From Disease Surveillance To Image Monitoring: Challenges in Modern SPC Applications

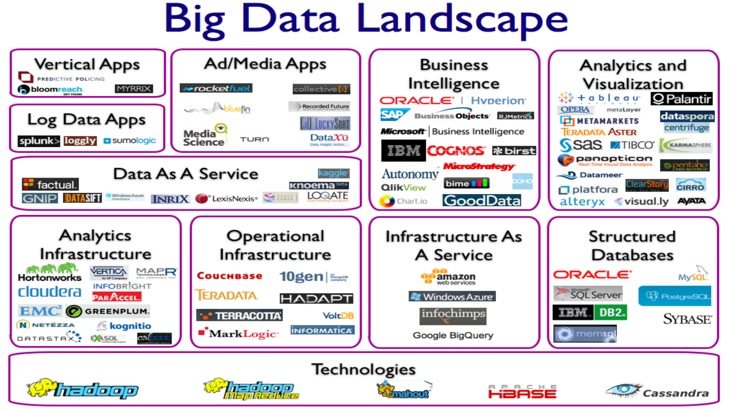
Peihua Qiu

Department of Biostatistics
University of Florida
E-mail: pqiu@ufl.edu

December 2018
SCPE-2018

Big Data


DATA


Big Data Landscape
Vertical Apps
Ad/Media Apps
Business Intelligence
Analytics and Visualization
Log Data Apps
Data As A Service
Analytics Infrastructure
Operational Infrastructure
Infrastructure As A Service
Structured Databases
Technologies
Google BigQuery

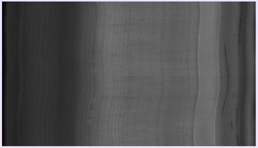
Big Data (con'd)

- One common feature: data streams (i.e., new data keep coming)
- One fundamental task: monitor the behavior/performance of the underlying processes
- One useful tool: statistical process control (SPC)

Applications in Industry

(a)

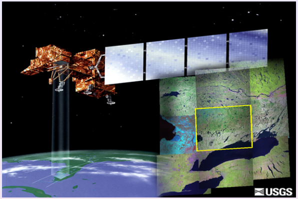
(b)



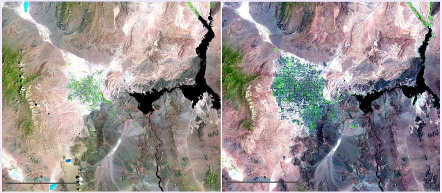
(c)

Figure 1: (a) An imaging system for checking steel surface and detecting cracks and defects. (b) A camera. (c) An observed image of steel surface.

- Stress and strain analysis of products (Patterson and Wang 1991)
- Anomaly detection of rolling processes (Jin et al. 2004, Yan et al. 2016)
- Inspection of composite material fabrication (Sohn et al. 2004)
- Quality control in liquid crystal display manufacturing (Jiang et al. 2005)



(a)

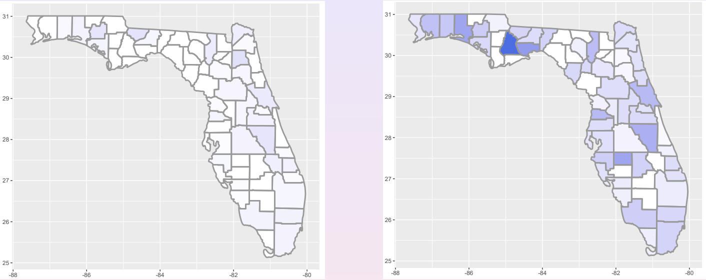


(b)

Figure 2: (a) Landsat satellite imaging system. (b) Two Landsat images of Las Vegas (top-left) taken in 1984 and 2007.

Earth surface and earth resources in land use, forest science, climate science, agriculture forecasting, ecological and ecosystem monitoring, water resources, and other natural science disciplines.

Applications in Health Sciences



(a) (b)

Figure 3: Observed influenza-like illness incidence rates in Florida on 06/01/2012 (left) and 12/01/2012 (right).

In the nation, there are thousands of spatial locations of observations.

Statistical Process Control (SPC)

- Basic SPC charts
  - Shewhart chart (Shewhart 1931)
  - CUSUM chart (Page 1954)
  - EWMA chart (Roberts 1959)
  - CPD chart (Hawkins, Qiu and Kang 2003)
- Monitoring **production lines** in manufacturing industries
- Assumptions: process observations are **independent** and **identically distributed** with a **parametric** (e.g., normal) distribution.
- Conventional charts are **unreliable** whe their assumptions are violated.

