Big Data? Statistical Process Control Can Help!

Peihua Qiu
Department of Biostatistics, University of Florida
2004 Mowry Road, Gainesville, FL 32610

Abstract

"Big data" is a buzzword these days due to an enormous amount of data-rich applications in different industries and research projects. In practice, big data often take the form of data streams in the sense that new batches of data keep being collected over time. One fundamental research problem when analyzing big data in a given application is to monitor the underlying sequential process of the observed data to see whether it is longitudinally stable, or how its distribution changes over time. To monitor a sequential process, one major statistical tool is the statistical process control (SPC) charts, which have been developed and used mainly for monitoring production lines in the manufacturing industries during the past several decades. With many new and versatile SPC methods developed in the recent research, it is our belief that SPC can become a powerful tool for handling many big data applications that are beyond the production line monitoring. In this paper, we introduce some recent SPC methods, and discuss their potential to solve some big data problems. Certain challenges in the interface between the current SPC research and some big data applications are also discussed.

Key Words: Correlation; Covariates; Data-rich applications; Dynamic processes; Feature extraction; Image data; Nonparametric methods; Spatio-temporal data.

1 Introduction

Sensors, mobile devices, satellites, and many other modern technologies make data acquisition more and more convenient. Consequently, data-rich applications are common in practice. Computer scientists, informaticians, statisticians, and other data scientists have developed many tools for handling such applications (cf., e.g., Maheshwari 2019, Siegel 2016). In this paper, we introduce some recent methods in the research area of statistical process control (SPC), which should be fundamentally important to many big data applications but have not got much attention from the big data communities yet. Related discussions on similar topics can be found in Megahed and Jones-Farmer (2015) and Reis and Gins (2017).

Let us first discuss several data-rich applications. The Landsat project of the US Geological Survey (USGS) and NASA has launched 8 satellites since 1972 to continuously provide scientifically valuable images of the Earth's surface. These images can be freely accessed by researchers around the world (cf., Zanter 2016). The 47-year archive of the Landsat images has become a major data resource for scientific research about the Earth's surface in different scientific disciplines and research areas, including the land use research, forest science, climate science, agriculture forecasting, ecological and ecosystem monitoring, fire science, water resources, biodiversity conservation, and more. As a demonstration, the top-left panel of Figure 1 shows two images of the Las Vegas area in Nevada taken in 1984 and 2007, respectively. These two images clearly show the increasing urban sprawl in the Las Vegas area during the 23-year time period. Consequently, the environment in that region has changed quite dramatically in that time period. For instance, Lake Mead on the border of Nevada and Arizona has shrunk, due mainly to the increasing demand for water resource. The two images in the top-right panel of Figure 1 show an area of the tropical dry forest lying northeast of Santa Cruz de la Sierra of Bolivia in 1986 and 2000, respectively. The deforestation in that area during the 15-year time period is clearly seen in the images. The current satellite of the Landsat project (i.e., the Landsat 8) can deliver a new image of a given region roughly every 16 days. It is therefore fundamentally important to sequentially monitor the image data stream of the given region, and give a signal for further scientific research each time that a significant difference is detected between the current image and the images taken in the past. In the manufacturing industry, images have been used widely for quality control purposes, partly because they are convenient and economic to acquire. See plots in the second row of Figure 1 for an example. Applications of image monitoring in industry include stress and strain analysis of products (Patterson and Wang 1991), defect inspection of rolling processes (Jin et al. 2004), inspection of composite material fabrication (Sohn et al. 2004), and more. Many data-rich applications involve data from different sources. As an example, the Landsat image data mentioned above are often used together with meteorological data (Hausner et al. 2018), Moderate Resolution Imaging Spectroradiometer (MODIS) data (e.g., Weng et al. 2014), and data from other sources. In medical studies, data from different sources, including administrative claim records, clinical registries, electronic health records, biometric data, patient-reported data, medical imaging data, biomarker data, and more, are often used together for developing effective new medical treatments (e.g., Lee and Yoon 2017). In many cases, data from different sources would have different modes, formats, scales, or even quality. It is therefore challenging to integrate them and analyze the integrated data properly.

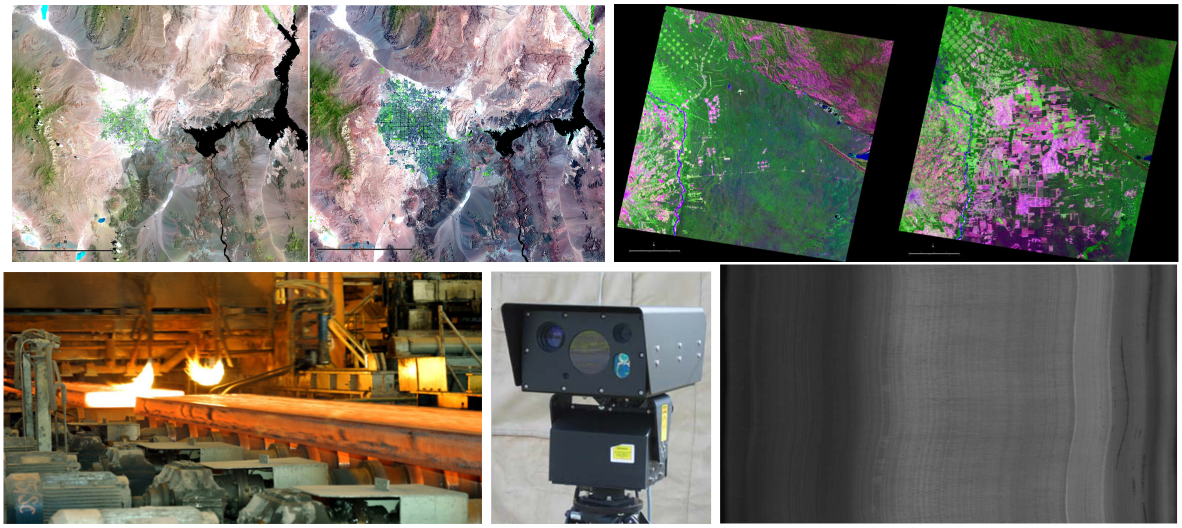

Figure 1: Two Landsat images of the Las Vegas area (top-left) taken in 1984 and 2007, two images of a forest area in Bolivia (top-right) taken in 1986 and 2000, an imaging system for detecting steel surface defects (low-left), the camera used by the system (low-middle), and an observed steel surface image (low-right).

One common feature of many big data applications is that observations of one or more longitudinal processes are collected sequentially over time and take the form of data streams in the sense that new batches of data keep coming over time. For such applications, one fundamental question to answer is whether the underlying longitudinal processes are temporally stable, or how their distributions change over time. As an example, by comparing satellite images of a specific region taken at different times, scientists can study the temporal change of important Earth resources and Earth environment, such as the land cover (e.g., Vittek et al. 2014), water resource (e.g., Frazier and Page 2000), land surface temperature (e.g., Parastatidis et al. 2017), and many more.

To sequentially monitor a longitudinal process, a major statistical tool is SPC (Hawkins and Olwell 1998, Montgomery 2012, Qiu 2014). Traditional SPC concepts and methods are developed mainly for monitoring production lines in the manufacturing industry, to detect any *special cause variation* (e.g., process mean shift or drift) in the observed data. In cases when the data variation is mainly due to random noise, it is often called *common cause variation*, and the process under monitoring is considered to be in statistical control, or simply *in-control (IC)*. When a process has a special cause variation present, the process is considered to be *out-of-control (OC)*. An SPC chart will give a signal and claim the process to be OC once special cause variation is detected. After the signal, the process needs to be stopped immediately for us to figure out the root causes of the detected special cause variation, and the process needs to be adjusted accordingly before it is re-started. When a production process is first started, we do not know much about its performance yet. In such cases, we usually let the process produce a small number of products and then analyze the observed data to see whether they meet the designed requirements. If the answer is "no", then the process should be adjusted accordingly. This adjust-and-control step is usually repeated several times until the quality of the production process meets the designed requirements. This phase of SPC is often called *Phase I SPC*. After Phase I SPC, we let the production process keep produce products, monitor its quality at the same time by periodically sampling the products and analyzing the observed quality variables of the sampled products using a control chart, and give a signal once the distribution of the observed data is found to be different from the IC distribution of the quality variables that has been estimated during Phase I SPC. This phase of SPC is often called *Phase II SPC* or *online* process monitoring. For a nice discussion about these concepts and terminologies, see Woodall (2000). Traditional SPC charts can be roughly divided into four categories: Shewhart charts (Shewhart 1931), cumulative sum (CUSUM) charts (Page 1954), exponentially weighted moving average (EWMA) charts (Roberts 1959), and change-point detection (CPD) charts (Hawkins et al. 2003). Some of them are designed for monitoring multiple quality variables (e.g., Hawkins 1991, Lowry et al. 1992), or quality variables taking binary, count or categorical values (e.g., Gan 1993, Megahed et al. 2011). To use these control charts, there are many issues that need to be properly addressed regarding their design and implementation. For instance, a Phase II control chart often requires certain IC process parameters to be estimated beforehand from an IC dataset collected during or immediately after Phase I SPC. The impact of the IC data size on the performance of the Phase II control chart has been discussed extensively in the literature. See, for instance, Jeske (2016), Zwetsloot and Woodall (2017) and the references cited therein.

Traditional SPC charts mentioned above are based on the assumptions that process observations are independent and identically distributed (i.i.d.) with a common parametric distribution (e.g., normal) when the process in concern is IC. While SPC has found more and more applications, these assumptions can hardly be satisfied in many of these applications, especially those involving big data. To meet the need of such applications, many new SPC charts have been developed in recent decades, which work well in cases when one or more of these assumptions are violated. Some new SPC methods have been developed specifically for data-rich applications. In this paper, we introduce some recent SPC concepts and methods that are flexibly designed and some others that are proposed for big data applications. Some challenges in the interface of existing SPC methods and certain big data applications will also be discussed.

The remainder of the article is organized as follows. Some recent SPC charts that are flexibly designed are discussed in Section 2. Several recent SPC charts that are developed specifically for big data applications are described in Section 3. Some challenges to handle big data applications using SPC methods are discussed in Section 4. Finally, several remarks conclude the article in Section 5.

2 Recent SPC Research for Monitoring Processes with Complicated Data

The four types of basic SPC charts discussed in Section 1 are based on the assumptions that IC process observations at different time points are i.i.d. and follow a parametric distribution (e.g., normal). In practice, these assumptions would hardly be valid, especially in data-rich applications. In the literature, it has been well demonstrated that the basic SPC charts are unreliable to use when one or more of their model assumptions are violated (Hawkins and Olwell 1998, Qiu 2014). So, much recent SPC research focuses on developing new control charts that are appropriate to use without these assumptions. In this section, we introduce some representative recent SPC methodologies that are developed to handle cases when one or more conventional assumptions are invalid. For simplicity of presentation, our introduction focuses mainly on Phase II SPC. For recent research on Phase I SPC, see papers such as Capizzi and Masarotto (2013), Graham et al. (2010), Jones-Farmer et al. (2009, 2014), and Ning et al. (2015).

