IIE Transactions

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

Monitoring wafers' geometric quality using an additive Gaussian process model

Linmiao Zhang, Kaibo Wang & Nan Chen

To cite this article: Linmiao Zhang, Kaibo Wang & Nan Chen (2016) Monitoring wafers' geometric quality using an additive Gaussian process model, IIE Transactions, 48:1, 1-15, DOI: 10.1080/0740817X.2015.1027455

To link to this article: http://dx.doi.org/10.1080/0740817X.2015.1027455

IIE TRANSACTIONS
2016, VOL. 48, NO. 1, 1–15
http://dx.doi.org/10.1080/0740817X.2015.1027455

Taylor & Francis
Taylor & Francis Group

Monitoring wafers' geometric quality using an additive Gaussian process model

Linmiao Zhang[a], Kaibo Wang[b] and Nan Chen[a]

[a]Department of Industrial and Systems Engineering, National University of Singapore, Singapore; [b]Department of Industrial Engineering, Tsinghua University, Beijing, People's Republic of China

ABSTRACT
The geometric quality of a wafer is an important quality characteristic in the semiconductor industry. However, it is difficult to monitor this characteristic during the manufacturing process due to the challenges created by the complexity of the data structure. In this article, we propose an Additive Gaussian Process (AGP) model to approximate a standard geometric profile of a wafer while quantifying the deviations from the standard when a manufacturing process is in an in-control state. Based on the AGP model, two statistical tests are developed to determine whether or not a newly produced wafer is conforming. We have conducted extensive numerical simulations and real case studies, the results of which indicate that our proposed method is effective and has potentially wide application.

ARTICLE HISTORY
Received 31 March 2013
Accepted 27 February 2015

KEYWORDS
Geometric quality; Gaussian process; T^2 test; generalized likelihood ratio test

1. Introduction

A wafer's geometric quality, which can be manifested by the thickness, roughness, or flatness profile of the entire surface layer, is an important quality feature in the semiconductor industry. More specifically, in the manufacture of electronic chips, a silicon ingot is usually sliced into sections using wire saws. After several flattening steps, including lapping, polishing, and cleaning, the wafers are sent to front-end and back-end processes to form the final chips (O'Mara et al., 1990). An undesired geometric quality often results in a large number of defective dies on the wafer during front-end processes (Doering and Nishi, 2007), causing production delays or economic loss. Due to the importance of geometric quality, people in the semiconductor industry are looking for effective methods to monitor and control quality.

Statistical testing methods enable us to quantitatively monitor quality characteristics. Prior to conducting statistical tests, several preparation procedures, such as data sampling and data modeling, may be applied to help construct the tests. In industrial practice, engineers have developed relatively systematic approaches to determine the geometric conformity of sliced wafers. As shown in Fig. 1, either the contact (using mechanical probe) or non-contact (using capacitance probe or wavelength scanning interferometer) method is able to produce numerous measurements on a single wafer that contain rich information about the geometric quality.

Then several indicators are derived from these measurements to measure the geometric quality based on the International Technology Roadmap for Semiconductors. These indicators include total thickness variation, non-linear thickness variation, bow, warp, and sori (Schmitz et al., 2003). Despite their importance in safeguarding the geometric quality, these summary indicators cannot provide a comprehensive view of the geometric quality for several reasons. First, the aggregated indicators are usually summary statistics, which lose the majority of the rich information the metrology equipment may provide. Second, although the aggregated indicators are effective in screening out nonconforming units, the efficiency of the indicators for identifying process changes is usually unsatisfactory. Jin et al. (2012) reported that the contact method may take more than 8 hours to measure a typical batch (400 in one production run) of wafers. The non-contact method could take an even longer time. Third, and more important, when quality deterioration is noticed from the aggregated indicators, they cannot provide detailed insight about the failure patterns or root causes due to their loss of measurement information. Therefore, a more systematic and efficient method to utilize these data to model and monitor the geometric quality of a wafer is desired. However, there are several difficulties that make this task a challenge.

First, as demonstrated in Fig. 1, the thickness profile is rather complex. No simple patterns or trends can be visually identified, and it is difficult to accurately model using some parametric functions. As a result, traditional profile monitoring techniques (Zou, Tsung, and Wang, 2007; Zou, Zhou, Wang, and Tsung, 2007; Jensen et al., 2008) that approximate the profile by a parametric function and then monitor the parameter vector are difficult to apply in this case. Second, the measurement locations on different wafers may not be perfectly aligned due to different crystal orientations of the ingots and wafer rotation during the measurement. Therefore, conventional multivariate monitoring schemes such as a T^2 chart are not suitable as the variables being monitored are essentially varying from one wafer to another. Third, the measurements are spatially correlated due to the similar conditions experienced by physically adjacent points. As a result, methods with the assumptions that errors are independently and identically distributed (i.i.d.) are no longer

CONTACT Nan Chen ✉ isecn@nus.edu.sg
Supplemental data for this article can be accessed on the publisher's website at http://www.tandfonline.com/uiie.

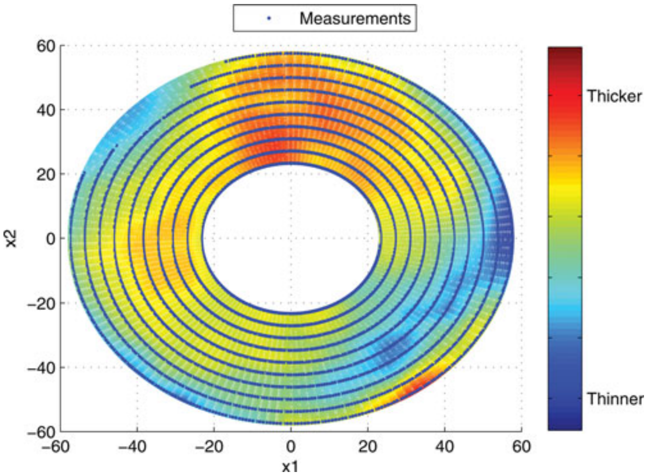


Figure 1. Example of a wafer thickness profile and possible measurement locations.

applicable. Last, but by no means least, not only do changes in the mean or variance of the deviations the reflect potential process shifts, but the changes in its spatial correlation are also a symptom of unexpected process shifts and should be monitored. Therefore, we need a comprehensive monitoring scheme that is effective in detecting all types of changes that may occur in the complex geometric data.

Due to these challenges, only a limited number of papers have been published. In particular, Jin et al. (2012) suggested using a Gaussian process to model the thickness profile of the entire wafer. To speed up the process, they proposed a sequential measurement strategy that adaptively determines the next measurement locations. Using their method, only a small set of measurements needs to be taken, and then a Gaussian process model is built to accurately characterize the entire geometric profile. As a result, the measurement time can be significantly reduced. Despite its importance, their method was developed for measuring a single wafer. Therefore, it is not suitable for quality monitoring, as each geometric profile is modeled as an independent Gaussian process, and there are no statistical rules to determine whether or not the fitted Gaussian process is in control. Zhao et al. (2011) proposed a partial differential equation–constrained Gaussian process model to predict the wafer thickness profile. The model integrates physical knowledge of the slicing process and the observed data to better characterize the geometric quality. However, their method also focuses on modeling a single wafer and lacks a quantification of the variations when the manufacturing process is in control.

Although the aforementioned works did not solve the problem, they have demonstrated that the Gaussian process is a suitable model for spatially correlated data (Cressie, 1993), and it can also characterize complex geometric profiles. In addition, compared with other non-parametric methods such as B-splines (e.g., De Boor (2001)) or kernel smoothing (e.g., Hastie and Loader (1993)), the Gaussian process is much easier to extend to higher input dimensions when the manufacturing process involves other controllable or uncontrollable variables. One thing to be noted is that a single Gaussian process model may not help to detect point-wise deviation between two profiles since they can be independent realizations generated from a single Gaussian process. However, in the early stage of wafer manufacturing, the point-wise deviation is not crucial. Instead, it is the spatial pattern of the surface that influences the downstream manufacturing. For example, Fig. 2 shows two curves (solid line and dashed line representing two profiles) that are point-wise different. However, in terms of their impact on the downstream quality, they are indistinguishable because their spatial patterns are the same. On the other hand, we also want to make sure that the surface does not significantly deviate from the desired profile. Therefore, considering the advantages of the Gaussian process and targeting the limitations in existing works, we propose an Additive Gaussian Process (AGP) model to characterize the geometric profile of a wafer using data measured on a group of wafers. The AGP model is composed of two independent Gaussian processes with different covariance structures. The first Gaussian process is used to approximate the unknown desired (or standard) geometric profile, whereas the second one is used to quantify the "distribution" of spatially correlated deviations from the standard profile when the manufacturing process is in control. By using this approach to construct the model, we are able to detect point-wise changes in the standard profile and spatial pattern changes in the deviations. We would like to highlight that we are not the only ones proposing this "additive" concept. Ba and Joseph (2012) proposed a similar structure called the Composite Gaussian Process (CGP) model. However, the CGP model is mainly focused on modeling non-stationary output data from computer simulations, which is different from our AGP model. A detailed discussion on the differences between the CGP and AGP models will be provided in the next section.

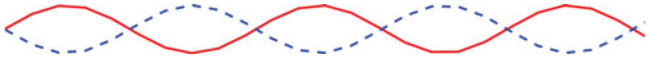


Figure 2. Two curves with the same spatial pattern and point-wise difference.

Since our AGP model considers a group of geometric profiles collaboratively, it allows distinct measurement locations on different wafers. In addition, only a small set of measurements need to be taken from each wafer. Therefore, compared with traditional methods, the required measurement time is significantly shortened. Based on the AGP model, we also develop two statistical monitoring methods, the T^2 chart and the Generalized Likelihood Ratio (GLR) chart, to rigorously analyze whether a tested wafer conforms to the standard within an acceptable variation. The proposed monitoring schemes are able to detect different patterns in changes on the geometric profile including mean shifts, variance shifts, and correlation shifts.

The remainder of this article is organized as follows: Section 2 formulates the problem and introduces the AGP model used to model the standard geometric profile of a wafer and its deviation; Section 3 describes the statistical monitoring schemes we propos to monitor the geometric conformity of a wafer; Section 4 uses extensive simulation studies to investigate the performance of the proposed method; Section 5 presents an application in wafer manufacturing, which monitors the thickness of sliced wafers; Section 6 concludes the article and discusses future directions.

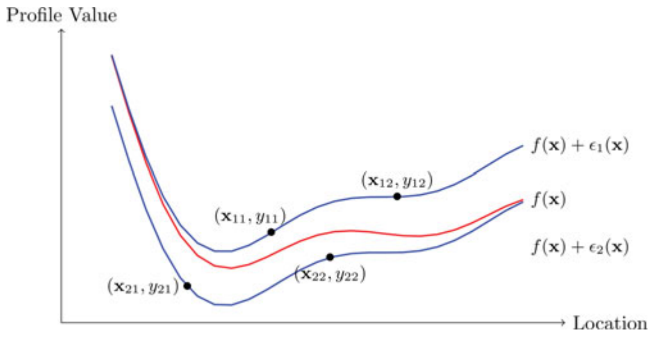

Profile Value
Location
$f(\mathbf{x})+\epsilon_1(\mathbf{x})$
$f(\mathbf{x})$
$f(\mathbf{x})+\epsilon_2(\mathbf{x})$
$(\mathbf{x}_{11}, y_{11})$
$(\mathbf{x}_{12}, y_{12})$
$(\mathbf{x}_{21}, y_{21})$
$(\mathbf{x}_{22}, y_{22})$

Figure 3. Demonstration of notation.