Statistical Process Control (con'd)

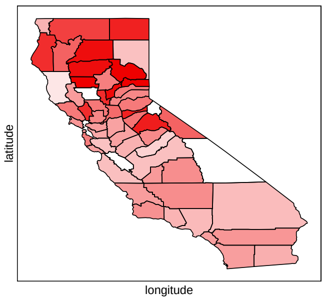
- Nonparametric SPC: data ranking or data categorization (Capizzi 2015, Chakrabort et al. 2001, Qiu and Hawkins 2001, Qiu 2018)
- Dynamic screening system (Qiu and Xiang 2014, 2015, Qiu et al. 2018)
- SPC for correlated data (Apley and Tsung 2002, Li and Qiu 2018, Runger 2002)
- Profile monitoring
  - Univariate profile data (Kim et al. 2003, Qiu et al. 2010, Zou et al. 2008)
  - Multiple profile data (Paynabar et al. 2013, 2016)
  - Spatial profile data: Gaussian process modelling (Colosimo et al. 2014, Wang et al. 2014), nonparametric surface estimation (Zang and Qiu 2018a,b)

Disease Surveillance

Spatio-temporal (ST) disease incidence rate:

$$\lambda(t, \mathbf{s}) = \lim_{dt \to 0, d\mathbf{s} \to 0} \frac{E(N(t, \mathbf{s}; dt, d\mathbf{s}))}{M(t, O(\mathbf{s}))dtd\mathbf{s}}, \qquad \text{for } \mathbf{s} \in \Omega, \quad t \in [0, T].$$

Lung cancer incidence rates in 58 counties of California in 06/2006



- Spatio data variation, ST data correlation, seasonality
- Covariates (e.g., PM 2.5, temperature)
- In-control (IC) vs out-of-control (OC)

Spatio-Temporal Modelling of IC Data

- Observed spatial data within a given time interval that are considered IC.
- Existing methods
  - Log-Gaussian Cox process (LGCP) modelling (e.g., Cressie and Huang 1999, Diggle 2013)
  - Bayesian dynamic spatio-temporal modelling (e.g., Finley et al. 2015)
  - Parametric regression (Lindström et al. 2015)
  - Kernel smoothing by assuming independence (Kafadar 1996, Kelsall and Diggle 1998)
- Yang and Qiu (2018, *Statistics in Medicine*)

- Model: for $j = 1, 2, \ldots, m_i$ and $i = 1, 2, \ldots, n$,

$$y(t_i, \mathbf{s}_{ij}) = \lambda(t_i, \mathbf{s}_{ij}) + \varepsilon(t_i, \mathbf{s}_{ij})$$

- $t_i \in [0, T]$, $\mathbf{s}_{ij} \in \Omega$
- Covariance function: $V(\mathbf{u}, \mathbf{v}) = \sigma(\mathbf{u})\sigma(\mathbf{v})Corr(\varepsilon(\mathbf{u}), \varepsilon(\mathbf{v}))$, for $\mathbf{u}, \mathbf{v} \in [0, T] \times \Omega$
- Model estimation: local linear ST kernel (LLSTK) smoothing

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3 \in R} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{y(t_i, \mathbf{s}_{ij}) - [a + b(t_i - t) + c(s_{x,ij} - s_x) + d(s_{y,ij} - s_y)]\}^2 K_1\left(\frac{t_i - t}{h_t}\right) K_2\left(\frac{d_E(\mathbf{s}_{ij}, \mathbf{s})}{h_s}\right)$$

- LLSTK estimator: $\widehat{\lambda}(t, \mathbf{s})$

Spatio-Temporal Modelling of IC Data (con'd)

- K_1 and K_2: Epanechnikov kernel function $K_e(u) = \frac{3}{4}(1-u^2)I(|u| \le 1)$
- Selection of bandwidths h_t and h_s: modified cross-validation (CV)
  - $\text{CV}(h_t, h_s) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{1}{m_i}\sum_{j=1}^{m_i}[\widehat{\lambda}_{-(ij)}(t_i, \mathbf{s}_{ij}) - y(t_i, \mathbf{s}_{ij})]^2\}$
  - $\widetilde{K}_\epsilon(u) = \frac{4}{4-3\epsilon-\epsilon^3}\begin{cases} \frac{3}{4}(1-u^2)I(|u| \le 1), & \text{if } |u| \ge \epsilon; \\ \frac{3(1-\epsilon^2)}{4\epsilon}|u|, & \text{if } |u| < \epsilon, \end{cases}$
- **Theorem:** Under some regularity conditions,

$$a_N^2 E\left(\widehat{\lambda}(t,\mathbf{s}) - \lambda(t,\mathbf{s})\right)^2 = 0, \qquad \text{for any } (t,\mathbf{s}) \in (0,1)\times(0,1)^2,$$

where $a_N = \min\{\frac{1}{h_t^{2-\delta}}, \frac{1}{h_s^{2-\delta}}, (\frac{N|\mathbf{H}|}{C''_{N,\mathbf{w}}})^{0.5-\delta}\}$ and $0 < \delta < 0.5$.