2.1 Serially correlated data monitoring

In practice, process observations at different time points are usually correlated with each other (Apley and Tsung 2002). In the literature, there has been some discussion about process monitoring of serially correlated data. Many existing methods are based on parametric time series modelling of the observed data, and on sequential monitoring of the residuals obtained from the time series modelling. For instance, it is assumed in Capizzi and Masarotto (2008) that IC process observations $\{X_n, n \geq 1\}$ follow the ARMA(p,q) model

$$X_n = \mu_0 + \sum_{j=1}^{p} \phi_j (X_{n-j} - \mu_0) + \varepsilon_n - \sum_{j=0}^{q} \theta_j \varepsilon_{n-j}, \qquad (1)$$

where μ_0 is the IC mean, $\{\varepsilon_n\}$ are i.i.d. random errors with the distribution $N(0, \sigma^2)$, and $\{\phi_j\}$ and $\{\theta_j\}$ are coefficients. Then, after the ARMA(p,q) model (1) is estimated from an IC dataset, residuals of the Phase II process observations can be computed. Capizzi and Masarotto then suggested a CPD chart for monitoring the residuals. For related discussions, see Apley and Shi (1999), Apley and Tsung (2002), Berthouex et al. (1978), Lee and Apley (2011), Loredo et al. (2002), Montgomery and Mastrangelo (1991), Runger and Willemain (1995), Vander Wiel (1996), Wardell et al. (1994), and more.

Recently, Qiu et al. (2019) suggested a more flexible approach, described briefly below. Instead of using a parametric time series model (e.g., (1)), it is assumed that $\gamma(q) = Cov(X_i, X_{i+q})$ depends only on q when i changes, and $\gamma(q) = 0$ when $q > T_{max}$, where X_i and X_{i+q} are two process observations obtained at times i and $i + q$ when the process is IC, and T_{max} is an integer. The first assumption says that the covariance among process observations is stationary, which should be reasonable in most applications and is true when the ARMA(p,q) model (1) is valid. The second assumption says that serial data correlation exists only when two observations are within $T_{max} > 0$ in their observation indices, which should be (approximately) true in many applications, as long as T_{max} is not chosen too small. Then, the covariance structure described by $\gamma(q)$ can be estimated from an IC dataset, the Phase II process observations can be de-correlated using the estimated covariance structure, and a CUSUM chart can be applied to the de-correlated data for process monitoring. Li and Qiu (2019) suggested a self-starting version of this approach, where estimates of the IC parameters could be updated recursively during Phase II process monitoring. In this modified version, the charting statistic is based on data categorization, and thus it is more robust to the IC process distribution. An alternative nonparametric CUSUM chart based on wavelet transformations was proposed recently in Li et al. (2019a) for monitoring autocorrelated processes. Monitoring of autocorrelated count data was discussed in Fu and Jeske (2014), Xu and Jeske (2017), and Weib and Testik (2009).

2.2 Dynamic process monitoring

For many longitudinal processes, their distributions could change over time, even when their performance is considered IC. One example is about our medical indices, such as blood pressure readings and cholesterol levels. Distributions of these indices would change when people get older. To monitor such dynamic processes, the traditional SPC charts are obviously inappropriate to use because they require the IC process distribution to be unchanged over time, and they would give false signals when the (cumulative) difference between the IC process observations and the estimated IC distribution exceeds their control limits. In the past several years, we have developed a new method, called dynamic screening system (DySS), for sequential monitoring of dynamic processes (e.g., Li and Qiu 2016, 2017, Qiu and Xiang 2014, 2015, Qiu et al. 2018, Qiu et al. 2019, You and Qiu 2019, 2020), which is briefly described below.

The DySS method will be introduced using the example to early detect stroke by sequentially monitoring a person's total cholesterol level readings. It consists of the following three main steps:

(i) **Estimation of the regular longitudinal pattern:** We first estimate the IC longitudinal pattern of the total cholesterol level from the observed total cholesterol level data of a set of non-stroke people.

(ii) **Cross-sectional comparison:** For a specific person to monitor, we standardize her/his total cholesterol level observations using the estimated regular longitudinal pattern obtained in step (i).

(iii) **Sequential monitoring:** Apply a conventional control chart to the standardized data obtained in step (ii) for sequential process monitoring.

Each of these three steps will be briefly described below using a dataset from the SHARe Framingham Heart Study (Qiu and Xiang 2014). In the dataset, there are 1,028 non-stroke patients and 27 stroke patients. Each patient is observed at seven follow-up times at which observations of the total cholesterol level are collected. So, the observed data of the 1,028 non-stroke patients can be used as an IC dataset for estimating the regular longitudinal pattern of the total cholesterol level y, and they are assumed to follow the nonparametric longitudinal model

$$y(t_{ij}) = \mu(t_{ij}) + \varepsilon(t_{ij}), \qquad \text{for } i = 1, 2, \ldots, 1,028,\ j = 1, 2, \ldots, 7, \tag{2}$$

where t_{ij} is the jth observation time of the ith non-stroke patient, $\mu(t_{ij})$ is the mean of $y(t_{ij})$, and $\varepsilon(t_{ij})$ is the error term. For simplicity, observation times are re-scaled to be in the design interval $[0, 1]$. For any $s, t \in [0, 1]$, the covariance function of $\varepsilon(\cdot)$ is denoted as $V(s, t) = \text{Cov}(\varepsilon(s), \varepsilon(t))$. In model (2), observations of different people are assumed to be independent. Then, by the four-step model estimation procedure described in Qiu and Xiang (2014), the estimates of $\mu(t)$ and $\sigma_{y(t)}$ can be obtained, denoted as $\widehat{\mu}(t)$ and $\widehat{\sigma}_{y(t)}$, respectively. The pointwise 95% confidence band of $\mu(t)$, defined as $\widehat{\mu}(t) \pm 1.96\widehat{\sigma}_{y(t)}$, is shown in Figure 2 by the bold-dashed curves, along with the estimated mean $\widehat{\mu}(t)$ (bold-solid curve) and the observed data of the 27 stroke patients (thin curves). For a specific patient to monitor, assume that his/her observations of y are $y(t_j^*)$, for $t_j^* \in [0, 1]$ and $j \geq 1$. Then, the standardized observations are defined as

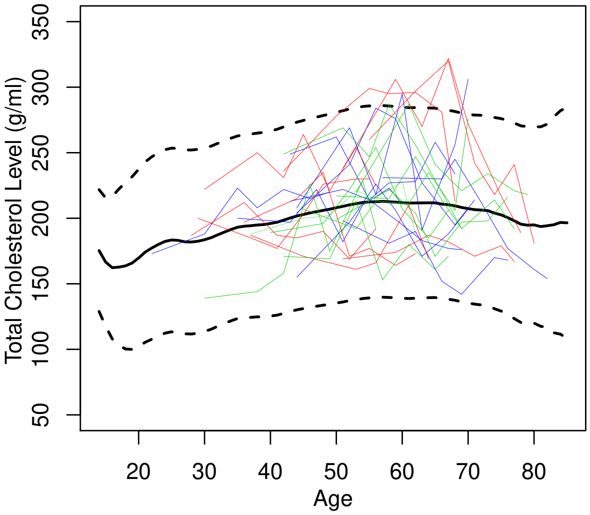

Figure 2: The 95% pointwise confidence band of the mean total cholesterol level $\mu(t)$ (bold-dashed curves), the point estimator $\widehat{\mu}(t)$ of $\mu(t)$ (bold-solid curve), and the observed total cholesterol levels of the 27 stroke patients (thin curves).

$$\widehat{\epsilon}(t_j^*) = \frac{y(t_j^*) - \widehat{\mu}(t_j^*)}{\widehat{\sigma}_{y(t_j^*)}}, \qquad \text{for } j \geq 1.$$

By using the standardized observations, the observed data of the specific patient have actually been compared to the observed data of the non-stroke patients cross-sectionally at different observation times $\{t_j^*\}$. In cases when the observations $\{y(t_j^*)\}$ are independent at different time points and their distributions are normal, we can define the upward CUSUM charting statistic to be

$$C_j^+ = \max\left[0, C_{j-1}^+ + \widehat{\epsilon}(t_j^*) - k_C^+\right], \qquad \text{for } j \geq 1,$$

where $C_0^+ = 0$ and $k_C^+ > 0$ is a constant. The upward CUSUM charting statistic C_j^+ is used here because the upward mean shifts are our main concern in this example about the total cholesterol level. Then, a signal is given by the chart when C_j^+ is larger than a properly chosen control limit. In the setting considered in Qiu and Xiang (2014), the control charts for monitoring the 27 stroke patients are shown in Figure 3, from which 22 patients get signals from the charts. The cases when the observations $\{y(t_j^*)\}$ are serially correlated, multivariate, and/or their distributions are non-Gaussian have been discussed in the literature. See, for instance, Li and Qiu (2016, 2017), Qiu et al. (2018) and the references cited therein.

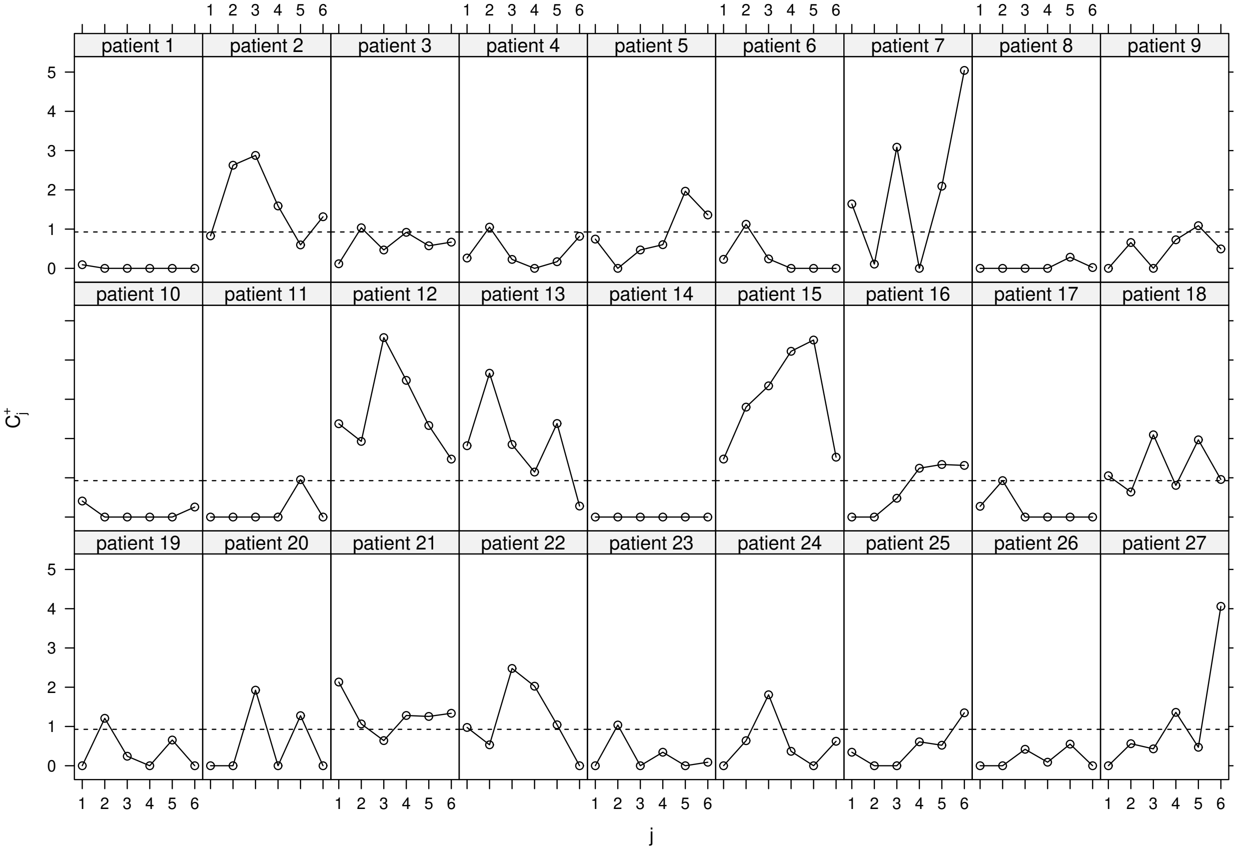

Figure 3: CUSUM charts for monitoring the 27 stroke patients in the setup considered in Qiu and Xiang (2014). The dashed horizontal lines denote the control limit.