2. Statistical quantification using AGP

2.1. *Problem formulation and notation*

We assume a group of N_0 wafers have been produced when the manufacturing process was in control. On the ith wafer, we take measurements at n_i different locations $\mathbf{X}_i \equiv [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{in_i}]^T$ with corresponding measurement values $\mathbf{Y}_i \equiv [y_{i1}, y_{i2}, \cdots, y_{in_i}]^T$, where $\mathbf{x}$ denotes the two-dimensional coordinate on the wafer.

We use the function $f(\mathbf{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ to denote the standard or desired quality value (e.g., thickness, roughness, or flatness) we expect at location $\mathbf{x}$. Due to different sources of variation in the process, each produced wafer can be modeled as the summation of a standard profile $f(\mathbf{x})$ and a random error; that is,

$$y_{ij} = f(\mathbf{x}_{ij}) + \epsilon_i(\mathbf{x}_{ij}), \qquad \forall i = 1, 2, \cdots, N_0; \quad j = 1, 2, \cdots, n_i, \tag{1}$$

where $\epsilon_i(\mathbf{x}_{ij})$ is the deviation of the quality measurement at location $\mathbf{x}_{ij}$ on wafer i from the standard value $f(\mathbf{x}_{ij})$, including both process variations and measurement errors. Different from conventional models, in Equation (1) $\epsilon_i(\mathbf{x}_{ij})$ and $\epsilon_i(\mathbf{x}_{ik})$, $k \neq j$ are typically correlated because points on the same wafer undergo similar processing conditions, which induces inherent spatial correlations of the deviations between locations $\mathbf{x}_{ij}$ and $\mathbf{x}_{ik}$. In contrast, $\epsilon_i(\mathbf{x}_{ij})$ and $\epsilon_{i'}(\mathbf{x}_{i'k})$, $i \neq i'$ can be considered as independent from each other because different wafers are produced independently. Figure 3 uses a one-dimensional example to demonstrate our notation.

When a new wafer is produced, we take measurements at locations $\mathbf{X}_l \equiv [\mathbf{x}_{l1}, \mathbf{x}_{l2}, \cdots, \mathbf{x}_{ln_l}]^T$ with values $\mathbf{Y}_l \equiv [y_{l1}, y_{l2}, \cdots, y_{ln_l}]^T$. Using the quality measurements $(\mathbf{X}_l, \mathbf{Y}_l)$, we want to develop a systematic monitoring scheme to detect whether or not the manufacturing process is in control. If abnormal deviations from the standard are discovered, appropriate diagnostic and corrective actions need to be taken to improve the process quality.

2.2. *Gaussian process regression*

Gaussian process regression is a popular model in spatial statistics (Cressie, 1993), and it also plays an important role in metamodeling to approximate complex functions (see, e.g., Sacks et al. (1989) and Ankenman et al. (2010)). A typical Gaussian process regression model can be expressed as

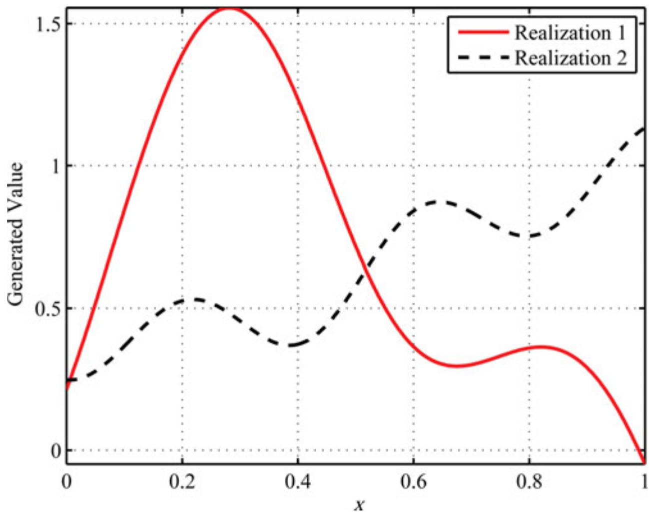


Figure 4. Two independent realizations of the same Gaussian process.

$$y(x) = \lambda(x) + Z(x), \tag{2}$$

where $\lambda(x)$ is a deterministic function, and $Z(x)$ is realization of a Gaussian process with zero mean. The Gaussian process can be viewed as a distribution over a set of continuous functions, and any finite samples from the Gaussian process follow a multivariate normal distribution. As a result, $[y(x_1), y(x_2), \cdots, y(x_n)]^T$ are normally distributed with mean vector $[\lambda(x_1), \lambda(x_2), \cdots, \lambda(x_n)]^T$ and covariance matrix $\mathbf{C} = [k(x_i, x_j)]_{n\times n}$, where $k(x_i, x_j)$ is a positive-definite kernel function. Commonly used kernel functions include the squared exponential function, Matérn functions, etc. (Rasmussen and Williams, 2006).

The Gaussian process is attracting increasing interest due to its high levels of flexibility and conceptual simplicity. More important, it can provide a unique statistical view on the prediction errors. This feature also makes it useful in simulation optimization (Jones et al., 1998; Huang et al., 2006) and sequential samplings and experimental design (Jin et al., 2012). We also want to highlight that Equation (2) is essentially a nonparametric model. In other words, only knowing the parameters of $\lambda(x)$ and $Z(x)$ is insufficient to completely determine $y(x)$. The observed data (whether noisy or not) are also required to provide meaningful predictions.

To demonstrate this, Fig. 4 illustrates two independent realizations of the same Gaussian process with radically different characteristics. As a result, simply monitoring the parameters of the Gaussian process models that are fitted for each individual wafer is insufficient to effectively detect changes in the geometric profiles.

2.3. *AGP model*

In Section 2.1, we use $f(\mathbf{x})$ to represent the standard geometric profile as the desired or designed output from the process. However, the exact function is often unknown and needs to be estimated from historical data. Considering the flexibility requirement on approximation and the characteristics of spatial correlation, in this article we propose an AGP model to quantify the geometric variations when the process is in control. More specifically, we assume $f(\mathbf{x})$ is a realization of the Gaussian process with mean μ and covariance function

$$\begin{aligned} & s(\mathbf{x}_{ij}, \mathbf{x}_{ik}|\boldsymbol{\theta}_1) \\ & \quad = \sigma^2 \exp\left[-(\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \times \operatorname{diag}(\boldsymbol{\theta}_1) \times (\mathbf{x}_{ij} - \mathbf{x}_{ik})\right], \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_1$ is the two-dimensional correlation parameter, and $\operatorname{diag}(\boldsymbol{\theta}_1)$ is the diagonal matrix with diagonal vector $\boldsymbol{\theta}_1$. For demonstration purposes, in this article we use the squared exponential covariance function. Other covariance functions suggested in Rasmussen and Williams (2006) can also be easily applied in the AGP model. In addition, to model the spatial correlation in $\epsilon_i(\mathbf{x})$, we assume $\epsilon_i(\mathbf{x})$ $i = 1, 2, \cdots, N_0$, are independent realizations of another Gaussian process with mean zero and covariance function

$$\begin{aligned} & v(\mathbf{x}_{ij}, \mathbf{x}_{ik}|\boldsymbol{\theta}_2) \\ & \quad = \tau^2 \exp\left[-(\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \times \operatorname{diag}(\boldsymbol{\theta}_2) \times (\mathbf{x}_{ij} - \mathbf{x}_{ik})\right]. \end{aligned} \tag{4}$$

As a result, the observed quality measurements are simply the sum of the realizations of two Gaussian processes, which we refer to as the AGP.

Using the measurement data from in-control wafers, we can approximate the standard profile $f(\mathbf{x})$ and quantify the amount of variation $\epsilon_i(\mathbf{x})$, this provides us with a baseline to monitor newly manufactured wafers. Combining all of the in-control measurements, we denote $\mathbf{X}_{IC} \equiv [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{N_0}]^T$ and $\mathbf{Y}_{IC} \equiv [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{N_0}]^T$, which have the dimensions $M_0 \times 2$ and $M_0 \times 1$, respectively. Here $M_0 = \sum_{i=1}^{N_0} n_i$ is the total number of measurements from all in-control wafers. Also, we denote $\boldsymbol{\beta} \equiv [\mu, \sigma^2, \boldsymbol{\theta}_1, \tau^2, \boldsymbol{\theta}_2]$ as the entire set of parameters in the AGP model. Given $\boldsymbol{\beta}$, we note that $\mathbf{Y}_{IC}$ follows a multivariate normal distribution based on the property of the Gaussian process, with joint density function

$$\begin{aligned} f(\mathbf{Y}_{IC}|\boldsymbol{\beta}) = {} & (2\pi)^{-M_0/2} (\det \boldsymbol{\Sigma}_0)^{-1/2} \\ & \times \exp\left[-\frac{(\mathbf{Y}_{IC} - \mu \mathbf{1}_{M_0})^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{Y}_{IC} - \mu \mathbf{1}_{M_0})}{2}\right], \end{aligned} \tag{5}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector with all ones, and $\boldsymbol{\Sigma}_0$ is the $M_0 \times M_0$ covariance matrix of $\mathbf{Y}_{IC}$. Based on the AGP model, the covariance between elements of $\mathbf{Y}_{IC}$ takes the form

$$\begin{aligned} & \operatorname{cov}(y_{ij}, y_{i'k}) = \\ & \begin{cases} s(\mathbf{x}_{ij}, \mathbf{x}_{i'k}|\boldsymbol{\theta}_1) + v(\mathbf{x}_{ij}, \mathbf{x}_{i'k}|\boldsymbol{\theta}_2), & \forall i = i' \\ & i, i' = 1, 2, \cdots, N_0, \\ s(\mathbf{x}_{ij}, \mathbf{x}_{i'k}|\boldsymbol{\theta}_1), & \forall i \neq i' \end{cases} \end{aligned} \tag{6}$$

because the deviation $\epsilon_i(\mathbf{x})$ is assumed to be independent from $\epsilon_j(\mathbf{x})$, $j \neq i$, and within one wafer $\epsilon_i(\mathbf{x}_{ij})$ and $\epsilon_i(\mathbf{x}_{ik})$ are spatially correlated. Because of this special structure, $\boldsymbol{\Sigma}_0$ is in fact the sum of two covariance matrices, as illustrated in Fig. 5.