A Simulation Example

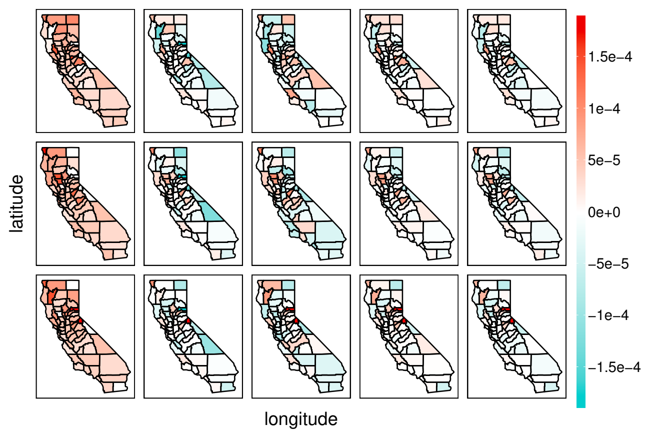
- $\lambda(t, \mathbf{s}) =$
  $0.5 + 0.3 \sin\left(\frac{\pi}{2} + \pi s_x\right) \sin\left(\frac{\pi}{2} + \pi s_y\right) + 0.15 \cos\left(\frac{3\pi}{2} + 2\pi t\right)$
- Spatial data are generated by using R package $neuRosim$ with parameters $(\rho_t, \rho_\mathbf{s})$

MSE values in scale 10^{-3}

$(\rho_t, \rho_\mathbf{s})$	DSTM	LGCP	WAS	LLSTK
(0.5,0.5)	5.95	7.07	4.85	1.73
(0.5,0.3)	5.26	6.90	3.84	0.92
(0.3,0.5)	5.34	7.11	4.87	0.69
(0.3,0.3)	4.54	6.91	3.85	0.37

Revisit the Lung Cancer Data

Observed lung cancer data (1st column) in CA: March 2007 (1st row), Oct 2010 (2nd row) and Dec 2011 (3rd column); Residual plots of LGCP (2nd column), DSTM (3rd column), WAS (4th column) and LLSTK (5th column).



- Estimated regular pattern from IC data: $\widehat{\lambda}(t,\mathbf{s})$, $\widehat{V}(\boldsymbol{u},\boldsymbol{v})$
- Disease incidence rates to monitor: $\{y(t_i^*,\mathbf{s}_{ij}^*), j=1,2,\ldots,m_i^*, i=1,2,\ldots\}$
- For each i, $\boldsymbol{y}(t_i^*)=(y(t_i^*,\mathbf{s}_{i1}^*), y(t_i^*,\mathbf{s}_{i2}^*),\ldots,y(t_i^*,\mathbf{s}_{im_i^*}^*))'$, $\widehat{\boldsymbol{\lambda}}(t_i^*)=(\widehat{\lambda}(t_i^*,\mathbf{s}_{i1}^*),\widehat{\lambda}(t_i^*,\mathbf{s}_{i2}^*),\ldots,\widehat{\lambda}(t_i^*,\mathbf{s}_{im_i^*}^*))'$, $\widehat{\boldsymbol{\varepsilon}}(t_i^*)=\boldsymbol{y}(t_i^*)-\widehat{\boldsymbol{\lambda}}(t_i^*)$.
- Sequential data decorrelation using covariance Cholesky decomposition
- Decorrelated and standardized data: $\widehat{\widetilde{\boldsymbol{e}}}(t_1^*),\widehat{\widetilde{\boldsymbol{e}}}(t_2^*),\ldots,\widehat{\widetilde{\boldsymbol{e}}}(t_i^*)$.

Disease Surveillance (con'd)

- CUSUM chart for detecting upward shift: for $i \geq 1$,

$$C_i^+ = \max\left(0, C_{i-1}^+ + \frac{\widehat{\widetilde{\boldsymbol{e}}}(t_i^*)'\widehat{\widetilde{\boldsymbol{e}}}(t_i^*) - m_i^*}{\sqrt{2m_i^*}} - k\right),$$

  where $C_0^+ = 0$ and $k > 0$ is an allowance constant.
- The CUSUM chart gives a signal when $C_i^+ > \gamma$, where $\gamma > 0$ is a control limit.
- Block bootstrap procedure for determining γ from the IC data.