2.3 Nonparametric SPC

Classic control charts, such as those introduced in Section 1, are based on the assumption that process observations follow a normal or another parametric distribution, which is rarely valid in practice. It has been well demonstrated that the classic control charts are unreliable to use when their distributional assumption is violated (e.g., Qiu and Hawkins 2001). So, nonparametric SPC has been in rapid development in recent years, and many nonparametric control charts have been developed. For recent overviews on this topic, see Chakraborti and Graham (2019) and Qiu (2018). In this part, some representative nonparametric control charts will be introduced.

Let us first discuss the univariate case. Many existing univariate nonparametric control charts are based on the ranking information within different batches of process observations. Assume that process observations at the nth time point are $X_{n1}, X_{n2}, \ldots, X_{nm}$, for $n \geq 1$, where m is the batch size. For the nth batch of observations, let R_{nj} be the rank of $|X_{nj} - \widetilde{\mu}_0|$ in the sequence $\{|X_{n1} - \widetilde{\mu}_0|, |X_{n2} - \widetilde{\mu}_0|, \ldots, |X_{nm} - \widetilde{\mu}_0|\}$, where $\widetilde{\mu}_0$ is the IC median of the process distribution. Then, the sum of the Wilcoxon signed-ranks is

$$\psi_n = \sum_{j=1}^{m} \operatorname{sign}\left(X_{nj} - \widetilde{\mu}_0\right) R_{nj}, \tag{3}$$

where sign(u) is the sign function defined to be -1, 0, 1, respectively, when $u < 0, = 0, > 0$. Intuitively, a mean shift in the process observations would be reflected in the value of ψ_n. For instance, an upward mean shift would make the value of ψ_n positively large, and a downward shift would make it negatively large. Thus, the classic control charts discussed in Section 1 can be applied to $\{\psi_n\}$ to make the resulting charts robust to the IC process distribution. Many nonparametric control charts have been developed in this way. See, for instance, Chakraborti et al. (2015), Graham et al. (2011), Hawkins and Deng (2010), Li et al. (2010), and Zou and Tsung (2010). An alternative nonparametric CUSUM chart based on kernel density estimation was developed by Ambartsoumian and Jeske (2015).

In multivariate cases, process observations $\{\mathbf{Z}_n = (Z_{n1}, Z_{n2}, \ldots, Z_{np})', n \geq 1\}$ are vectors, where $p > 1$ is the number of dimensions and $Z_{n1}, Z_{n2}, \ldots, Z_{np}$ are observations of p quality variables. Without loss of generality, assume that these observations have been standardized to have mean $\mathbf{0}$ and variance $I_{p \times p}$. In practice, the IC mean and variance could be estimated from an IC dataset for data standardization. In this setting, Qiu and Hawkins (2001) suggested a multivariate nonparametric SPC chart, based on *cross-component* ranking of the observed data. To be more specific, for the observed vector $\mathbf{Z}_n$, the first anti-rank A_{n1} is defined to be the index of the smallest component of $\mathbf{Z}_n$ that takes its values in $(1, 2, \ldots, p)$, the last anti-rank A_{np} is the index of the largest component, and so forth. Unlike the p ranks of the components of $\mathbf{Z}_n$ that are equally important in detecting mean shifts in $\mathbf{Z}_n$, the first anti-rank is particularly sensitive to downward mean shifts in one or a small number of components of $\mathbf{Z}_n$, and the last anti-rank is particularly sensitive to upward mean shifts in one or a small number of components of $\mathbf{Z}_n$. In cases when no prior information is available about the direction of a future shift, the first and last anti-ranks can be used jointly. Then, a nonparametric CUSUM chart was proposed in Qiu and Hawkins (2001) based on the anti-ranks, which was robust to the original process distribution. Qiu (2018) pointed out that "In cases when the joint distribution of $\mathbf{Z}_n$ is not normal, its marginal distributions and the relationship between any two subsets of the components of $\mathbf{Z}_n$ could be complicated, which explains the main reason why multivariate non-Gaussian distributions are difficult to describe." To overcome this difficulty in monitoring multivariate non-Gaussian data, Qiu (2008) proposed a general scheme to construct multivariate nonparametric SPC charts, by first categorizing the original data $\mathbf{Z}_n$ and then describing the joint distribution of the categorized data using a log-linear model (Agresti 2002). A nonparametric control chart can then be constructed based on the difference between the categorized Phase II observed data and the IC distribution of the categorized data that can be estimated from an IC dataset. A univariate version of this approach was discussed in Qiu and Li (2011). It has been shown in that paper by several large simulation studies that this approach performs favorably, compared to some representative nonparametric charts based on data ranking. Besides the multivariate nonparametric SPC charts described above, alternative charts based on data depth, spatial sign, and spatial rank can be found in Holland and Hawkins (2014), Li et al. (2013), Liu (1995), Zou and Tsung (2011), and Zou et al. (2012).

3 Recent SPC Research for Handling Data-Rich Applications

In recent years, SPC has found many data-rich applications, including those involving a large number of quality variables, processes with image or network data, and more. To properly handle such applications, many new SPC methods have been developed in the literature. Some representative ones are described below.

3.1 Sequential monitoring of high-dimensional multivariate processes

Semiconductor manufacturing is a complicated process with a series of steps, which can be roughly divided into the following four groups: blank wafer creation, diffusion and deposition, photolithography, and etching and metallization (May and Spanos 2016). See Figure 4 for a demonstration. At each step, observations of many quality variables are collected and monitored, regarding the physical or electronic properties of a product or its components, such as film thickness, film uniformity, and electronic resistance. As an example, in a real semiconductor manufacturing dataset maintained by the University of California at Irvine (UCI) Machine Learning Repository (http://archive.ics.uci.edu), observations of a total of 590 quality variables are collected. Production of many other products, especially durable products (e.g., cars, airplanes), all consists of many steps, with a large number of quality variables being monitored. For such high-dimensional multivariate processes, it is important to develop effective and efficient SPC charts for quality control purposes.

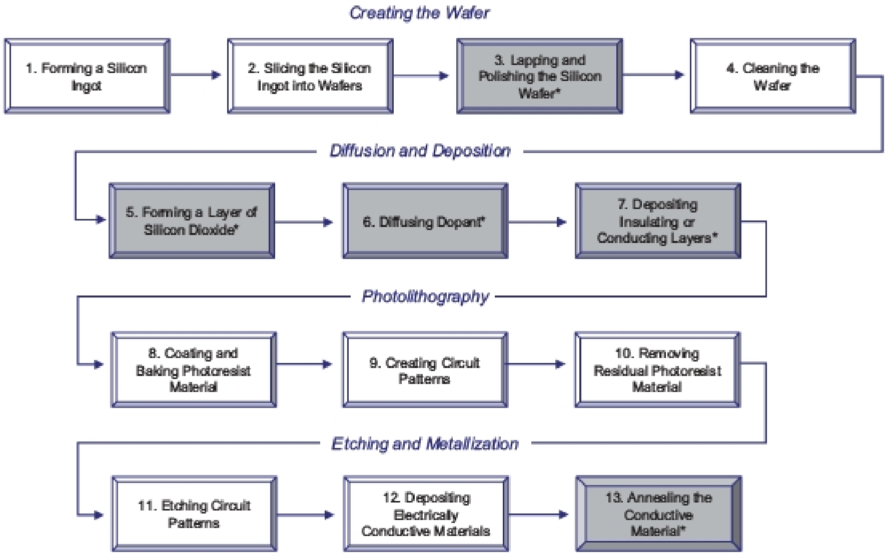
Creating the Wafer
1. Forming a Silicon Ingot
2. Slicing the Silicon Ingot into Wafers
3. Lapping and Polishing the Silicon Wafer*
4. Cleaning the Wafer
Diffusion and Deposition
5. Forming a Layer of Silicon Dioxide*
6. Diffusing Dopant*
7. Depositing Insulating or Conducting Layers*
Photolithography
8. Coating and Baking Photoresist Material
9. Creating Circuit Patterns
10. Removing Residual Photoresist Material
Etching and Metallization
11. Etching Circuit Patterns
12. Depositing Electrically Conductive Materials
13. Annealing the Conductive Material*

Figure 4: Demonstration of a semiconductor manufacturing process.