Given the in-control measurements $\mathbf{Y}_{IC}$, we can compute the distribution of the measurements at *any* location on a new wafer if the process is still in control. In particular, the measurements $\mathbf{Y}_l$ at location $\mathbf{X}_l$ on a new wafer follow a jointly normal distribution with $\mathbf{Y}_{IC}$:

$$\begin{bmatrix} \mathbf{Y}_l \\ \mathbf{Y}_{IC} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \mathbf{1}_{n_l} \\ \mu \mathbf{1}_{M_0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_l & \boldsymbol{\Sigma}_{l,0} \\ \boldsymbol{\Sigma}_{l,0}^T & \boldsymbol{\Sigma}_0 \end{bmatrix}\right), \tag{7}$$

where $\boldsymbol{\Sigma}_{l,0}$ is the $n_l \times M_0$ covariance matrix between $\mathbf{Y}_l$ and $\mathbf{Y}_{IC}$. Since $\epsilon_l(\mathbf{x})$ is independent from previous deviations, the elements of $\boldsymbol{\Sigma}_{l,0}$ are simply $s(\mathbf{x}_{lj}, \mathbf{x}_{ik}|\theta_1), \forall j = 1, \cdots, n_l, i = 1, \cdots, N_0, k = 1, \cdots, n_i$. Similarly, $\boldsymbol{\Sigma}_l$ is the $n_l \times n_l$ covariance matrix of $\mathbf{Y}_l$ with elements $s(\mathbf{x}_{lj}, \mathbf{x}_{lk}|\theta_1) + v(\mathbf{x}_{lj}, \mathbf{x}_{lk}|\theta_2), \forall j, k = 1, 2, \cdots, n_l$. Following Equation (7), the conditional distribution of $\mathbf{Y}_l$ given $\mathbf{Y}_{IC}$ still follows a multivariate normal distribution with mean vector and covariance matrix

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_l &= \mu \mathbf{1}_{n_l} + \boldsymbol{\Sigma}_{l,0} \boldsymbol{\Sigma}_0^{-1} (\mathbf{Y}_{IC} - \mu \mathbf{1}_{M_0}), \\ \tilde{\boldsymbol{\Sigma}}_l &= \boldsymbol{\Sigma}_l - \boldsymbol{\Sigma}_{l,0} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{l,0}^T. \end{aligned} \tag{8}$$

In other words, when the process is in control, we expect $\mathbf{Y}_l$ to follow the normal distribution with mean $\tilde{\boldsymbol{\mu}}_l$ and variance $\tilde{\boldsymbol{\Sigma}}_l$. We can use this information to develop monitoring statistics to detect changes in the process.

If $\boldsymbol{\beta}$ is unknown, it can be substituted by the estimated value $\hat{\boldsymbol{\beta}}$, which can be obtained by maximizing the log-likelihood function (up to a constant) of the in-control samples:

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \Bigg\{ & -\frac{1}{2} \log(\det \boldsymbol{\Sigma}_0) \\ & - \frac{1}{2} (\mathbf{Y}_{IC} - \mu \mathbf{1}_{M_0})^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{Y}_{IC} - \mu \mathbf{1}_{M_0}) \Bigg\}. \end{aligned} \tag{9}$$

In the case of two-dimensional location data, $\boldsymbol{\beta}$ has seven dimensions, and direct optimization of $\boldsymbol{\beta}$ might be difficult. In Appendix A, we propose the maximum profile likelihood method that reduces the dimension of $\boldsymbol{\beta}$ and has better numerical stability.

Remark 1. For general Gaussian process regression, the inverse of the covariance matrix can become numerically unstable when the sample size n is large. In addition, the computational complexity is of the order of $O(n^3)$, which significantly increases as n increases. These problems are well recognized in the literature. On the other hand, because of the many nice properties of the Gaussian process, there have been significant developments on the large-scale computation of the Gaussian processes (Cressie and Johannesson, 2008; Haaland and Qian, 2011; Ranjan et al., 2011). These improvements have extended its application to very large datasets. However, in our application, we do not require a large data set to construct the AGP model. Also, as shown in Fig. 5, our covariance matrix contains block diagonal components. This structural advantage can also help to improve the numerical stability. Moreover, given the in-control sample data, our AGP model estimation the most expensive computations, such as inverting $\boldsymbol{\Sigma}_0$, only need to be performed once. All subsequent predictions only involve matrix multiplication. This further improves the computational stability.

Remark 2. It is also interesting to note that our AGP model is indeed different from the CGP model proposed by Ba and Joseph (2012), although they look similar to each other. The CGP model is mainly focused on tackling non-stationary simulation output. The authors use two Gaussian process covariance

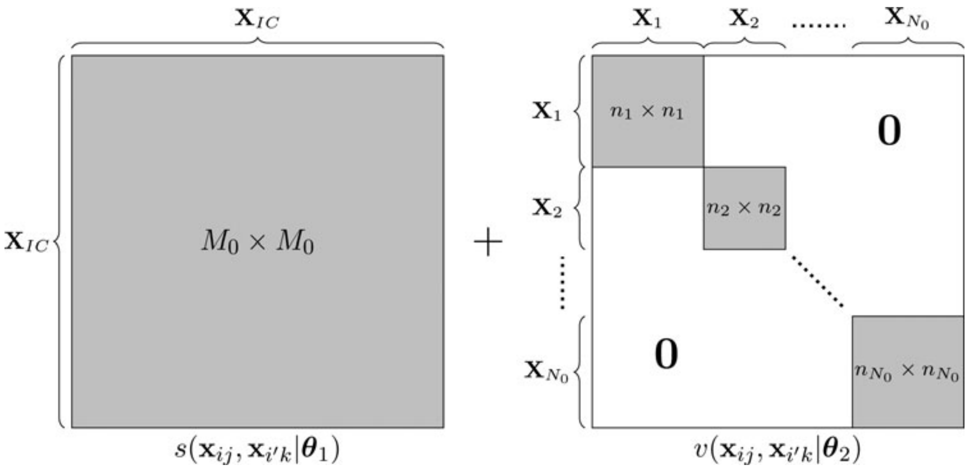


Figure 5. Structure of $\boldsymbol{\Sigma}_0$. The grey areas are the non-zero matrix blocks with corresponding dimensions indicated in the middle.

structures, one to model global correlation and one to model local correlation. Even though the resulting Gaussian process is still stationary, it performs much better in modeling heterogeneous simulation output compared with a single covariance structure. In fact, similar ideas were used in Haaland and Qian (2011), where even more layers of covariance structures were used to improve the accuracy. However, both models are used to approximate a single realization of a Gaussian process. In contrast, our model was motivated by a radically different setup. In our model, each surface corresponds to a different realization of the underlying Gaussian Process. As a result, the first component of the Gaussian Process is used to characterize the shared mean surface, whereas the second component of the Gaussian Process reflects the characteristics of the deviation surface. Different from the CGP model, which accepts measurements from one surface (realization) as input and finds covariance functions for global and local correlations, the AGP model needs measurements from multiple surfaces (realizations). It then estimates the common mean function and the distribution of the deviations from the mean. In summary, our AGP model is different from the CGP model in both motivation and mathematical details.

3. Statistical monitoring of geometric quality

The AGP model provides a quantification of the geometric profiles when the process is in control. Based on the model predictions, we can further setup control charts to monitor the geometric quality. In this article, we only consider simple Shewhart-type control charts. In other words, each new wafer is tested independently without information aggregation as in the Cumulative Sum (CUSUM) chart or Exponentially Weighted Moving Average (EWMA) chart. As a result, studying the Average Run Length (ARL) of the charts is equivalent to studying the α, β errors of the statistical testing procedure.

3.1. T^2 test

As previously mentioned, conditioned on the historical in-control measurements $\mathbf{Y}_{IC}$, the measurements on a new wafer follow a multivariate normal distribution if the process is in control. A natural choice to test whether or not $\mathbf{Y}_l$ conforms with the predicted distribution can be stated as

$$H_0 : \mathbf{Y}_l \sim N(\tilde{\boldsymbol{\mu}}_l, \tilde{\boldsymbol{\Sigma}}_l) \qquad H_1 : \mathbf{Y}_l \not\sim N(\tilde{\boldsymbol{\mu}}_l, \tilde{\boldsymbol{\Sigma}}_l).$$

A commonly used test statistic is the T^2 statistic: $T_l^2 = (\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l)^T \tilde{\boldsymbol{\Sigma}}_l^{-1} (\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l)$, which follows $\chi^2_{n_l}$ when H_0 is true (the process is in control). When T_l^2 is larger than the control limit H_T, we can reject H_0 (the process is out of control). The control limit can be determined such that the α error of the T^2 test meets a specified value ARL_0.

When the number of measurements taken from each wafer are different, the distribution of T_l^2 also varies according to n_l. In this case, we can use the p-value of the T^2 statistic as the monitoring statistic $p_l = 1 - \mathbb{F}_{\chi^2}(T_l^2|\nu)$, where $\mathbb{F}_{\chi^2}(\cdot|\nu)$ denotes the Cumulative Distribution Function (CDF) of the χ^2_ν distribution with ν degrees of freedom. When p_l is smaller than a given limit H_p, we can declare that the process is out of control.

Remark 3. In this article, we focus on Phase II analysis; i.e., we assume the parameters of the AGP model are known or have already been accurately estimated. In this case, the test statistics have an exact χ^2 distribution. Unfortunately, when the parameters of the model are unknown and need to be estimated from limited samples, the test statistics do not have known distributions. This issue is beyond the scope of this article, and we will investigate it in our future work.

3.2. GLR test

Despite the simplicity of the T^2 test, it is designed to detect omnibus changes. However, in our application, when changes are detected, we may want to further analyze the root causes of these changes. Therefore, the specific change types are expected to be known. Based on engineering knowledge, there are three typical change scenarios in surface fabrication: mean shift, variance change, and roughness change. Figure 6 demonstrates these change scenarios using one-dimensional curves as an example. Under this circumstance, a proposed GLR test that is able to provide change type information can be applied.

In this section, we illustrate the procedure using one example involving multiple types of changes. In more detail, we assume that when the process is out of control, another geometric deviation is added to the model ([1]), leading to

$$y_{lj} = f(\mathbf{x}_{lj}) + \epsilon_l(\mathbf{x}_{lj}) + \xi_l(\mathbf{x}_{lj}), \qquad \forall j = 1, 2, \cdots, n_l, \quad (10)$$

where $\xi_l(\mathbf{x})$ denotes the additional geometric deviation due to the out-of-control manufacturing process. Again, without prior knowledge on the forms of the deviation, we assume that it is an independent realization of another Gaussian process with mean δ and covariance function $w(\mathbf{x}_{lj}, \mathbf{x}_{lk}|\theta_l) = \gamma^2 \exp[-(\mathbf{x}_{lj} - \mathbf{x}_{lk})^T \times \mathrm{diag}(\theta_l) \times (\mathbf{x}_{lj} - \mathbf{x}_{lk})]$, the structure of which is consistent with Equations (3) and (4). More important, each parameter of this new Gaussian process component corresponds to different scenarios in Fig. 6. For example, the mean shifts lead to a non-zero δ; increased variance leads to larger γ^2, etc. For notational simplicity, we use $\mathbf{\Sigma}_w$ that depends on θ_l, γ^2 to represent the covariance matrix of $\xi_l(\mathbf{x})$ evaluated at locations $\mathbf{X}_l$.

