zhang, G.M. and Kapoor, S.G. (1990) Dynamic generation of machined surfaces part 1: description of a random excitation system. Technical report. Institute for Systems Research, TR 1990-36.

Zhao, H., Jin, R. and Shi, J. (2011) PDE-constrained Gaussian process model on material removal rate of wiresaw slicing process modeling. *ASME Transactions*, *Journal of Manufacturing Science and Engineering*, **133**(2), 021012-1–021012-9.

Zhu, L. and Kao, I. (2005) Galerkin-based modal analysis on the vibration of wire–slurry system in wafer slicing using a wire saw. *Journal of Sound and Vibration*, **283**, 589–620.

Biographies

Ran Jin is a Ph.D. student at the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. He received a B.Eng. in Electronics Information Engineering at Tsinghua University, Beijing, in 2005, an M.S. in Industrial Engineering, and an M.A. in Statistics at the University of Michigan, Ann Arbor, in 2007 and 2009, respectively. His research interests include data mining, engineering knowledge–enhanced statistical modeling of complex systems, process monitoring, diagnosis, and control.

Chia-Jung Chang is a Ph.D. student at the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. She received both her B.S. and M.S. degrees from the Industrial Engineering and Engineering Management Department at National Tsing-Hua University in Taiwan in 2005 and 2007, respectively. Her research interests focus on quality engineering and applied statistics, system informatics and control of complex systems, and design of experiments.

Jianjun Shi is the Carolyn J. Stewart Chair Professor at the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. Before joined Georgia Tech in 2008, he was the G. Lawton and Louise G. Johnson Professor of Engineering at the University of Michigan. He was awarded B.S. and M.S. degrees in Electrical Engineering at the Beijing Institute of Technology in 1984 and 1987, respectively, and a Ph.D. in Mechanical Engineering at the University of Michigan in 1992. His research interests focus on the fusion of advanced statistics, signal processing, control theory, and domain knowledge to develop methodologies for modeling, monitoring, diagnosis, and control of complex systems in data-rich environments. He is the founding chairperson of the Quality, Statistics and Reliability subdivision at INFORMS. He currently serves as the Focus Issue Editor of *IIE Transactions on Quality and Reliability Engineering*. He is a Fellow of the Institute of Industrial Engineering, a Fellow of the American Society of Mechanical Engineering, a Fellow of the Institute of Operations Research and Management Science, and also a life member of ASA.