Early multivariate SPC charts were designed mainly for monitoring multivariate processes with a relatively small number of quality variables (e.g., Hawkins 1991, Lowry et al. 1992). These charts would become ineffective when the number of quality variables is large, partly because all quality variables are treated equally and monitored jointly in these charts and thus they cannot react promptly to shifts in a small number of quality variables in such cases. Also, their computation when monitoring high-dimensional processes is so intensive that they become inappropriate for such applications. Almost at the same time, Zou and Qiu (2009) and Wang and Jiang (2009) proposed two multivariate control charts for monitoring high-dimensional processes. The major idea behind these control charts is that shifts in high-dimensional processes often occur in a small number of quality variables in practice. Thus, it could improve the effectiveness of a multivariate control chart if the quality variables that most probably have shifts at the current time point can first be identified by a variable selection procedure and then the control chart can just focus on the selected quality variables. Let $\{\mathbf{X}_n, n \geq 1\}$ be observations of a p-dimensional process, where p is large, and let $\overline{\mathbf{X}}_n$ be the sample mean of $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$. Then, Zou and Qiu (2009) considered the following adaptive LASSO penalized likelihood function:

$$PL(\boldsymbol{\delta}) = n\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}_0 - \boldsymbol{\delta}\right)' \Sigma_0^{-1}\left(\overline{\mathbf{X_n}} - \boldsymbol{\mu}_0 - \boldsymbol{\delta}\right) + n\gamma \sum_{j=1}^{p} \frac{1}{|\overline{X}_{nj} - \mu_{0,j}|}|\delta_j|,$$

where $\boldsymbol{\mu}_0$ and Σ_0 are the IC mean and the IC covariance matrix of the process, $\overline{X}_{nj}$, $\mu_{0,j}$ and δ_j are the j-th components of $\overline{\mathbf{X}}_n$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\delta}$, respectively, and $\gamma > 0$ is a smoothing parameter. Let the solution of the minimization procedure $\min_{\boldsymbol{\delta}} PL(\boldsymbol{\delta})$ be denoted as $\widehat{\boldsymbol{\delta}}_\gamma$. Then, by the oracle property of the adaptive LASSO procedure, if a given quality variable does not have a mean shift by the time point n, then the corresponding component of $\widehat{\boldsymbol{\delta}}_\gamma$ should be 0. Thus, process monitoring can just focus on the non-zero components of $\widehat{\boldsymbol{\delta}}_\gamma$. It has been shown in Zou and Qiu (2009) that the resulting control chart would be much more effective than some traditional multivariate control charts, especially in cases when p is large but the number of components having shifts is small. In the approach discussed in Wang and Jiang (2009), an L_0 penalty was recommended for the penalized likelihood function, while the penalty in $PL(\boldsymbol{\delta})$ defined above is L_1. Several modifications and generalizations have been proposed. See, for instance, Capizzi and Masarotto (2011), Yan et al. (2018), and Zou et al. (2015). One good property of such multivariate control charts based on variable selection is that the shifted components can be easily specified after the related chart gives a signal. Recent discussions on post-signal diagnostics when monitoring high-dimensional processes can be found in Li et al. (2019b) and the references cited therein. Process monitoring and post-signal diagnostics about the semiconductor manufacturing dataset mentioned above were discussed in Zou et al. (2015) and Li et al. (2019b).

Another type of control charts that are potentially useful for monitoring high-dimensional multivariate processes is based on machine learning approaches. A key component of a control chart is an appropriate decision rule to decide whether the process under monitoring is IC or not at each observation time point. So, the process monitoring problem could be regarded as a classification problem in which the process status should be classified into either the IC or the OC status after the process observations at a given time point are analyzed properly by a control chart. To develop a classification rule from a training dataset, usually it requires that the training dataset contains both IC and OC process observations. Namely, the training data can be denoted as $\{(\mathbf{X}_t, y_t), t = 1, 2, \ldots, m\}$, where $\mathbf{X}_t$ are observations of the p-dimensional quality vector $\mathbf{X}$ and

$y_t \in \{-1, 1\}$ are group labels with "$y_t = -1$" denoting the IC group and "$y_t = 1$" the OC group. In SPC applications, however, we usually have an IC dataset only, obtained before Phase II process monitoring. Thus, the above mentioned *labelled data* are often unavailable. To overcome this difficulty, Tuv and Runger (2003) proposed the idea of *artificial contrasts*. By this idea, some artificial observations of $\mathbf{X}$ can be generated from a given distribution (e.g., the distribution of each component of $\mathbf{X}$ is assumed Uniform in a range and all components of $\mathbf{X}$ are assumed to be independent). Then, these artificial observations can be used as OC observations. A classification rule can then be obtained from the training dataset that consists of the original IC data and the artificial OC data by a machine learning approach, such as support vector machines (SVM), random forests (RF), multiple additive regression trees (MART), and more. For generalizations and improvements of the original artificial contrasts idea, see Deng et al. (2012) and the references cited therein. To accommodate the fact that only IC data are available before Phase II process monitoring, Sun and Tsung (2003) suggested a multivariate control chart based on *one-class classification*. More specifically, a boundary curve (or surface) of the IC data can be first specified using the SVM algorithm for one-class classification. To increase the flexibility of the boundary curve (or surface), a kernel-distance is used when measuring the distance from a given data point to the center of the IC data. Then, a Phase II process observation is claimed to be OC if it is located outside the boundary curve (or surface). Some modifications of this approach can be found in He and Zhang (2011), Ning and Tsung (2013) and the references cited therein.

There are also some multivariate SPC charts based on the principal component analysis (PCA). See, e.g., Ferrer (2007), Jachson (1991), Kourti and MacGregor (1996), and the references cited therein. PCA is a popular statistical tool for reducing the dimensionality of an observed dataset (Johnson and Wichern 2007). The first principal component of the p-dimensional random vector $\mathbf{X}$ is defined to be the linear combination $Y_1 = \mathbf{a}_1'\mathbf{X}$ with the largest variance, where $\mathbf{a}_1$ is a coefficient vector with unit length, the second principal component is defined to be the linear combination $Y_2 = \mathbf{a}_2'\mathbf{X}$ with the largest variance, where $\mathbf{a}_2$ is a coefficient vector with unit length that is orthogonal to $\mathbf{a}_1$ (i.e., $\mathbf{a}_1'\mathbf{a}_2 = 0$), and so forth. In practice, the principal components can be obtained from the eigenvalue-eigenvector decomposition of the covariance matrix Σ of $\mathbf{X}$, where Σ can be estimated from an IC data. By using the first several principal components, majority variability in the original distribution of $\mathbf{X}$ can be preserved. Thus, the dimension can be reduced. For the observed p-dimensional process observations $\{\mathbf{X}_n, n \geq 1\}$, if q principal components are used, where

q is often much smaller than p, then we can simply monitor $\{\mathbf{Y}_n = (\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X} \ldots, \mathbf{a}_q'\mathbf{X})', n \geq 1\}$. However, some existing PCA-based multivariate SPC charts depend on the normality assumption in their chart designs. Also, process shifts in the directions that are orthogonal to the directions of the adopted principal components cannot be detected effectively by such control charts. In the literature, there are also some multiscale control charts developed for monitoring multivariate processes (e.g., Bakshi 1998, Ganesan et al. 2004, Guo et al. 2012, Reis and Saraiva 2006). These charts are usually based on the wavelet transformations (Donoho and Johnstone 1994) of the original process observations. The wavelet coefficients at different scales are used for constructing control charts for detecting process mean and/or variance shifts. In the construction of the control charts, PCA is often applied to the wavelet coefficients to reduce dimensionality (e.g., Reis and Saraiva 2006).

3.2 Sequential monitoring of univariate and multivariate profiles

All the control charts discussed in the previous parts are for monitoring observations of a single or multiple quality variables of a sequential process. In many applications, performance of a process is reflected by the relationship between a set of response variables and a set of predictors. Observations of the response variables versus the predictors are called *profiles* in the SPC literature. As an example, the deep reactive ion etching (DRIE) process is critical to the output wafer quality in semiconductor manufacturing and requires careful control and monitoring. In the DRIE process, the desired etching profile is the one with smooth and straight side walls and flat bottoms. Ideally, the side walls of a trench should be perpendicular to the bottom of the trench with a certain degree of smoothness around the corners, as shown by the plot in the middle of the top panel of Figure 5. Various other profile shapes, such as the positive and negative ones shown in the top panel of Figure 5 due to underetching and overetching, are considered to be unacceptable (Rauf et al. 2002). The lower-left panel of Figure 5 shows a multi-operation forging machine with progressive dies. In the forging processes, tonnage force exerted on all dies are measured by four strain sensors mounted at four corners. The lower-right panel of Figure 5 shows the profile data recorded by the four strain sensors during one forging process.

To monitor the DRIE process automatically, observed data of each etching profile can be acquired by the scanning electron microscope (SEM). Because the etching profiles are usually symmetric, we can focus on one half of each profile (e.g., the left half). To make that part of the

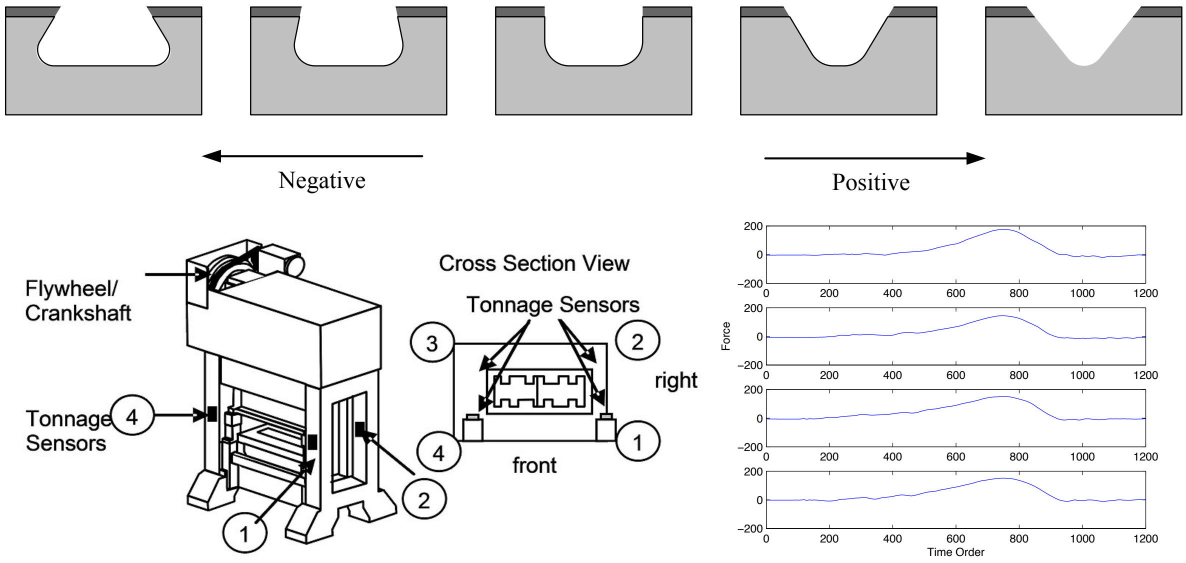

Figure 5: Several etching profiles of a deep reactive ion etching process (upper panel), a multi-operation forging machine (lower-left panel), and profile data recorded by the four strain sensors of the forging machine during one forging process (lower-right panel).

profile convenient to describe by a mathematical function, it can be rotated by 45 degrees along a given reference point in a pre-specified coordinate system, before profile readings are collected by SEM. Then, the observed profile data can be described by the following model:

$$y_{ij} = g(x_{ij}) + \varepsilon_{ij}, \text{ for } j = 1, 2, \ldots, n_i, \; i = 1, 2, \ldots, \tag{4}$$

where $g(x)$ is the mean profile function, $\{(x_{ij}, y_{ij}), j = 1, 2, \ldots, n_i\}$ are observations of the ith profile, and $\{\varepsilon_{ij}\}$ are random errors. When a profile is IC, its mean profile function is denoted as $g_0(x)$. This function and other IC quantities can usually be estimated from an IC dataset that consists of a number of observed IC profiles. Then, Phase II profile monitoring is mainly for detecting any systematic deviation of the observed profiles from the IC profile pattern that is described by $g_0(x)$ and other IC quantities. For instance, if our main concern is about the mean profile function, then our goal is to detect any deviation of the actual mean profile function of a future profile from the IC mean profile function $g_0(x)$. Early profile monitoring methods assume that $g(x) = a + bx$, where a and b are regression coefficients (e.g., Kang and Albin 2000, Kim et al. 2003, Zou et al. 2006). In such cases, a and b can be estimated from observations of individual profiles, and then the sequence of estimated values of each coefficient can be monitored by a conventional control chart. For an overview on this topic, see Woodall (2007). More recent research on profile monitoring does not require any parametric form for describing $g(x)$, and the within-profile data correlation can also be accommodated in some methods (e.g., Qiu et al. 2010, Zou et al. 2008). For a discussion about monitoring of real DRIE profile data by a nonparametric approach, see Zou et al. (2009). In the forging process monitoring problem shown by the two lower panels of Figure 5, four univariate profiles need to be monitored simultaneously. This is the multivariate profile monitoring problem discussed in the SPC literature. For Phase I and Phase II methods on multivariate profile monitoring, see Paynabar et al. (2016), Ren (2019) and the references cited therein. Some methods for monitoring profiles with multiple predictors will be discussed in the next subsection.