According to this assumption, to test whether or not the wafer is in conformance is equivalent to testing whether or not the deviation $\xi_l(\mathbf{x})$ is significantly different from zero. In other words, the hypothesis can be stated as

$$H_0 : \mathbf{Y}_l \sim N(\tilde{\boldsymbol{\mu}}_l, \tilde{\mathbf{\Sigma}}_l), \quad H_1 : \mathbf{Y}_l \sim N(\tilde{\boldsymbol{\mu}}_l + \delta \mathbf{1}_{n_l}, \tilde{\mathbf{\Sigma}}_l + \mathbf{\Sigma}_w)$$
$$\text{for some non-zero}\delta, \gamma^2, \theta_l.$$

Consequently, the GLR statistic in this context can be expressed as

$$R_l = 2 \ln \frac{\sup_{\delta,\gamma^2,\theta_l} \det(\tilde{\mathbf{\Sigma}}_l + \mathbf{\Sigma}_w)^{-1/2} \exp\left[-(\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l - \delta \mathbf{1}_{n_l})^T (\tilde{\mathbf{\Sigma}}_l + \mathbf{\Sigma}_w)^{-1} (\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l - \delta \mathbf{1}_{n_l})/2 \right]}{\det(\tilde{\mathbf{\Sigma}}_l)^{-1/2} \exp\left[-(\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l)^T \tilde{\mathbf{\Sigma}}_l^{-1} (\mathbf{Y}_l - \tilde{\boldsymbol{\mu}}_l)/2 \right]}. \quad (11)$$

To find the distribution of R_l when H_0 is true (process is in control), we can reformulate the hypothesis as

$$H_0 : \ \delta = 0, \gamma^2 = 0 \qquad H_1 : \ \delta \neq 0, \text{ or } \ \gamma^2 > 0.$$

It is noted that when $\gamma^2 = 0$, θ_l is meaningless and does not need to appear in H_0. In other words, we only require the non-negative θ_l in both H_0 and H_1. Since the condition that $\gamma^2 = 0$ in H_0 is on the boundary of the parameter space, R_l approximately follows a 50{: 50% mixture of χ_1^2 and χ_2^2 distribution when n_l is large according to Self and Liang (1987). In other words, the CDF of R_l can be expressed as

$$\mathbb{P}(R_l \leq t) = 0.5 \times \mathbb{F}_{\chi^2}(t|1) + 0.5 \times \mathbb{F}_{\chi^2}(t|2).$$

It is noted that the distribution of R_l does not depend on n_l or $\mathbf{X}_l$. Therefore, the same control limit can be used regardless of the number or locations of the measurements. When R_l is larger than the control limit H_R, we can reject H_0 (the process is out of control).

Remark 4. When an out-of-control signal is received, parameters $(\delta, \gamma^2, \theta_l)$ in statistic R_l can be used to diagnose the specific type of change. The GLR test generally performs well when the changes from H_0 are sufficiently characterized by the alternative hypothesis. However, when the changes are different from the types stated in the alternative hypothesis, the performance of the GLR test might not be satisfactory. The statistic (11) is only one possible choice of the GLR statistic and is designed to characterize shifts occuring globally on the surface. If the process engineers have relevant knowledge on the shift patterns and regions, a more specific alternative hypothesis could be used in Equation (10), and the GLR method can still be applied based on the new hypothesis.

Remark 5. Generally speaking, the T^2 test does not require prior knowledge on the possible shift patterns. It is suitable when knowledge on shift patterns is limited. On the other hand, the GLR test is developed to detect specific types of change determined through the alternative hypothesis. A specifically designed GLR test is expected to be more powerful than a T^2 test toward certain types of changes. Unfortunately, the increase in its detection power comes at a price: it is not robust against other types of changes. As a result, which test to use largely depends on the practical problems and available information.

4. Simulation studies

In this section, we present some simulation studies to demonstrate the effectiveness of our proposed method. We first illustrate that the AGP model is indeed effective in approximating complex profiles, and the estimation procedure is numerically

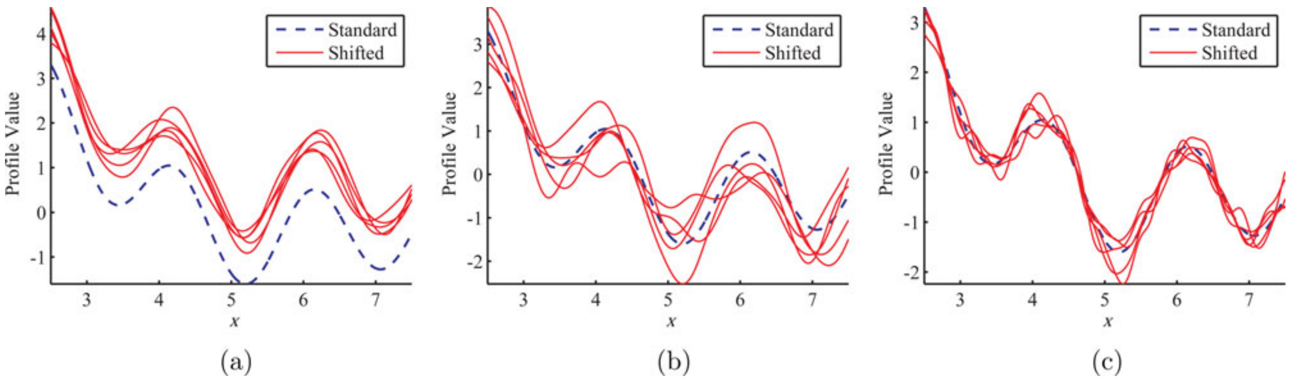
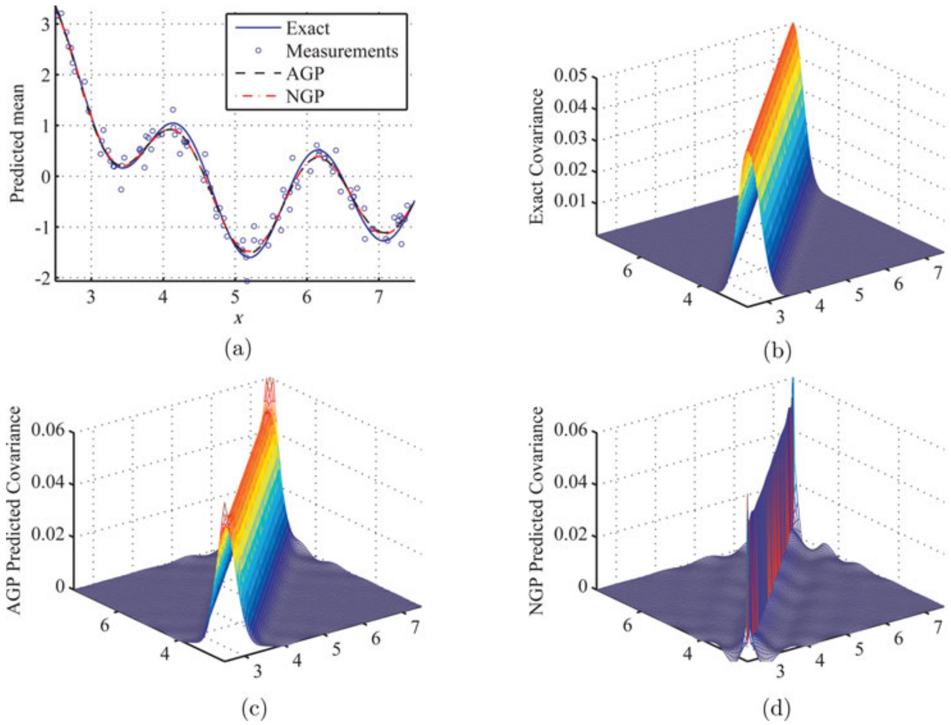


Figure 6. Three typical change scenarios when a process is out of control: (a) mean shift; (b) variance change; and (c) roughness change.



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

Figure 7. Predicted means and covariances using AGP and NGP models: (a) exact and predicted means; (b) exact covariance; (c) predicted covariance using AGP; and (d) predicted covariance using NGP.

stable. Then we further analyze the performance of the statistical monitoring schemes based on the AGP model. For easier illustration, we use a one-dimensional curve instead of a two-dimensional geometric profile in the demonstrations in this section.

4.1. *Approximation by the AGP and its estimation performance*

We first demonstrate that the AGP model is sufficient to approximate the complex standard profile and quantify the in-control variations from a group of samples with spatially correlated errors. In the simulation, we use a one-dimensional function (Shpak, 1995)

$$y = \sin(x) + \sin(10x/3) + \ln(x) - 0.84x + 3, \qquad 2.5 \le x \le 7.5 \tag{12}$$

as the standard curve. The spatially correlated error $\epsilon(x)$ was generated from a Gaussian process with mean $\eta = 0$, $\tau^2 = 0.05$, and $\theta_2 = 5$ in its covariance function (4). We generated $N_0 = 10$ in-control curves, with $n_i = 10$ measurements taken from each curve. The measurement locations were randomly selected according to a Latin Hypercube Sampling (LHS) strategy. We used the maximum profile likelihood method to estimate the parameter $\hat{\boldsymbol{\beta}}$ for the AGP model and then predict y at different locations.

Table 1. Bias and RMSE of the MLE of the AGP model.

(N_0, n_0)		$\mu = 1$	$\sigma^2 = 0.2$	$\theta_1 = 3$	$\tau^2 = 0.05$	$\theta_2 = 10$
(10,10)	Bias	−0.0043	−0.0189	0.4375	−0.0002	0.7089
	RMSE	0.1824	0.1001	1.6348	0.0080	4.3834
(10,20)	Bias	−0.0013	−0.0189	0.1756	0.0001	0.0011
	RMSE	0.1831	0.0975	0.9608	0.0066	0.9204
(20,10)	Bias	0.0106	−0.0103	0.2528	0.0000	0.4140
	RMSE	0.1903	0.1038	1.1990	0.0056	3.1826
(20,20)	Bias	0.0015	−0.0169	0.1317	0.0002	0.0001
	RMSE	0.1850	0.0920	0.7562	0.0045	0.5976

As a comparison, we use a Gaussian process with noisy observations (Rasmussen and Williams, 2006), which we call the Noisy Gaussian Process (NGP) model in this article. The NGP model assumes $y_{ij} = f(\mathbf{x}_{ij}) + \epsilon, i = 1, 2, \cdots, N_0; j = 1, 2, \cdots, n_i$ to fit the data. More specifically, it still uses a Gaussian process to approximate the standard profile $f(\mathbf{x}_{ij})$ but it simply uses an i.i.d. noise to model the deviations between individual samples and the standard profile. Figure 7 compares the predicted means and covariances at different locations using both AGP and NGP models.

It is shown in Fig. 7(a) that the difference between the mean predictions from AGP and NGP is not significant. Both predictions are very close to the exact function. More quantitatively, the mean prediction from the AGP model has an Integrated Mean Squared Error (IMSE) of 0.0052, whereas the IMSE of NGP is 0.0077. However, when predicting covariance, Fig. 7(d) clearly shows that the NGP model failed to predict the correct structure. This is simply because our simulation model includes a correlated noise, whereas the NGP model with an i.i.d. noise assumption will no longer closely fit the simulated data. In contrast, covariance prediction from AGP (Fig. 7(c)) is much closer to the exact case (Fig. 7(b)). This comparison demonstrates that although the NGP is equally effective in mean prediction, our AGP model is overall more appropriate to model complex profiles with spatially correlated deviations.