3.3 Sequential monitoring of spatial data

The top-left panel of Figure 6 shows a 3D printed product. For this product, we are mainly concerned about the shape of its top surface. To monitor its top surface, a laser scanner shown in the top-middle panel of Figure 6 can scan the top surface and record the positions of the points on the top surface. The recorded data of the top surface of one product are shown in the top-right panel of Figure 6. In such recorded data, the number of scanned points on one top surface is about 150,000 and their (x, y) positions in the coordinate system may deviate from regularly spaced rows and columns. The two lower panels of Figure 6 show the incidence rates of an infectious disease in Florida on 06/01/2012 (a summer time) and 12/01/2012 (a winter time). Because of the great threat of infectious diseases to the public health, some global, national or regional disease reporting and surveillance systems have been established to collect disease occurrence data (often on a daily basis), and the collected data are then monitored continuously in these systems for detecting disease outbreaks early. The observed data shown in Figure 6 can be regarded as spatial profile data, since the data at each time point describe the relationship between a response variable (i.e., the surface height and the disease incidence rate in the two examples) and the location variables x and y.

In the two applications shown in Figure 6, a sequence of spatial data needs to be monitored in order to monitor the production of the 3D printed products or the infectious disease incidence rates. This is the so-called spatio-temporal process monitoring problem in the literature, and its major goal is to check whether the spatial data have a significant distributional shift over time. Besides the two applications shown in Figure 6, spatio-temporal process monitoring has many other important applications, including environmental monitoring (e.g., temporal monitoring of the spatial PM 2.5 readings), weather surveillance, and many more. In many applications, it is

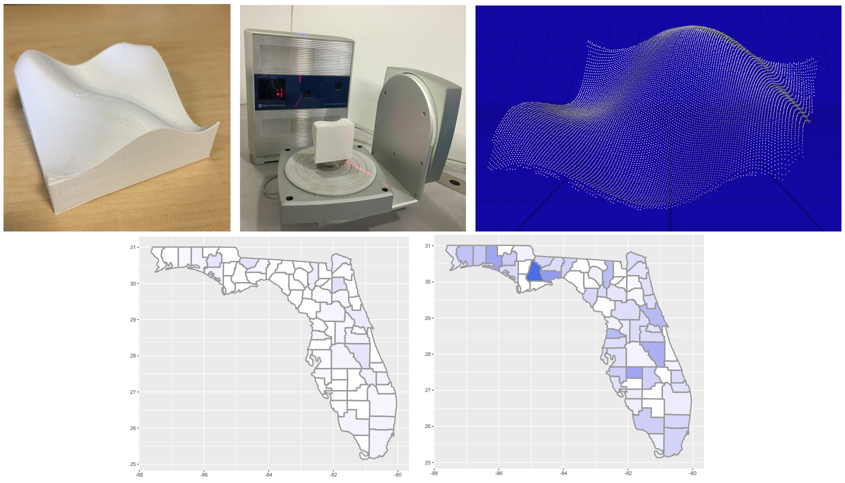

Figure 6: Top panels: A 3D printed product (left), a laser scanner (middle), and the observed data of the product's top surface. Bottom panels: Incidence rates of an infectious disease in Florida on 06/01/2012 (left) and 12/01/2012 (right).

often reasonable to assume that the true spatial mean function of an individual spatial profile is a continuous function. But, the spatio-temporal data are often spatio-temporally correlated, with observations closer in time and/or location are more correlated. Also, the distribution of the spatial data often changes over time even when the process is IC, due to seasonality and other reasons. These features make the spatio-temporal data challenging to analyze and monitor properly. In a national or regional disease reporting system, some conventional control charts, such as the CUSUM and EWMA charts, are usually included. However, the spatio-temporal data correlation and the time-varying nature of the IC longitudinal pattern of the spatial profiles are usually ignored by them (Chen et al. 2010, Kite-Powell et al. 2010). In the epidemiology literature, there are some existing methods for detecting spatial or spatio-temporal disease clusters based on the Knox or scan statistic (e.g., Knox and Bartlett 1964, Kulldorff 1997). These methods, however, are all retrospective, and they require certain restrictive distributional assumptions (e.g., the disease incidence follows a Poisson or Negative Binomial distribution). Some control charts for prospective disease surveillance have been developed based on the Knox or scan statistic (e.g., Marshall et al. 2007, Woodall 2006). But, the spatio-temporal data correlation and the time-varying nature of the IC longitudinal pattern of the spatial profiles are also ignored in these methods. There are some other discussions about prospective online monitoring of spatial data, motivated mainly by certain manufacturing applications where it is reasonable to assume that the IC distribution of the observed spatial data does not change over time. For instance, Jiang et al. (2011) suggested a likelihood-ratio-based control chart for monitoring spatial data, by assuming the spatial data are jointly normally distributed. Colosimo et al. (2014) and Wang et al. (2014) suggested control charts based on the Gaussian process modelling framework. Zang and Qiu (2018a,b) suggested control charts for both Phase I and Phase II monitoring of spatio-temporal data obtained from a 3D printing production process (cf., Figure 6) by using the nonparametric local linear kernel surface estimation procedure, where the spatio-temporal data correlation was ignored. For monitoring the disease incidence rates, Zhang et al. (2015, 2016) suggested monitoring the observed data in individual regions of different scales. The suggested methods in these papers can accommodate the time-varying nature of the IC process distribution. But, they do not monitor the observed disease incidence rates in different regions simultaneously. So far, we could not find any existing methods for monitoring spatio-temporal data that can properly accommodate both the spatio-temporal data correlation and the time-varying IC longitudinal pattern of the spatial profiles. Therefore, much future research is needed on this topic.

Due to a rapid progress in image acquisition techniques, images have become a basic data format in many applications. The images shown in the top panels of Figure 1 are taken by the satellite of the Landsat project (cf., Section 1 for a related discussion). By monitoring the sequence of such images of a given region, we can study longitudinal changes of Earth's surface and/or the environment in that region. Images can also be regarded as spatial profiles because they consist of many spatially located pixels at which the image intensity values denote the brightness. But, besides data correlation and time-varying IC longitudinal pattern of the spatial profiles mentioned above, images have some special data structures, including the two major ones described below. First, images usually have edges, and the corresponding image intensity surfaces would have jumps and other singularities involved. Second, the geometric locations in a sequence of images of a given object (e.g., a specific region in the satellite images shown in Figure 1) are often misaligned, due to the relative move between the camera and the object at different image acquisition times. Therefore, the observed images in the sequence need to be geometrically aligned before process monitoring. Otherwise, the process monitoring results could be unreliable. To this end, *image registration* methods for image geometric alignment become critically important (e.g., Qiu and Xing 2013). To demonstrate the importance of image registration, Figure 7 shows the satellite images of the San Francisco bay area taken in 1990 and 1999, respectively, their difference before image registration, and their difference after image registration. It can be seen that the pattern in the image of the difference between the two original images is mainly due to a geometric misalignment, and this pattern mostly disappears after the image registration.

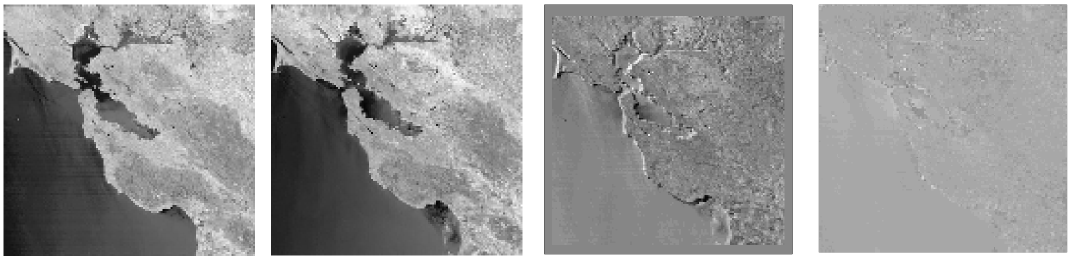

Figure 7: Two satellite images taken at the San Francisco bay area in 1990 and 1999, respectively, their difference before image registration, and their difference after image registration (from left to right).

It should be pointed out that image registration cannot solve all the problems caused by geometric misalignment. To make this point clear, let us consider a toy example shown in Figure 8. Plots (a) and (b) in the figure are two original images to compare. By checking the edge curves in the two images, it can be seen that their positions are different. For instance, the right end of the edge curve in plot (b) is higher than that in plot (a). After image registration, assume that the image in plot (b) becomes the one in plot (c). If we compare the images in plots (a) and (c) carefully, then we can see that the original difference between (a) and (b) has been mostly removed by image registration. But, the discrepancy is not completely removed yet, which can be seen from the pixelwise difference between the images in (a) and (c) that is shown in plot (d). It can be seen that the pixelwise differences around edge locations are large. In such cases, comparison based on such differences would likely lead to a false conclusion that, besides the geometric misalignment, the two original images are significantly different. This example shows that edge pixels and pixels in the continuity regions play quite different roles in image comparison. Thus, they should be considered separately. Based on these considerations, Feng and Qiu (2018) suggested several approaches for comparing two images of a same object, using either the observed image intensities around the detected edges, or the observed image intensities in the estimated continuity regions, or both. The methods were demonstrated using the steel surface example shown in Figure 1.

There is some existing research on sequential monitoring of image data, mainly in the chemical and industrial engineering literatures because images have been widely used in manufacturing industry in recent years (e.g., Megahed et al. 2011, Prats-Montalban 2014, Yan et al. 2015). Some existing methods for image monitoring proceed in two main steps. They first extract some features

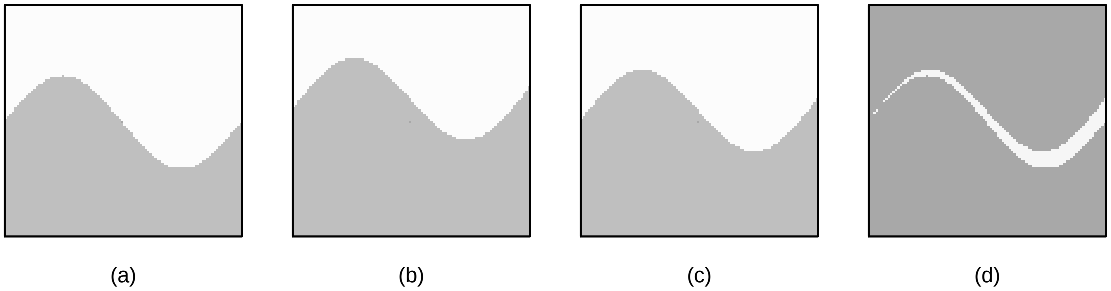

Figure 8: (a) First original image, (b) second original image, (c) second original image after image registration, and (d) pixelwise difference between (a) and (c).

from each observed image using methods such as the PCA, and then monitor the extracted features by a conventional control chart (e.g., Duchesne et al. 2012, Lin et al. 2008). Some other methods focus on certain pre-specified regions, called regions of interest (ROIs), in individual images, and then monitor the images by a control chart constructed based on a summary statistic (e.g., the average image intensity) of the ROIs (e.g., Jiang et al. 2011, Megahed et al. 2012). The first type of methods mentioned above have not taken into account the spatial data structure of the observed images, while the second type considers the spatial data structure partially by specifying ROIs in advance. All these methods do not take into account the edges and other complicated structure of the image intensity surfaces. They have not properly accommodated the possible geometric misalignment among different observed images and possible data correlation either. Therefore, much future research is needed for proper monitoring of image data.

3.4 Sequential monitoring of network data

The Enron email corpus is a well-known dataset in social network research. After the bankruptcy of the Enron Corporation in October 2001, all emails to and from the Enron employees during the period from 1998 and 2002 were made public by the ruling of the Federal Energy Regulatory Commission. A subset of this data is shown in Figure 9 as a network, which describes how all people in the dataset are connected by emails. For a network, people usually use the tool of graph theory to describe its status at a given time point or within a given time period, where discrete objects (i.e., employees in the Enron email example) are called *nodes* or vertices, and the pairwise relationship among nodes is described by *edges.* In the Enron email example, the name of an employee is attached to the corresponding node. Such nodes are called *labeled* nodes. The edges are denoted by *directed* lines to specify the email senders and recipients. The amount of emails between a pair of employees is denoted by the line thickness of an edge. Two employees are not connected by an edge if they have no email conversations during the specific time period of the data. In practice, the relationship among nodes in a network often changes over time, the related network process is called a *dynamic* network process. For a dynamic network process, it is often our interest to detect any distributional changes of the network data over time. In many cases, the temporal distributional changes are related to one or more sub-networks, which are also called *anomalies* in the literature (e.g., Savage et al. 2014).

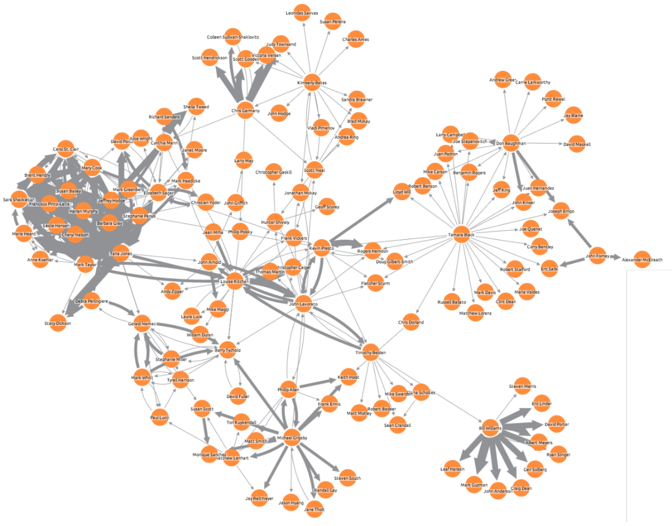

Figure 9: Network of a subset of the Enron email corpus.

There have been some existing SPC methods for monitoring a network process. These methods usually apply SPC charts to aggregated measures of the topological characteristics (e.g., density, degree, clustering coefficient, and scan statistic) of the entire network or relevant sub-networks (e.g., McCulloh and Carley 2011, Neil 2013). Some of them are based on parametric modelling while the others are nonparametric. For instance, Zou and Li (2017) suggested a network monitoring approach based on network state-space modelling. Dong et al. (2019) suggested a approach for monitoring multi-layer networks, using the so-called multi-layer weighted stochastic block modelling that was based on the assumption of a multivariate zero-inflated Poisson distribution for describing the correlated multi-layer interactions among knots. For other recent research on network modelling and monitoring, see Jeske et al. (2018), Savage et al. (2014), Woodall (2017), Yu et al. (2019), and the references cited therein. The Enron email corpus data shown in Figure 9 have been analyzed in many papers, including Dong et al. (2019).

4 Interface Between Big Data and SPC

When data get larger, they often have a more complicated structure. Consequently, proper analysis of them becomes more challenging. In this part, we discuss some challenging issues that are quite common to various big data monitoring problems, with the hope that they can be properly addressed in the future research.

Feature-based process monitoring. Because of the big data size in certain process monitoring problems, people tend to simplify the observed data, by first extracting certain features of the observed data and then monitoring the extracted features. One example is about image monitoring discussed in Section 3.3. Instead of monitoring a sequence of whole observed images, some authors suggest monitoring one or more types of image features. Commonly used image features include landmarks, edges, corners, regions, centroids, (functional) principal components, and more (Qiu and Nguyen 2008). By properly monitoring chosen features, the process monitoring problem and the involved computation are simplified. In such cases, a key question to answer is whether the original goals of process monitoring can still be achieved by using the simplified data. In some cases, the answer is unfortunately negative. As an example, principal components are commonly used in image monitoring. While these quantities ignore most spatial structures (e.g., edges) in the observed images, they are not sensitive to certain specific shifts as well. For instance, assume that the first principal component of a sequence of images has the expression

$$Y_{1n} = \mathbf{a}_1' \mathbf{X}_n, \qquad \text{for } n \geq 1,$$

where $\mathbf{X}_n$ is a long-vector of the observed image intensities of the nth image, and $\mathbf{a}_1$ is a coefficient vector. Then, by monitoring the sequence of the first principal component $\{Y_{1n}, n \geq 1\}$, it is impossible to detect all mean shifts $\boldsymbol{\delta}$ satisfying $\mathbf{a}_1' \boldsymbol{\delta} = 0$. Therefore, there are several important questions to answer before a feature-based big data monitoring method can be considered, which include: i) what kind(s) of features are appropriate to use for a specific big data monitoring problem, ii) how many features should be extracted for process monitoring, and iii) whether the original goals of process monitoring have been substantially compromised by using the selected features. These questions, however, have not been properly addressed yet in the literature. Instead of extracting features to simplify the original observed data and then monitor the extracted features, an alternative approach is to monitor the original process observations and use distributed parallel computing (cf., Chen et al. 2018) and other fast computing algorithms to handle the heavy computing burden.

More future research is needed in that direction.

Accommodation of complicated data structure. Big data often have complicated data structures. One such structure concerns data correlation. Usually, quality variables are correlated with each other, and their observations at different time points are also serially correlated. Such data correlation is often a reflection of the impact of certain confounding variables, such as weather and geographical conditions in the Landsat image example (cf., Figure 1). Because the impact of the confounding variables is often difficult to describe, the correlation in the observed data is difficult to describe as well. In some existing methods for monitoring images, networks, and other data streams with complicated structures, the observed data are assumed to be either independent or following some specific parametric models (e.g., ARMA models). These assumptions are rarely satisfied in practice, making the related methods unreliable or ineffective to use (Qiu et al. 2019). Process observations have many other complicated structures, especially when the data are big. For instance, images usually have edges and other spatial structures, and network graphs may have clusters involved. Proper accommodation of such data structures is also important to make the related process monitoring methods effective.

Accommodation of covariates. In practice, performance of a process is often affected by various covariates. Observations of some covariates could be available to us. Therefore, we should make use of the helpful information in covariates when monitoring the related process. As an example, the sequence of Landsat images of a given region may depend on the weather and geographical conditions of the region, and observations of these conditions can be obtained from the databases managed by the National Weather Service, National Centers for Environmental Information, US Census Bureau, and more. However, it has not been well discussed in the SPC literature yet regarding the proper use of the helpful information in covariates. Intuitively, the observed data of the process under monitoring and the covariates can be monitored jointly. But, this would not be a good plan because a signal from the joint monitoring scheme could be triggered by the covariates, and it is often inconvenient to distinguish this scenario from the one that the signal is actually triggered by the performance of the process itself. The alternative idea to first regress the quality variables of a process on the covariates and then monitor the resulting residuals would suffer the same limitation. Also, data from different sources could have different observation times, different data contamination types or levels, different data quality, and so forth, which would make it harder for us to use helpful information in covariates.

Dynamic process monitoring. As discussed in Section 2.2, many processes to monitor in practice are dynamic processes in the sense that their IC distributions would change over time. Although dynamic process monitoring by DySS provides a reasonable solution to this important process monitoring task, there are still some fundamentally important issues to address. First, we need to determine a time period from the history data of the process under monitoring, in which the performance of the process is believed to be IC. In some applications, there are existing scientific discussions about this issue. For instance, in the infectious disease surveillance literature, if all observed incidence rates of a given disease are below a specific level, then it can be concluded that there is no disease outbreak (e.g., Bie et al. 2010). Such scientific discussions should be helpful for us to determine an IC time period and/or an IC dataset for process monitoring. In some other applications, however, it is difficult to have this kind of scientific standards for defining IC performance of a process, or a standard does not even exist (e.g., Earth's surface and the environment keep changing over time in the Landsat image example). For such applications, probably we can simply specify a time period as a baseline time period and compare the future performance of a process with its performance in the baseline time period. When time goes by, the baseline time period can be re-specified as appropriate. In practice, many dynamic processes cannot be stopped once a signal is given by a control chart, although certain interventions can still be implemented (e.g., occurrence of an infectious disease). For such processes, what will be an appropriate strategy to detect the next process shift after a signal is obtained from a control chart? How can we evaluate the performance of the control chart in detecting multiple shifts? These issues need to be addressed properly in our future research.

5 Concluding Remarks

Many big data in practice are in the form of data streams, and they are sequential observations of certain underlying longitudinal processes. To study the patterns of these processes over time, SPC is a relevant statistical tool. In the previous sections, we have introduced some recent SPC methods for monitoring processes with complicated data and for handling some data-rich process monitoring problems. These SPC methods should be useful for many big data applications. However, the complicated data structure involved in certain big data applications raises many new challenging issues that the existing SPC methods cannot yet handle properly (cf., the related discussion in

Section 4). Therefore, much future research is needed to modify existing SPC charts or develop new SPC methods in order to address them adequately.

Acknowledgments: We thank the editor, the associate editor and two anonymous referees for many constructive comments and suggestions, which improved the quality of the paper greatly. This research is supported in part by an NSF grant.

References

Agresti, A. (2002), *Categorical Data Analysis (2nd edition),* New York: John Wiley & Sons.

Ambartsoumian, T., and Jeske, D.R. (2015), "Nonparametric CUSUM control charts and their use in two-stage SPC applications," *Journal of Quality Technology*, **47**, 264–277.

Apley D.W., and Shi, J. (1999), "The GLRT for statistical process control of autocorrelated processes," *IIE Transactions*, **31**, 1123–1134.