We also conducted extensive simulations to show that the estimation procedure of the AGP model is accurate and numerically stable. In each simulation replication, we generated N_0 curves from an AGP model with parameters $\boldsymbol{\beta} \equiv [\mu = 1,$

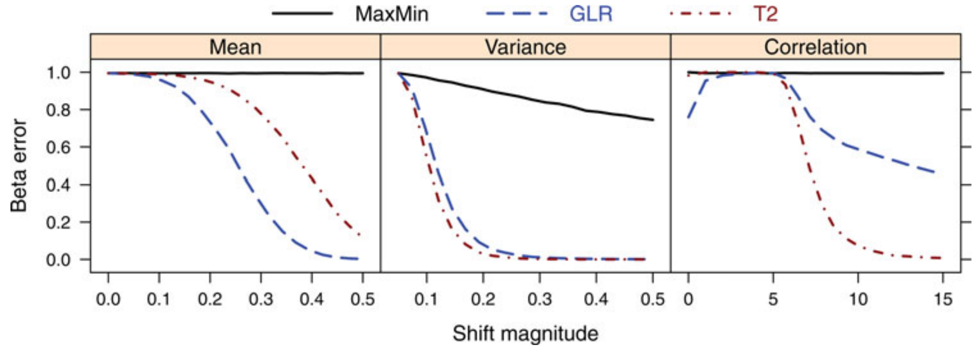


Figure 8. Comparisons of β error in detecting different types and magnitudes of shifts.

$\sigma^2 = 0.2, \theta_1 = 3, \tau^2 = 0.05, \theta_2 = 10]$. n_0 measurements were taken from each curve based on the LHS strategy. From these data, the AGP parameters were estimated using the profile Maximum Likelihood Estimator (MLE) as presented in Appendix A. We repeated this procedure for $K = 1000$ times for different pairs of (N_0, n_0) and calculated the bias and Root Mean Squared Error (RMSE) of each component inside $\hat{\boldsymbol{\beta}}$. Using μ as an example:

$$\text{Bias}(\hat{\mu}) = \frac{1}{K}\sum_{i=1}^{K}(\hat{\mu}_i - \mu),$$

$$\text{RMSE}(\hat{\mu}) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\hat{\mu}_i - \mu)^2},$$

where i is the replication index. Calculations for the other components followed the same manner. The numerical results are summarized in Table 1.

It clearly shows that increasing the sample size, either larger N_0 or larger n_0, can generally reduce the bias and RMSE. In addition, for the same sample size $N_0 \times n_0$, it is more helpful in terms of estimation to have a larger n_0 rather than a larger N_0. Table 1 also reveals that the parameters of the second Gaussian process component are much easier to estimate than that of the first one. In addition, it is expected that a more sophisticated selection method of the measurement locations could further improve the estimation performance.

4.2. *Performance of statistical monitoring schemes*

4.2.1. *Known in-control parameters β*

In this section, we further investigate the performance of the statistical monitoring schemes proposed in Section 3. We start by studying the charting performance when the standard function $f(x)$ and parameters of $\epsilon(x)$ are known exactly. In this case, when the process is in control, the measurements $\mathbf{Y}_l$ at locations $\mathbf{X}_l$ follow normal distribution with mean $\mathbf{f}_l \equiv [f(\mathbf{x}_{l1}), f(\mathbf{x}_{l2}), \cdots, f(\mathbf{x}_{ln_l})]^T$ and covariance matrix $\boldsymbol{\Sigma}_l = [v(\mathbf{x}_{li}, \mathbf{x}_{lj}|\theta_2)]_{n_l \times n_l}$. It is noted that since $f(x)$ is exactly known, the first Gaussian process component in the AGP model is not needed, and $\boldsymbol{\Sigma}_l$ is obtained using a single covariance function. Consequently, the T^2 statistic becomes $TE_l^2 = (\mathbf{Y}_l - \mathbf{f}_l)^T \boldsymbol{\Sigma}_l^{-1}(\mathbf{Y}_l - \mathbf{f}_l)$. Similarly, the GLR statistic can be

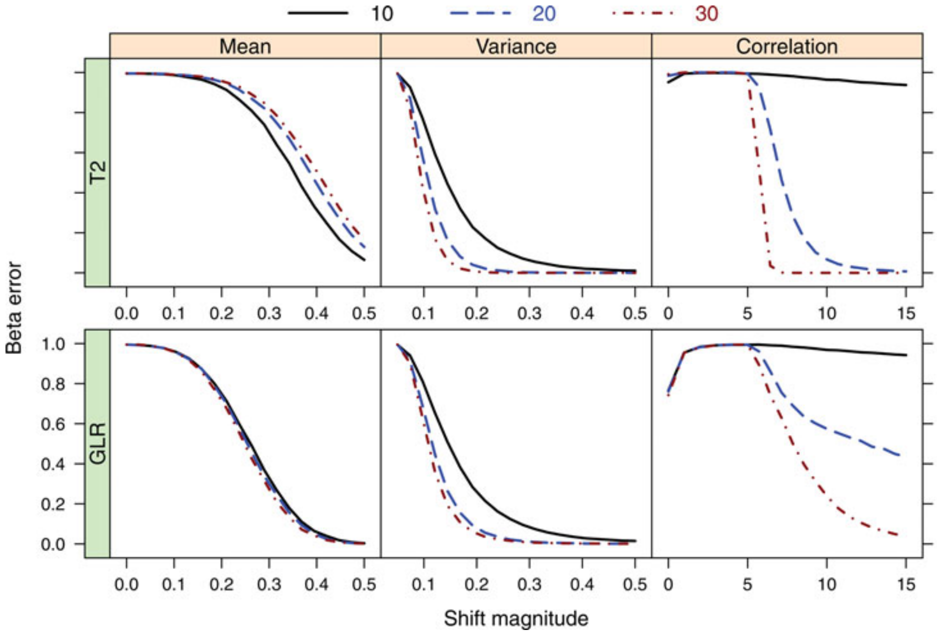


Figure 9. Comparison of detection performance when using a different number of measurements on each wafer.

simplified as

$$RE_l = 2\ln \frac{\sup_{\delta,\gamma^2,\theta_l} \det(\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_w)^{-1/2} \exp\left[-(\mathbf{Y}_l - \mathbf{f}_l - \delta \mathbf{1}_{n_l})^T (\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_w)^{-1} (\mathbf{Y}_l - \mathbf{f}_l - \delta \mathbf{1}_{n_l})/2\right]}{\det(\boldsymbol{\Sigma}_l)^{-1/2} \exp\left[-(\mathbf{Y}_l - \mathbf{f}_l)^T \boldsymbol{\Sigma}_l^{-1} (\mathbf{Y}_l - \mathbf{f}_l)/2\right]}.$$

For comparison, we also consider another simple method called the Max-Min statistic, which is currently used in the semiconductor industry to monitor the thickness profile of a wafer. The Max-Min statistic calculates the difference between the largest and smallest measurements among all measured locations on each wafer. When the difference exceeds a certain limit H_M, the process is considered to be out of control.

In this simulation, we used the same standard curve (12) and noise process as in Section 4.1. For each wafer to be tested, we randomly selected 20 locations to measure based on the LHS strategy. The control limits of the T^2 test and the GLR test can be analytically obtained from the χ^2_{20} distribution and the 50{: 50% mixture of χ^2_1 and χ^2_2 distribution, respectively. However, the control limit of the Max-Min statistic can only be obtained through simulation. In the simulation, we chose the control limits such that the alpha error $\alpha = 0.01$, and the corresponding control limits were $H_T = 37.57, H_R = 8.27,$ and $H_M = 5.39$.

Using these control limits, we compared the performance of the three tests in detecting different types of changes, including mean shifts, variance shifts, and correlation shifts. This translates to the changes in η, τ^2, θ_2 from the values listed in Section 4.1. In each change scenario, we estimated the β error of the tests using 20 000 simulation replications. The Operation Characteristic (OC) curves of different types of shifts are shown in Fig. 8.

It shows that both the T^2 test and GLR test can effectively detect all three types of changes. However, the Max-Min test is not able to detect mean shifts or correlation shifts because the difference between the largest and smallest measurement values remains the same (in distribution) when η or θ_2 changes. The Max-Min method also has a much larger β error in detecting variance shifts. Furthermore, GLR test is much more sensitive than the T^2 test in detecting the mean shift. However, it is not as good as the T^2 test in detecting variance shifts in spite of the small difference. When detecting correlation shifts, the GLRtest can detect both increasing and decreasing θ_2, whereas the T^2 test is only able to detect increasing θ_2, which corresponds to an increased roughness of the geometric profile. On the other hand, the T^2test performs much better than the GLR test in detecting large shifts of θ_2. This is mainly due to the fact that when θ_2 changes, the scenario is different from the alternative hypothesis (10) stated in the GLR test. As a result, the general-purpose GLR test is not very effective. However, if we are interested in faster detection of the correlation shifts, we can improve the GLR test by changing the alternative hypothesis. As we remarked in Section 3.2, in general the T^2 test is easy to implement and capable of detecting multiple types of changes without assumptions on the change scenario, whereas the GLR test is more complex yet flexible enough to be able to cater for different detection requirements and able to provide additional information on change type. Overall both the T^2 and GLR tests have their own advantages. We recommend that users choose between them based on practical considerations.

It is also interesting to note that the β error of the GLR test cannot decrease to zero even for a large magnitude of θ_2 shift. This might be explained by the Nyquist–Shannon sampling theorem (Shannon, 1949). Recall that the GLR test needs to estimate the shifted parameters from the data. Thus, with a finite sample size (20 in the simulation), it is unable to estimate large θ_2 values that correspond to high frequency changes in the observations. To confirm this, we chose different numbers of measurements ($n_l = 10, 20,$ and 30) on each wafer and adjusted the control limits such that their α errors were equal. The performance of these tests when different n_l is used is compared in Fig. 9.

It clearly indicates that a larger n_l leads to a better performance in detecting correlation shifts for both the T^2 test and the GLR test. Their performance in detecting variance shifts is also improved with a larger n_l, although the improvement is relatively small. As a result, if the correlation changes

Table 2. Average and standard error (in parentheses) of the real α error of GLR and T^2 tests with nominal $\alpha_0 = 0.05, 0.01$.

		$\alpha_0 = 0.05$		$\alpha_0 = 0.01$	
(N_0, n_0)	n_l	GLR	T^2	GLR	T^2
(10, 10)	10	0.051 (0.0020)	0.046 (0.0019)	0.019 (0.001 36)	0.020 (0.001 26)
	20	0.056 (0.0021)	0.035 (0.0022)	0.036 (0.002 15)	0.023 (0.001 68)
	30	0.051 (0.0021)	0.022 (0.0019)	0.032 (0.002 02)	0.017 (0.001 88)
(10, 20)	10	0.047 (0.0016)	0.050 (0.0018)	0.013 (0.001 12)	0.014 (0.000 84)
	20	0.056 (0.0016)	0.049 (0.0016)	0.019 (0.001 29)	0.021 (0.001 56)
	30	0.048 (0.0015)	0.025 (0.0017)	0.017 (0.001 33)	0.007 (0.000 62)
(20, 10)	10	0.049 (0.0016)	0.052 (0.0017)	0.013 (0.000 79)	0.015 (0.000 91)
	20	0.051 (0.0015)	0.041 (0.0019)	0.020 (0.001 31)	0.015 (0.001 28)
	30	0.046 (0.0014)	0.022 (0.0020)	0.021 (0.001 60)	0.012 (0.001 51)
(20, 20)	10	0.044 (0.0012)	0.048 (0.0014)	0.010 (0.000 51)	0.012 (0.000 69)
	20	0.049 (0.0014)	0.048 (0.0014)	0.013 (0.000 69)	0.016 (0.001 06)
	30	0.046 (0.0016)	0.011 (0.0009)	0.011 (0.000 66)	0.002 (0.000 21)

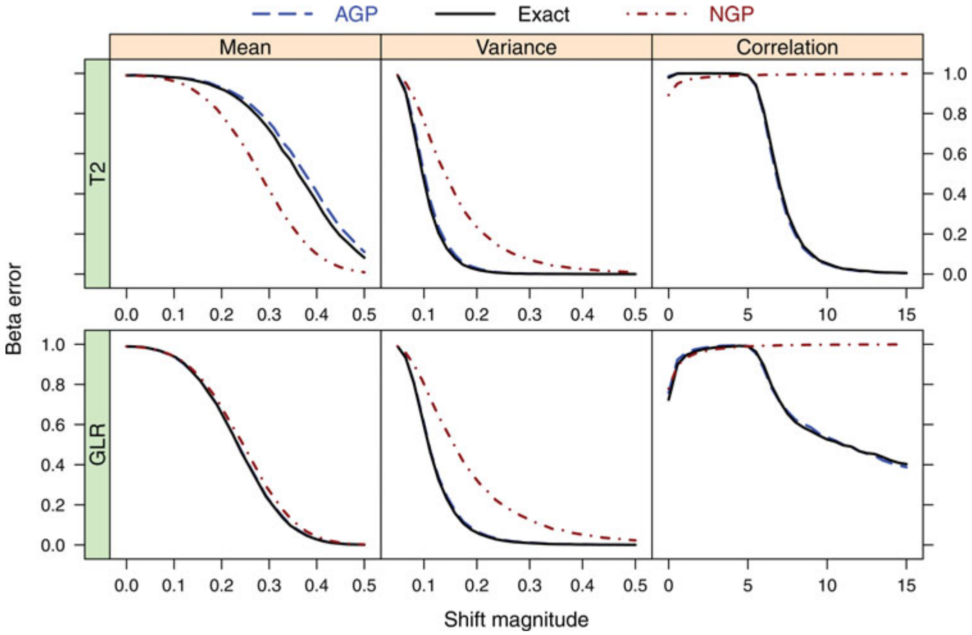


Figure 10. β errors of the T^2 and GLR tests when different models are used, with $N_0 = 20$, $n_0 = 20$, $n_l = 20$. Comparisons with different sample sizes are included in the online supplementary material.

(roughness changes) are of major interest, it is better to take more measurements from each wafer to conduct the statistical tests.

4.2.2. *Unknown in-control parameters*

The simulation studies in Section 4.2.1 compared the testing performance in the ideal case in which both $f(x)$ and the process parameters of $\epsilon(x)$ are known exactly. In practice, this information is typically unknown, and we need to use the AGP model and corresponding tests developed in Section 2.3 and Section 3. As we would expect, with less information the performance is not as good as the case in Section 4.2.1.

When the AGP model parameters $\boldsymbol{\beta}$ are known and predictions are correct, the T^2 test statistic exactly follows the $\chi^2_{n_l}$ distribution, whereas the GLR test statistic (11) asymptotically follows the mixture χ^2_1 and χ^2_2 distribution when n_l is large. However, when the parameters are estimated from historical data, these results are no longer valid. Simply using the critical values derived from the theoretical distributions can lead to different α errors from those designed (Jensen et al., 2006). To investigate the impact of parameter estimation and model approximation on the real α error, we designed the following simulation study.

In this simulation, we chose different combinations of in-control samples (N_0, n_0). We also considered different numbers of testing samples $n_l = 10, 20$, and 30. In each replication, the in-control data were generated based on the specific setting, and

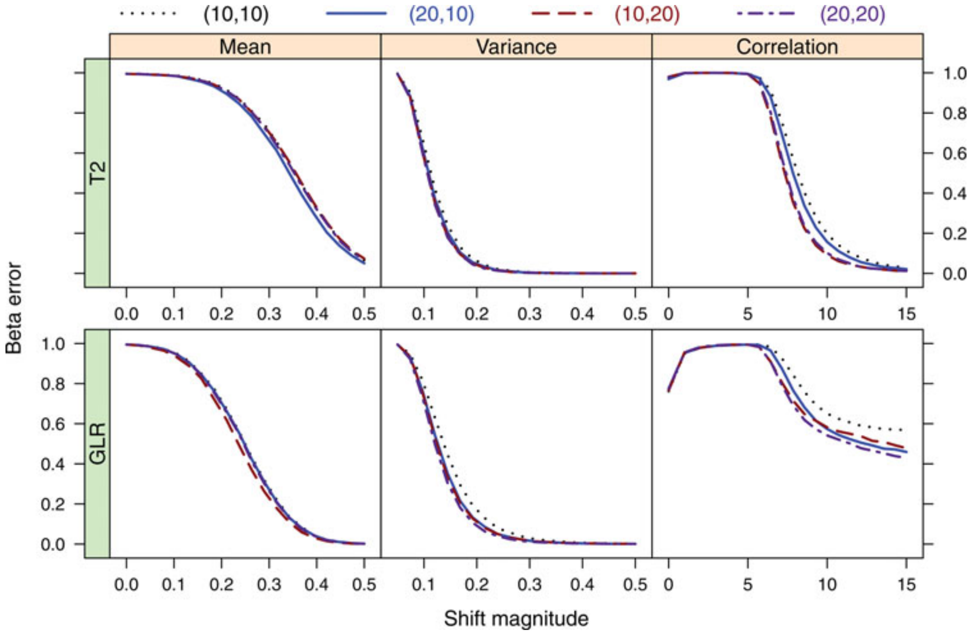


Figure 11. OC curves using different in-control sample sizes for the AGP model. The number of testing samples $n_l = 20$.

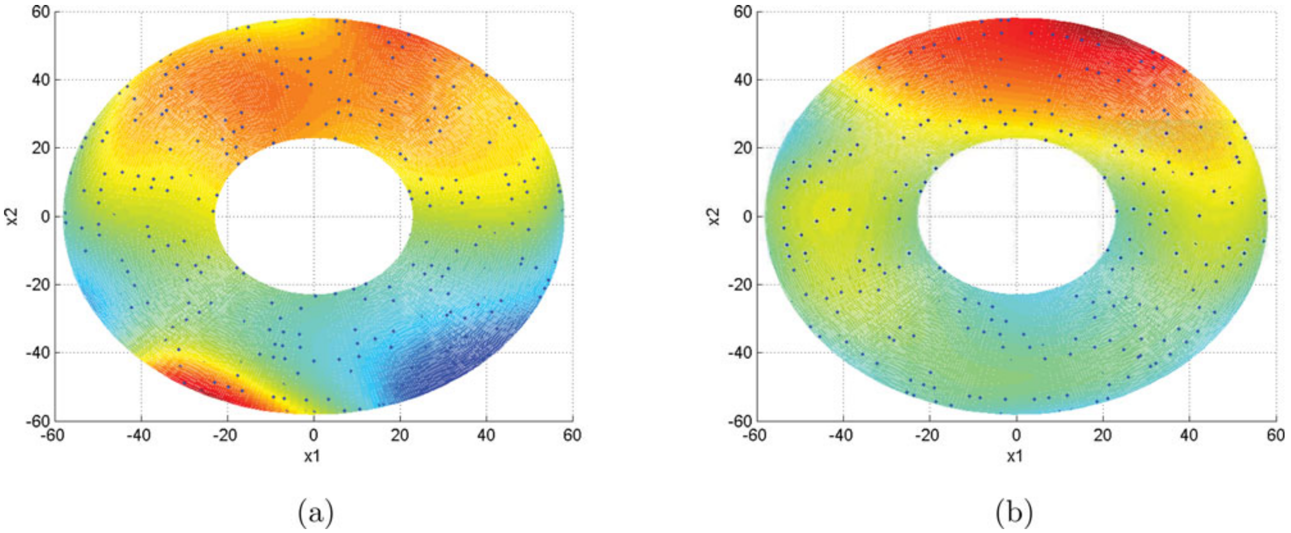


Figure 12. Examples of wafer thickness predictions using a single Gaussian process model: (a) wafer no. 2 and (b) wafer no. 7.

the parameters of the AGP model were estimated from the data. When the critical values were determined based on the theoretical distributions with nominal error α_0, we were able to estimate the real α error using 20 000 testing replications. This procedure was repeated 200 times, and the mean and standard error of the real α error are reported in Table 2.

The table reveals that more in-control samples or larger $M_0 = N_0 \times n_0$ can indeed lead to a smaller discrepancy between the real α error and the nominal α error. In particular, a larger N_0 rather than larger n_0 is more effective in reducing the difference. Also, it appears that n_l has a negative effect on the accuracy of the α error, especially when $n_l > n_0$. If a smaller α_0 is required, more in-control samples are needed to compensate for the effects of the estimated parameters.

When comparing their performance in detecting different types of shifts, we adjusted the control limits such that all of the tests had the same *real* α error of 0.01. Both the T^2 test and GLR test were constructed from each of the three cases: Exact, AGP, and NGP. Here Exact refers to the case when all the in-control parameters are known as in Section 4.2.1. The OC curves of these tests in detecting different shifts are shown in Fig. 10.

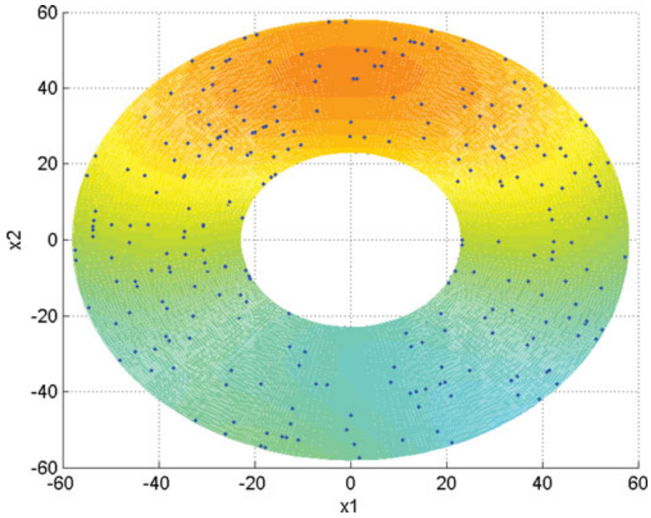


Figure 13. Predicted standard thickness profile using the AGP model.

It shows that the differences between the AGP model and the exact scenario are small in most cases. In contrast, tests based on the NGP model are generally worse than those based on the AGP model except for the mean shifts scenario. As compared in Section 4.1, the NGP model is as effective as the AGP model in prediction of the mean. Therefore, their GLR test performances in the mean-shift scenarios do not differ too much. However, as we can observe from the figure, the T^2 test using the NGP model seems to be better than that using the AGP model and even Exact case. This is because the NGP model does not account for spatial correlation between samples. As a result, the T^2 statistic is more sensitive when the mean changes in one direction. A similar phenomenon in the context of time series data monitoring has been reported in Zhang (1998). He found that when the time series are correlated, directly monitoring the samples (without accounting for the correlation) has a better performance in detecting mean shifts than monitoring the residuals (uncorrelated when the time series model is correct). In practice, however, it is impractical to adjust control limits because we cannot use simulation to evaluate the real α error. When using the limits based on nominal α errors, the real α error can be far away from the nominal values. Similar findings have been reported in the literature (Neuhardt, 1987; Montgomery et al., 1991). In summary, the presented results confirm that the AGP model is more suitable for approximating complex profiles with spatially correlated errors. Consequently, charts based on AGP models generally have a better performance in monitoring a variety of shifts.