Apley D.W., and Tsung, F. (2002), "The autoregressive T^2 chart for monitoring univariate autocorrelated processes," *Journal of Quality Technology*, **34**, 80–96.

Berthouex, P.M., Hunter, W.G., and Pallesen, L. (1978), "Monitoring sewage treatment plants: Some quality control aspects," *Journal of Quality Technology,* **10**, 139–149.

Bie, Q., Qiu, D., Hu, H., and Ju, B. (2010), "Spatial and temporal distribution characteristics of hand-foot-mouth disease in china," *Journal of Geo-Information Science*, **12**, 380–384.

Capizzi, G., and Masarotto, G. (2008), "Practical design of generalized likelihood ratio control charts for autocorrelated data," *Technometrics*, **50**, 357–370.

Capizzi, G., and Masarotto, G. (2011), "A least angle regression control chart for multidimensional data," *Technometrics*, **53**, 285–296.

Capizzi, G., and Masarotto, G. (2013), "Phase I distribution-free analysis of univariate data," *Journal of Quality Technology*, **45**, 273–284.

Chakraborti, S., and Graham, M.A. (2019), "Nonparametric (distribution-free) control charts: An updated overview and some results," *Quality Engineering*, **31**, 523–544.

Chakraborti, S., Qiu, P., and Mukherjee, A. (2015, ed.), "Special issue on nonparametric statistical process control charts," *Quality and Reliability Engineering International*, **31**, 1–152.

Chen, H., Zeng, D., and Yan, P. (2010), *Infectious Disease Informatics*, Springer: New York.

Chen, K., Hui, Y., and Kumara, S. (2018), "Parallel computing and network analytics for fast Industrial Internet-of-Things (IIoT) machine information processing and condition monitoring," *Journal of Manufacturing Systems*, **46**, 282–293.

Colosimo, B. M., Cicorella, P., Pacella, M., and Blaco, M. (2014), "From profile to surface monitoring: SPC for cylindrical surfaces via Gaussian processes," *Journal of Quality Technology*, **46**, 95–113.

Deng, H., Runger, G., and Tuv, E. (2012), "System monitoring with real-time contrasts," *Journal of Quality Technology*, **44**, 9–27.

Dong, H., Chen, N., and Wang, K. (2019), "Modeling and change detection for count-weighted multi-layer networks," *Technometrics*, DOI: 10.1080/00401706.2019.1625812.

Donoho, D.L., and Johnstone, I.M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, **81**, 425–455.

Duchesne, C., Liu, J., and MacGregor, J. (2012) "Multivariate image analysis in the process industries: A review," *Chemometrics and Intelligent Laboratory Systems,* **117**, 116–128.

Feng, L., and Qiu, P. (2018), "Difference detection between two images for image monitoring," *Technometrics*, **60**, 345–359.

Ferrer, A. (2007), "Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process," *Quality Engineering*, **19**, 311–325.

Frazier, P.S., and Page, K.J. (2000), "Water body detection and delineation with Landsat TM data," *Photogrammetric Engineering & Remote Sensing*, **66**, 1461–1467.

Fu, Y., and Jeske, D.R. (2014), "SPC methods for non-stationary correlated count data with application to network surveillance," *Journal of Applied Stochastic Models in Business and Industry,* **30**, 708–722.

Gan, F. (1993), "An optimal design of CUSUM control charts for binomial counts," *Journal of Applied Statistics*, **20**, 445–460.

Graham, M.A., Human, S.W., and Chakraborti, S. (2010), "A phase I nonparametric Shewhart-type control chart based on the median," *Journal of Applied Statistics*, **37**, 1795–1813.

Graham, M.A., Chakraborti, S., and Human, S.W. (2011), "A nonparametric exponentially weighted moving average signed-rank chart for monitoring location," *Computational Statistics and Data Analysis*, **55**, 2490–2503.

Gygi, C., Williams, B., and DeCarlo, N. (2012), *Six Sigma For Dummies (2nd Ed)*, New York: John Wiley & Sons.

Hausner, M.B., Huntington, J.L., Nash, C., Morton, C., McEvoy, D.J., Pilliod, D.S., Hegewisch, K.L., Daudert, B., Abatzoglou, J.T., and Grant, G. (2018), "Assessing the effectiveness of riparian restoration projects using Landsat and precipitation data from the cloud-computing application ClimateEngine.org," *Ecological Engineering*, **120**, 432–440.

Hawkins, D.M. (1991), "Multivariate quality control based on regression-adjusted variables," *Technometrics*, **33**, 61–75.

Hawkins, D.M., and Deng, Q. (2010), "A nonparametric change-point control chart," *Journal of Quality Technology*, **42**, 165–173.

Hawkins, D.M., and Olwell, D.H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer-Verlag.

Hawkins, D.M., Qiu, P., and Kang, C.W. (2003), "The changepoint model for statistical process control," *Journal of Quality Technology*, **35**, 355–366.

He, S.G., and Zhang, C.Y. (2011), "Support vector data description based multivariate cumulative sum control chart," *Advanced Materials Research*, **314**, 2482–2485.

Holland, M.D., and Hawkins, D.M. (2014), "A control chart based on a nonparametric multivariate change-point model," *Journal of Quality Technology*, **46**, 63–77.

Jackson, J.E. (1991), *A User's Guide to Principal Components,* vol. 587: New York: John Wiley & Sons.

Jeske, D.R. (2016), “Determining the Phase 1 study sample size to control the accuracy of the conditional in-control ARL of a normal-theory CUSUM,” *Quality and Reliability Engineering International*, **32**, 2499–2504.

Jeske, D.R., Stevens, N.T., Tartakovsky, A.G., and Wilson, J.D. (2018), “Statistical methods for network surveillance,” *Applied Stochastic Models in Business and Industry*, **34**, 425–445.

Jiang, W., Han, S.W., Tsui, K.L., and Woodall, W.H. (2011), “Spatiotemporal surveillance methods in the presence of spatial correlation,” *Statistics in Medicine*, **30**, 569-583.

Jin, N., Zhou, S., and Chang, T.S. (2004), “Identification of impacting factors of surface defects in hot rolling processes using multi-level regression analysis,” *Transactions of the North American Manufacturing Research Institute of SME*, **32**, 557–564.

Johnson, R.A., and Wichern, D.W. (2007), *Applied Multivariate Statistical Analysis (6th edition)*, Upper Saddle River, NJ: Pearson Prentice Hall.

Jones-Farmer, L.A., Jordan, V., and Champ, C.W. (2009), “Distribution-free phase I controlcharts for subgroup location,” *Journal of Quality Technology*, **41**, 304–317.

Jones-Farmer, L.A., Woodall, W.H., Steiner, S.H., and Champ, C.W. (2014), “An overview of Phase I analysis for process improvement and monitoring,” *Journal of Quality Technology*, **46**, 265–280.

Kang, L., and Albin, S. L. (2000), “On-Line Monitoring When the Process Yields a Linear Profile,” *Journal of Quality Technology*, 32, 418–426.

Kim, K., Mahmoud, M. A., and Woodall, W. H. (2003), “On the Monitoring of Linear Profiles,” *Journal of Quality Technology*, 35, 317–328.

Kite-Powell, A., Ofori-Addo, A., and Hamilton, J. (2010), *ESSENCE User Guide (Version 1.0)*, Florida Department of Health, Bureau of Epidemiology.

Knox, E., and Bartlett, M. (1964), “The detection of space-time interactions,” *Journal of the Royal Statistical Society (Series C)*, **13**, 25–30.

Kourti, T, MacGregor, J.F. (1996), “Multivariate SPC methods for process and product monitoring,” *Journal of Quality Technology*, **28**, 409–428.

Kulldorff, M. (1997), “A spatial scan statistic,” *Communications in Statistics-Theory and Methods*, **26**, 1481–1496.

Lee, H. C., and Apley, D. W. (2011), “Improved design of robust exponentially weighted moving average control charts for autocorrelated processes,” *Quality and Reliability Engineering International*, **27**, 337–352.

Lee, C.H., and Yoon, H.J. (2017), “Medical big data: promise and challenges,” *Kidney Research and Clinical Practice*, **36**, 3–11.

Li, J., Jeske, D.R., Zhou, Y., and Zhang, X. (2019a), “A wavelet-based nonparametric CUSUM control chart for autocorrelated processes with applications to network surveillance,” *Quality and Reliability Engineering International*, **5**, 644–658.

Li, J., and Qiu, P. (2016), “Nonparametric dynamic screening system for monitoring correlated longitudinal data,” *IIE Transactions*, **48**, 772–786.

Li, J., and Qiu, P. (2017), “Construction of an efficient multivariate dynamic screening system,” *Quality and Reliability Engineering International*, **33**, 1969–1981.

Li, J., Zhang, X., and Jeske, D.R. (2013), “Nonparametric multivariate CUSUM control charts for location and scale changes,” *Journal of Nonparametric Statistics,* **25**, 1–20.

Li, W., and Qiu, P. (2019), “A general charting scheme for monitoring serially correlated data with short-memory dependence and nonparametric distributions,” *IISE Transactions*, in press.

Li, W., Xiang, D., Tsung, F., and Pu, X. (2019b), “A diagnostic procedure for high-dimensional data streams via missed discovery rate control,” *Technometrics*, DOI: 10.1080/00401706.2019.1575284.

Li, S.Y., Tang, L.C., and Ng, S.H. (2010), “Nonparametric CUSUM and EWMA control charts for detecting mean shifts,” *Journal of Quality Technology*, **42**, 209–226.

Lin, H.D., Chung, C.Y., and Lin, W.T. (2008), “Principal component analysis based on wavelet characteristics applied to automated surface defect inspection,” *WSEAS Transactions on Computer Research*, **3**, 193–202.

Liu, R.Y. (1995), “Control charts for multivariate processes,” *Journal of the American Statistical Association*, **90**, 1380–1387.

Loredo, E., Jearkpaporn, D., and Borror, C. (2002), “Model-based control chart for autoregressive and correlated data,” *Quality and Reliability Engineering International*, **18**, 489–496.

Lowry, C.A., Woodall, W.H., Champ, C.W., and Rigdon, S.E. (1992), “A multivariate exponentially weighted moving average control chart,” *Technometrics*, **34**, 46–53.

Maheshwari, A. (2019), *Data Analytics Made Accessible*, Amazon Digital Services LLC.

Marshall, J.B., Spitzner, D.J., and Woodall, W.H. (2007), “Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time,” *Statistics in Medicine*, **26**, 1579–1593.

May, G., and Spanos, C. (2006), *Fundamentals of Semiconductor Manufacturing and Process Control*, New York: John Wiley & Sons.

McCulloh, I., and Carley, K.M. (2011), “Detecting change in longitudinal social networks,” *Journal of Social Structure*, **12**, 1–37.

Megahed, F.M., and Jones-Farmer, L.A. (2015), “Statistical perspectives on ’big data’,” In *Frontiers in statistical quality control 11* (eds Knoth S. and Schmid W.), 29–47, Cham: Springer.