Similar to the influence on α error when the in-control parameters are unknown, the number of in-control samples N_0, n_0 also has an impact on the detection performance. Using the same pairs of (N_0, n_0), we compare the OC curves of the tests constructed based on the AGP model in Fig. 11.

Again, it shows that a larger in-control sample size generally leads to a faster detection of different shifts. The improvement is especially evident in detecting variance and correlation shifts. In contrast, only a marginal improvement can be observed in

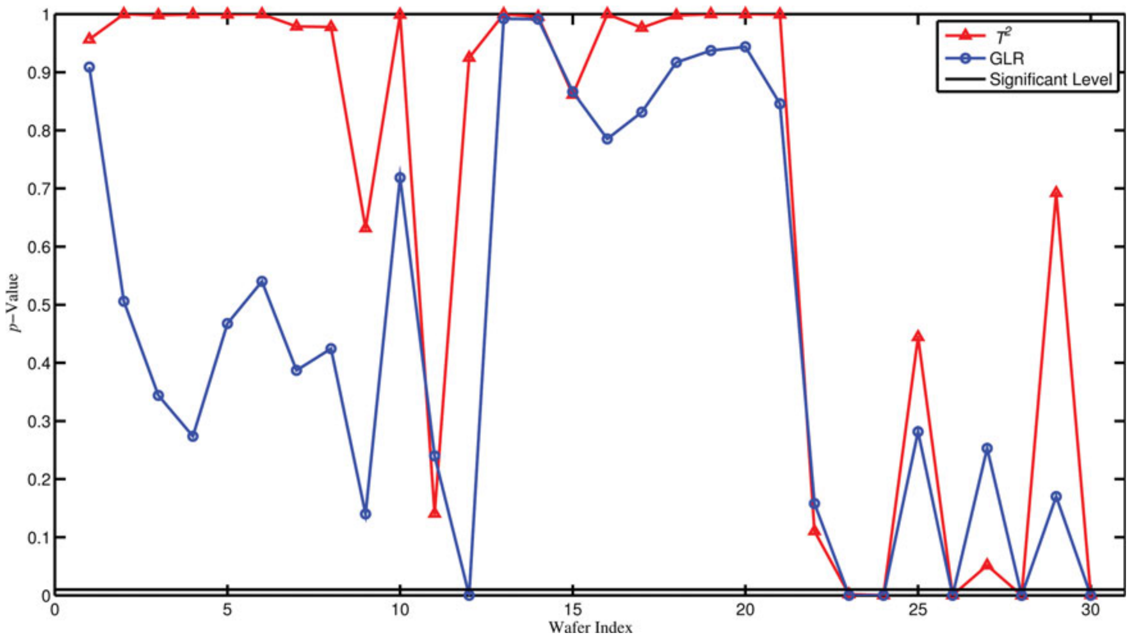


Figure 14. The p-values of the T^2 test and GLR test on the remaining 30 wafers.

detecting mean shifts. Together with earlier findings in this section, we can conclude that more in-control samples are beneficial for the monitoring performance and result in a more accurate α error and lower β error. In the case with limited in-control samples, a self-starting strategy (Sullivan and Jones, 2002) can be used to continuously improve the AGP model estimation during the monitoring process.

5. Application to the monitoring of wafer thickness profiles

In this section, we apply our proposed method to monitor the thickness profile of a silicon wafer after the slicing process. The data were collected from a real semiconductor fabrication plant (with preprocessing to remove sensitive information). A total of 38 wafers were measured, among which 8 wafers were identified as normal and used as in-control samples.

Although a perfectly flat wafer surface is desired, variations in the manufacturing process often cause roughness in its thickness profile. However, as long as the deviation from a flat surface is acceptable, the surface can be considered as being in control. Figure 12 demonstrates heat maps of the Gaussian process–predicted thickness profile of two different in-control wafers. Each prediction used 480 measurements at different locations. These two heat maps together with subsequent heat maps use the same color scale as in Fig. 1. Figure 12 clearly indicates that the wafer is not as flat as we expect and not very smooth due to process variations. More important, the thickness profiles of the in-control samples are quite different, which makes it a challenge to monitor the geometric quality using existing methods.

To monitor the remaining wafers, we used the eight in-control wafers to fit the AGP model. We collected 60 measurements from each of the in-control samples using a space-filling sampling strategy. These data were used to estimate the parameters of the AGP model. The MLE values were obtained as $\hat{\mu} = -0.0159$, $\hat{\sigma}^2 = 0.0043$, $\hat{\theta}_1 = [1.29 \times 10^{-4}, 3.82 \times 10^{-4}]$, $\hat{\tau}^2 = 0.0022$, $\hat{\theta}_2 = [0.0051, 0.0061]$. Figure 13 shows the predicted standard thickness profile from the AGP model. It is interesting to note that the standard profile is not a simple flat surface. This is because the raw silicon ingot is highly likely to undergo stress deformation during the slicing process, which results in the slicing direction being non-perpendicular to the axial direction of the ingot. Therefore, the thickness profile of a sliced wafer turns out to have a specific geometric feature, whose common pattern can be depicted by the standard profile. Despite the non-flatness, compared with Fig. 12 we can observe that the standard profile is much smoother because the process variation has been filtered out from the standard profile in the AGP model. In addition, the deviations between each in-control sample and the standard profile were obtained and used to quantify the process variations. These eight deviation profiles were numerically inspected. The results indicate that the spatial patterns of these deviations are similar and consistent with the assumptions of the AGP model. Please refer to the supplementary material for more details.

Based on the estimated AGP model, we can use the T^2 test and GLR test to determine whether or not the remaining 30 wafers are in conformance. From each wafer to be tested, 120 measurements were taken using the space-filling sampling strategy. Both test statistics were calculated using the procedures discussed in Section 3. To compare the T^2 test and the GLR test, we converted the test statistics to p-values, and the results are shown in Fig. 14. The test results indicate that most of the wafers is conform to the standard with acceptable variations. However, there are also a few wafers that fail both tests with $\alpha = 0.01$. In Fig. B1 in Appendix B, these thickness profiles are shown in more detail, and they indeed display discrepancies from the standard profile and other in-control wafers. These failed wafers are either much thicker or thinner in particular regions and overall much rougher compared with in-control samples and the standard profile. We notice that wafer no. 12 failed GLR but passed T^2. Analyzing the parameters of the GLR statistic we found that it is the variance shift causing that wafer 12 to fail the GLR test. As we can observe from Fig. B1(a), the thickness profile in the northeast region fluctuates a lot, which increases the overall variance. If the measurements happened to miss a region or did not sufficiently investigate a region, this test might not detect an abnormality on this wafer. This issue also raises the importance of sampling strategy. It is expected that by adaptively selecting the measurement points based on existing sampling information, the detection performance can be enhanced. We also reported the performance of the charts based on the NGP model in the supplementary material. Interested readers can refer to it for more discussion. In short, since the NGP model does not account for the spatial correlation of the data, and it is infeasible to adjust the control limit in practice through simulation, a significant number of false alarms are expected.

6. Conclusions and future directions

This article presented a systematic method to monitor the geometric quality of a wafer. We proposed an AGP model to approximate the unknown standard geometric profile and quantify the spatially correlated deviations during an in-control manufacturing process. Based on the AGP model, we developed two statistical tests, namely, the T^2 test and GLR test to determine whether or not newly produced wafer is conforming. Numerical simulations and real case studies have demonstrated that the proposed method is effective.

There are several topics worth further investigation. First of all, as demonstrated in Section 4.2.2, estimating the AGP parameters from the in-control samples often leads to an inaccurate α error of the developed test. This problem is especially important if the number of in-control samples is limited. Therefore, adjusting the control limit to account for parameter uncertainty may improve the accuracy of the test. Second, in this research all of the measurements are randomly sampled using a space-filling strategy such as an LHS plan. However, a more sensible way is to sequentially determine the measurement locations based on the current AGP model and detection objective. This adaptive sampling strategy is expected to improve the efficiency and effectiveness of the tests. Finally, other types of monitoring schemes that can aggregate information from multiple wafers can be investigated to allow faster detection of changes.

Acknowledgement

We would like to thank the reviewers and the editor for their constructive comments.

Funding

Kaibo Wang is partially supported by the National Natural Science Foundation of China under grants 71471096 and 71072012. Nan Chen is partially supported by the Singapore AcRF grant R-266-000-078-112.

Notes on contributors

Linmiao Zhang is currently a Ph.D. student in the Department of Industrial & Systems Engineering, National University of Singapore. He received his B.Eng. degree in Industrial Engineering from Nanjing University, China. His research topic is statistical modeling of complex engineering data. He is a student member of INFORMS.

Kaibo Wang is an Associate Professor in the Department of Industrial Engineering, Tsinghua University, Beijing, China. He received his B.S. and M.S. degrees in Mechatronics from Xi'an Jiaotong University, Xi'an, China, and his Ph.D. in Industrial Engineering and Engineering Management from the Hong Kong University of Science and Technology, Hong Kong. He has published papers in journals such as *IEEE Transactions on Automation Science and Engineering*, *Journal of Quality Technology*, *IIE Transactions*, *Quality and Reliability Engineering International*, *International Journal of Production Research*, and others. His research is devoted to statistical quality control and data-driven complex system modeling, monitoring, diagnosis, and control, with a special emphasis on the integration of engineering knowledge and statistical theories to solve problems from industry. He is a member of INFORMS and IIE and a senior member of ASQ.

Nan Chen is an Assistant Professor in the Department of Industrial and Systems Engineering at the National University of Singapore. He obtained his B.S. degree in Automation from Tsinghua University, M.S. degree in Computer Science, M.S. degree in Statistics, and Ph.D. degree in Industrial Engineering from the University of Wisconsin–Madison. His research interests include statistical modeling and surveillance of service systems, simulation modeling design, condition monitoring, and degradation modeling. He is a member of INFORMS, IIE, and IEEE.

References

Ankenman, B., Nelson, B. and Staum, J. (2010) Stochastic kriging for simulation metamodeling. *Operations Research*, **58**, 371–382.

Ba, S. and Joseph, V.R. (2012) Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, **6**, 1838–1860.

Cressie, N. (1993) *Statistics for Spatial Data*, Wiley,New York, NY.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 209–226.

De Boor, C. (2001) *A Practical Guide to Splines*, SpringerVerlag, New York, NY.

Doering, R. and Nishi, Y. (2007) *Handbook of Semiconductor Manufacturing Technology*, CRCPress, Boca Raton, FL.

Haaland, B. and Qian, P.Z. (2011) Accurate emulators for large-scale computer experiments. *The Annals of Statistics*, **39**, 2974–3002.

Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry. *Statistical Science*, **8**, 120–129.

Huang, D., Allen, T., Notz, W. and Zeng, N. (2006) Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, **34**, 441–466.

Jensen, W.A., Birch, J.B. and Woodall, W.H. (2008) Monitoring correlation within linear profiles using mixed models. *Journal of Quality Technology*, **40**, 167–183.

Jensen, W.A., Jones-Farmer, L.A., Champ, C.W. and Woodall, W.H. (2006) Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality Technology*, **38**, 349–364.

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

Jones, D., Schonlau, M. and Welch, W. (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**, 455–492.