Megahed, F.M., Kensler, J.L., Bedair, K., Woodall, W.H. (2011), “A note on the ARL of two-sided Bernoulli-based CUSUM control charts,” *Journal of Quality Technology*, **43**, 43–49.

Megahed, F.M., Woodall, W.H., and Camelio, J. A. (2011), “A review and perspective on control charting with image data,” *Journal of Quality Technology*, **43**, 83–98.

Megahed, F.M., Wells, L.J., Camelio, J.A., and Woodall, W.H. (2012), “A spatiotemporal method for the monitoring of image data,” *Quality Reliability and Engineering International*, **28**, 967–980.

Montgomery, D.C. (2012), *Introduction to Statistical Quality Control,* New York: John Wiley & Sons.

Montgomery, D.C., and Mastrangelo, C.M. (1991), “Some statistical process control methods for autocorrelated data,” *Journal of Quality Technology*, **23**, 179–193.

Moustakides, G.V. (1986), “Optimal stopping times for detecting changes in distributions,” *The Annals of Statistics*, **14**, 1379–1387.

Neil, J., Hash, C., Brugh, A., Fisk, M., and Storlie, C.B. (2013), "Scan statistics for the online detection of locally anomalous subgraphs," *Technometrics*, **55**, 403–414.

Ning, W., Yeh, A.B., Wu, X., and Wang, B. (2015), "A nonparametric phase I control chart for individual observations based on empirical likelihood ratio," *Quality and Reliability Engineering International*, **31**, 37–55.

Ning, X, and Tsung, F. (2013), "Improved design of kernel distance–based charts using support vector methods," *IIE Transactions*, **45**, 464–476.

Page, E.S. (1954), "Continuous inspection scheme," *Biometrika*, **41**, 100–115.

Parastatidis, D., Mitraka, Z., Chrysoulakis, N., and Abrams, M. (2017), "Online global land surface temperature estimation from Landsat," *Remote Sensing*, **9**, 2072–4292.

Patterson, E., and Wang, Z. (1991), "Towards full field automated photoelastic analysis of complex components," *Strain*, **27**, 49–53.

Paynabar, K., Qiu, P., and Zou, C. (2016), "A change point approach for Phase-I analysis in multivariate profiles monitoring and diagnosis," *Technometrics*, **58**, 191–204.

Prats-Montalban, J.M., and Ferrer, A. (2014), "Statistical process control based on Multivariate Image Analysis: A new proposal for monitoring and defect detection," *Computers and Chemical Engineering,* **71**, 501–511.

Qiu, P. (2008), "Distribution-free multivariate process control based on log-linear modeling," *IIE Transactions*, **40**, 664–677.

Qiu, P. (2014), *Introduction to Statistical Process Control,* Boca Raton, FL: Chapman Hall/CRC.

Qiu, P. (2018), "Some perspectives on nonparametric statistical process control," *Journal of Quality Technology*, **50**, 49–65.

Qiu, P., and Hawkins, D.M. (2001), "A rank based multivariate CUSUM procedure," *Technometrics*, **43**, 120–132.

Qiu, P., and Li, Z. (2011), "On nonparametric statistical process control of univariate processes," *Technometrics*, **53**, 390–405.

Qiu, P., Li, W., and Li, J. (2019), "A new process control chart for monitoring short-range serially correlated data," *Technometrics*, DOI: 10.1080/00401706.2018.1562988.

Qiu, P., and Nguyen, T. (2008), "On image registration in magnetic resonance imaging," *IEEE Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics*, 753–757.

Qiu, P., Xia, Z., and You, L. (2019), "Process monitoring ROC curve for evaluating dynamic screening methods," *Technometrics*, DOI: 10.1080/00401706.2019.1604434.

Qiu, P., and Xiang, D. (2014), "Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behavior," *Technometrics*, **56**, 248–260.

Qiu, P., and Xiang, D. (2015), "Surveillance of cardiovascular diseases using a multivariate dynamic screening system," *Statistics in Medicine*, **34**, 2204–2221.

Qiu, P., and Xing, C. (2013), "On nonparametric image registration," *Technometrics*, **55**, 174–188.

Qiu, P., Zi, X., and Zou, C. (2018), "Nonparametric dynamic curve monitoring," *Technometrics*, **60**, 386–397.

Qiu, P., Zou, C., and Wang, Z. (2010), "Nonparametric profile monitoring by mixed effects modeling (with discussions)," *Technometrics*, **52**, 265–277.

Rauf, S., Dauksher, W.J., Clemens, S.B., and Smith, K.H. (2002), "Model for a multiple-step deep si etch process," *Journal of Vacuum Science and Technology A*, **20**, 1177–1190.

Reis, M.S., and Gins, G. (2017), "Industrial process monitoring in the big data/industry 4.0 era: from detection, to diagnosis, to prognosis," *Processes*, **5**, 35.

Ren, H., Chen, N., and Wang, Z. (2019), "Phase-II monitoring in multichannel profile observations," *Journal of Quality Technology*, DOI: 10.1080/00224065.2018.1507556.

Roberts, S.V. (1959), "Control chart tests based on geometric moving averages," *Technometrics*, **1**, 239–250.

Runger, G.C., and Willemain, T.R. (1995), "Model-based and model-free control of autocorrelated processes," *Journal of Quality Technology*, **27**, 283–292.

Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014), "Anomaly detection in online social networks," *Social Networks*, **39**, 62–70.

Shewhart, W.A. (1931), *Economic Control of Quality of Manufactured Product,* New York: D. Van Nostrand Company.

Siegel, E. (2016), *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die,* New York: John Wiley & Sons.

Sohn, H., Park, G., Wait, J. R., Limback, N. P., and Farrar, C. R. (2004), "Wavelet-based active sensing for delamination detection in composite structures," *Smart Materials and Structures*, **13**, 153–160.

Sun, R, Tsung, F. (2003), "A kernel-distance-based multivariate control chart using support vector methods," *International Journal of Production Research*, **41**, 2975–2989.

Tuv, E., and Runger, G. (2003), "Learning patterns through artificial contrasts with application to process control," *WIT Transactions on Information and Communications Technologies*, **29**, 63–72.

Vander Wiel, S.A. (1996), "Monitoring processes that wander using integrated moving average models," *Technometrics*, **38**, 139–151.

Vittek, M., Brink, A., Donnay, F., Simonetti, D., and Desclee, B. (2014), "Land cover change monitoring using Landsat MSS/TM satellite image data over West Africa between 1975 and 1990," *Remote Sensing*, **6**, 658–676.

Wadsworth, H.M., Stephens, K.S., and Godfrey, A.B. (2002), *Modern Methods for Quality Control and Improvement,* New York: John Wiley & Sons.

Wang, K., and Jiang, W. (2009), "High-dimensional process monitoring and fault isolation via variable selection," *Journal of Quality Technology*, **41**, 247–258.

Wang, A., Wang, K., and Tsung, F. (2014), "Statistical surface monitoring by spatial-structure modeling," *Journal of Quality Technology*, **46**, 359–376.

Wardell, D.G., Moskowitz, H., and Plante, R.D. (1994), "Run length distributions of special-cause control charts for correlated processes," *Technometrics*, **36**, 3–17.

Weib, C.H., and Testik, M.C. (2009), “CUSUM monitoring of first-order integer-valued autoregressive processes of Poisson counts,” *Journal of Quality Technology*, **41**, 389–400.

Weng, Q., Fu, P., and Gao, F. (2014), “Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data,” *Remote Sensing of Environment*, **145**, 55-67.

Woodall, W.H. (2000), “Controversies and contradictions in statistical process control,” *Journal of Quality Technology*, **32**, 341–350.

Woodall, W.H. (2006), “Use of control charts in health-care and public-health surveillance (with discussions),” *Journal of Quality Technology*, 89–104.

Woodall, W.H. (2007), “Current research on profile monitoring,” *Producão*, **17**, 420–425.

Woodall, W.H., Zhao, M., Paynabar, K., Sparks, R., and Wilson, J.D. (2017), “An overview and perspective on social network monitoring,” *IISE Transactions*, **49**, 354–365.

Xu, S., and Jeske, D.R. (2017), “Repeated SPRT charts for monitoring INAR(1) processes,” *Quality Reliability Engineering International*, **33**, 2615–2624.

Yan, H., Paynabar, K., and Shi, J. (2015), “Image-based process monitoring using low-rank tensor decomposition,” *IEEE Transactions on Automation Science and Engineering,* **12**, 216–227.

Yan, H., Paynabar, K., and Shi, J. (2018), “Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition,” *Technometrics*, **60**, 181–197.

You, L., and Qiu, P. (2019), “Fast computing for dynamic screening systems when analyzing correlated data,” *Journal of Statistical Computation and Simulation*, **89**, 379–394.

You, L., and Qiu, P. (2020), “An effective method for online disease risk monitoring,” *Technometrics*, DOI: 10.1080/00401706.2019.1625813.

Yu, L., Zwetsloot, I.M., Stevens, N.T., Wilson, J.D. and Tsui K.L. (2019), “Monitoring dynamic networks: a simulation-based strategy for comparing monitoring methods and a comparative study,” arXiv preprint: 1905.10302.

Zang, Y., and Qiu, P. (2018a), “Phase I monitoring of spatial surface data from 3D printing,” *Technometrics*, **60**, 169–180.

Zang, Y., and Qiu, P. (2018b), "Phase II monitoring of free-form surfaces: an application to 3D printing," *Journal of Quality Technology*, **50**, 379–390.

Zanter, K. (2016), *Landsat 8 (L8) Data Users Handbook (version 2),* Department of the Interior, U.S. Geological Survey.

Zou, C., and Qiu, P. (2009), "Multivariate statistical process control using LASSO," *Journal of the American Statistical Association*, **104**, 1586–1596.

Zou, C., Qiu, P., and Hawkins, D. (2003), "Nonparametric control chart for monitoring profiles using change point formulation and adaptive smoothing," *Statistica Sinica*, **19**, 1337–1357.

Zou, C., and Tsung, F. (2010), "Likelihood ratio-based distribution-free EWMA control charts," *Journal of Quality Technology*, **42**, 1-23.

Zou, C., and Tsung, F. (2011), "A multivariate sign EWMA control chart," *Technometrics*, **53**, 84–97.

Zou, C., Tsung, F., and Wang, Z. (2008), "Monitoring profiles based on nonparametric regression methods," *Technometrics*, 50, 512–526.

Zou, C., Wang, Z., and Tsung, F. (2012), "A spatial rank-based multivariate EWMA control chart," *Naval Research Logistics*, **59**, 91–110.

Zou, C., Wang, Z., Zi, X., and Jiang, W. (2015), "An efficient online monitoring method for high-dimensional data streams," *Technometrics*, **57**, 374–387.

Zou, C., Zhang, Y., and Wang, Z. (2006), "Control chart based on change-point model for monitoring linear profiles," *IIE Transactions*, 38, 1093–1103.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, N., and Li, J. (2017), "Modeling and change detection of dynamic network data by a network state space model," *IISE Transactions*, **49**, 45–57.

Zwetsloot, I.M., and Woodall, W.H. (2017), "A head-to-head comparative study of the conditional performance of control charts based on estimated parameters," *Quality Engineering*, **29**, 244–253.