Montgomery, D., Mastrangelo, C., Faltin, F.W., Woodall, W.H., MacGregor, J.F. and Ryan, T.P. (1991) Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, **23**, 179–193.

Neuhardt, J.B. (1987) Effects of correlated sub-samples in statistical process control. *IIE Transactions*, **19**, 208–214.

O'Mara, W.C., Herring, R.B. and Hunt, L.P. (1990) *Handbook of Semiconductor Silicon Technology*, NoyesPublications, Park Ridge, NJ.

Ranjan, P., Haynes, R. and Karsten, R. (2011) A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics*, **53**, 366–378.

Rasmussen, C.E. and Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*, MITPress, Boston, MA.

Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989) Design and analysis of computer experiments. *Statistical Science*, **4**, 409–423.

Schmitz, T., Davies, A., Evans, C. and Parks, R. (2003) Silicon wafer thickness variation measurements using the National Institute of Standards and Technology infrared interferometer. *Optical Engineering*, **42**, 2281–2290.

Self, S.G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Shannon, C.E. (1949) Communication in the presence of noise. *Proceedings of the IRE*, **37**, 10–21.

Shpak, A. (1995) Global optimization in one-dimensional case using analytically defined derivatives of objective function. *Computer Science Journal of Moldova*, **3**, 168–184.

Sullivan, J.H. and Jones, L.A. (2002) A self-starting control chart for multivariate individual observations. *Technometrics*, **44**, 24–33.

Zhang, N.F. (1998) A statistical control chart for stationary process data. *Technometrics*, **40**, 24–38.

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**, 21012.1–21012.9.

Zou, C., Tsung, F. and Wang, Z. (2007) Monitoring general linear profiles using multivariate exponentially weighted moving average schemes. *Technometrics*, **49**, 395–408.

Zou, C., Zhou, C., Wang, Z. and Tsung, F. (2007) A self-starting control chart for linear profiles. *Journal of Quality Technology*, **39**, 364–375.

Appendixes

A. *Maximum profile likelihood of the AGP model*

To estimate the parameters of the AGP model from in-control measurements, we need to maximize the likelihood function (9). However, direct optimization is easily trapped in local optima. The scales of each dimension are also generally quite different, making the optimization more difficult. To improve the optimization performance, we can reduce its dimension by maximizing the profile likelihood.

In more details, given $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the two correlation matrices are completely determined. The first correlation matrix is denoted as $\mathbf{S}$, with elements $s(\mathbf{x}_{ij}, \mathbf{x}_{i'k}|\boldsymbol{\theta}_1)$. Because of the independence of $\epsilon_i(\mathbf{x})$ for different i, the second correlation matrix is a block diagonal matrix $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{N_0})$, where $\mathbf{V}_i$ has $v(\mathbf{x}_{ij}, \mathbf{x}_{ik}|\boldsymbol{\theta}_2)$ $j, k = 1, \cdots, n_i$ in each entry. According to the physics of the process and observed data, often $\sigma^2 > \tau^2$ and $\boldsymbol{\theta}_2 > \boldsymbol{\theta}_1$, where the inequality between the two vectors is interpreted using an element-wise comparison. This is because the standard profile often has a larger variation but smoother transitions compared with the deviation profile due to process variations. Consequently, we can define $\tau^2 = \rho \times \sigma^2$, with $0 \le \rho \le 1$.

Using these notations, the log-likelihood with respect to μ, σ^2 and ρ can be expressed as

$$l_r = -M_0 \ln\sigma - \frac{1}{2}\ln\det(\mathbf{S}+\rho\mathbf{V}) - \frac{(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})^T(\mathbf{S}+\rho\mathbf{V})^{-1}(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})}{2\sigma^2}. \tag{A1}$$

Taking the partial derivative of l_r and setting the gradient to zero results in

$$\begin{aligned}
\mu &= \frac{\mathbf{1}_{M_0}^T(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{Y}_{IC}}{\mathbf{1}_{M_0}^T(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{1}_{M_0}},\\
\sigma^2 &= \frac{(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})^T(\mathbf{S}+\rho\mathbf{V})^{-1}(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})}{M_0},\\
\sigma^2\mathrm{Tr}[(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{V}] &= \mathrm{Tr}[(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{V}(\mathbf{S}+\rho\mathbf{V})^{-1}\\
&\quad \times(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})^T],
\end{aligned} \tag{A2}$$

where $\mathrm{Tr}(\mathbf{S})$ denotes the trace of the matrix $\mathbf{S}$. The first two expressions in Equation (A2) are self-explanatory, and the third one can be transformed to

$$\begin{aligned}
&\frac{(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})^T(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{V}(\mathbf{S}+\rho\mathbf{V})^{-1}(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})}{(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})^T(\mathbf{S}+\rho\mathbf{V})^{-1}(\mathbf{Y}_{IC}-\mu\mathbf{1}_{M_0})}\\
&= \frac{\mathrm{Tr}[(\mathbf{S}+\rho\mathbf{V})^{-1}\mathbf{V}]}{M_0}
\end{aligned} \tag{A3}$$

by plugging in the expression for σ^2. When M_0 is large, the inverse of $\mathbf{S}+\rho\mathbf{V}$ may still take some time for each different ρ. We can significantly shorten the computational time by noting that $\mathbf{S}+\rho\mathbf{V} = \mathbf{V}^{1/2}(\mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2}+\rho\mathbf{I}_{M_0})\mathbf{V}^{1/2}$, where $\mathbf{I}_{M_0}$ is the identity matrix of dimension $M_0 \times M_0$, and $\mathbf{V} = \mathbf{V}^{1/2} \times \mathbf{V}^{1/2}$. Taking the singular value decomposition $\mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$, we have $(\mathbf{S}+\rho\mathbf{V})^{-1} = \mathbf{V}^{-1/2}\mathbf{P}(\boldsymbol{\Lambda}+\rho\mathbf{I}_{M_0})^{-1}\mathbf{P}^T\mathbf{V}^{-1/2}$. As a result, ρ only appears in the diagonal matrix $(\boldsymbol{\Lambda}+\rho\mathbf{I}_{M_0})^{-1}$, and all of the computationally intensive operations such as Cholesky decomposition, singular value decomposition, and most matrix multiplications only need to be calculated once for different ρ.

Using this computationally efficient procedure, we can find the solution to Equation (A3) in the interval [0, 1]. If no solution exists in this interval, one of the end points $\rho = 0, 1$ with largest likelihood value will be selected. Denoting $\bar{\rho}$ as the value selected that maximizes l_r, we can obtain $\bar{\mu}$, $\bar{\sigma}^2$ based on the first two expressions in Equation (A2). Then the maximum profile log-likelihood becomes (up to a constant)

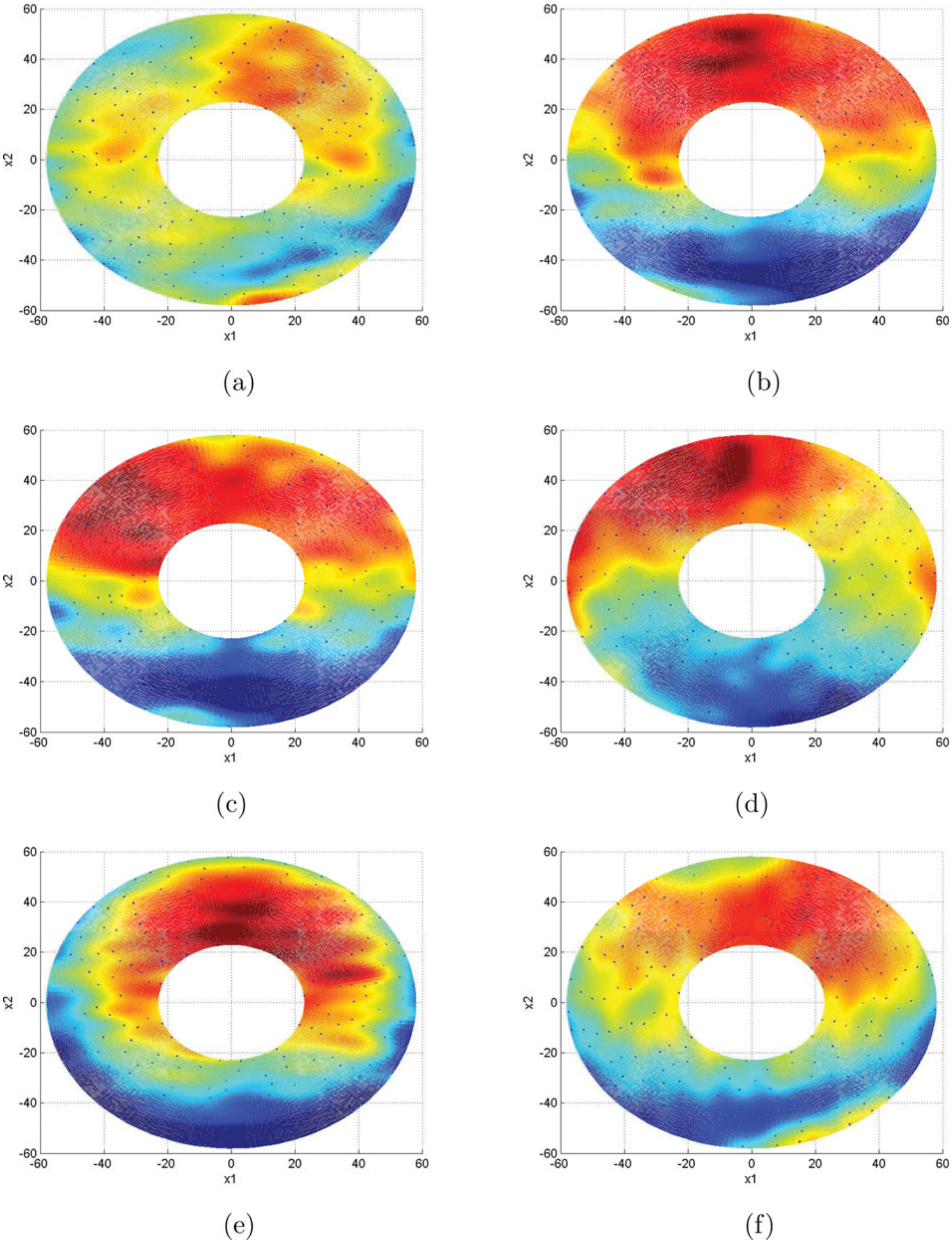
$$\bar{l}_r(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -M_0\ln\bar{\sigma} - \frac{1}{2}\ln\det(\mathbf{S}+\bar{\rho}\mathbf{V}), \tag{A4}$$

where all of the quantities depend on $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ explicitly or implicitly. As a result, the MLE estimator can be found by

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg\max_{\theta_1, \boldsymbol{\theta}_2}\left\{-M_0\ln\bar{\sigma} - \frac{1}{2}\ln\det(\mathbf{S}+\bar{\rho}\mathbf{V})\right\}. \tag{A5}$$

This optimization problem is much easier because the variables have similar scales, and the dimension is reduced. Thus, $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\rho}$ can be calculated using Equation (A2) with $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ plugged in.

B. Additional figures



(a)

(b)

(c)

(d)

(e)

(f)

Figure B1. Thickness profiles of the wafers that failed the tests: (a) wafer no. 12; (b) wafer no. 23; (c) wafer no. 24; (d) wafer no. 26; (e) wafer no. 28; and (f) wafer no. 30.