Application to the ILI Data

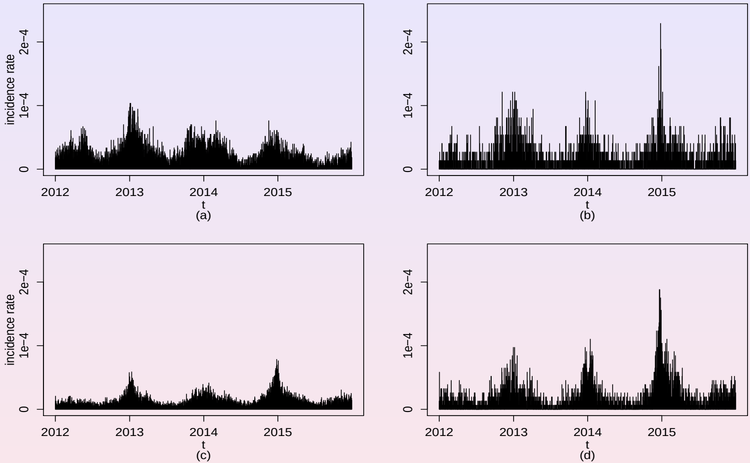


Figure 4: Observed ILI incidence rates in four representative counties of Florida: Collier (plot (a)), Nassau (plot (b)), Pinellas (plot (c)), and Santa Rosa (plot (d)).

MCUSUM: multivariate CUSUM (Crosier 1988); MEWMA: multivariate EWMA (Lowry et al. 1992); T^2: Hotelling's T^2.

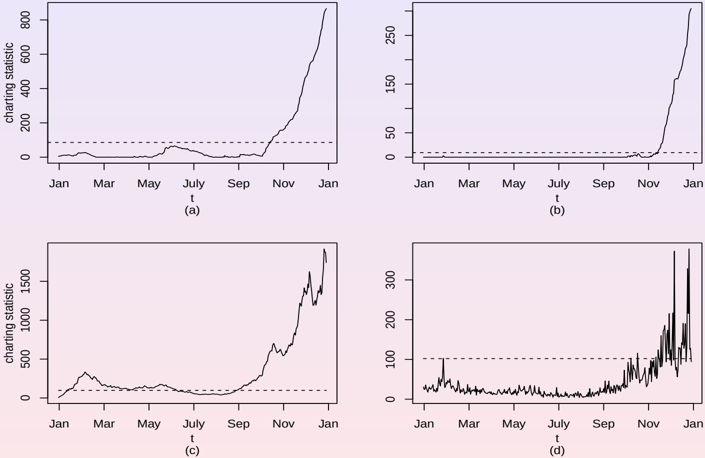


Figure 5: Charts of NEW (plot (a)), MCUSUM (plot (b)), MEWMA (plot (c)), and T^2 (plot (d)). The Signal time of NEW is Oct 16, 2014.

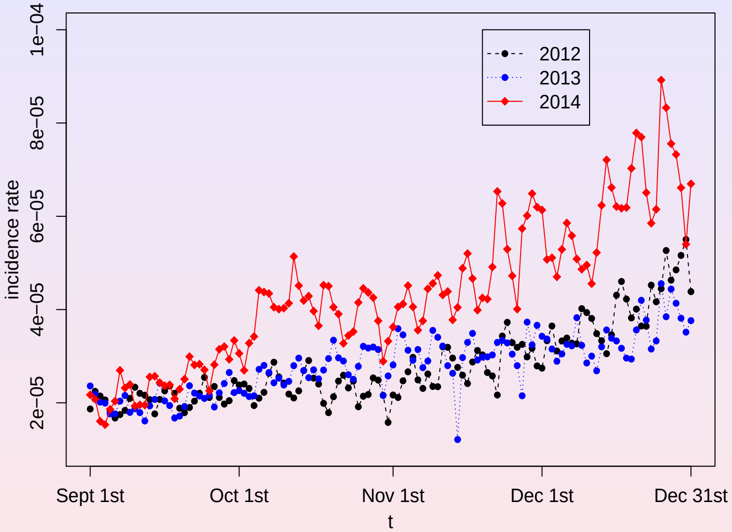


Figure 6: Disease incidence rates of the entire Florida state during Sept 1 and Dec 31 in years 2012, 2013 and 2014.

Application to the ILI Data (con'd)

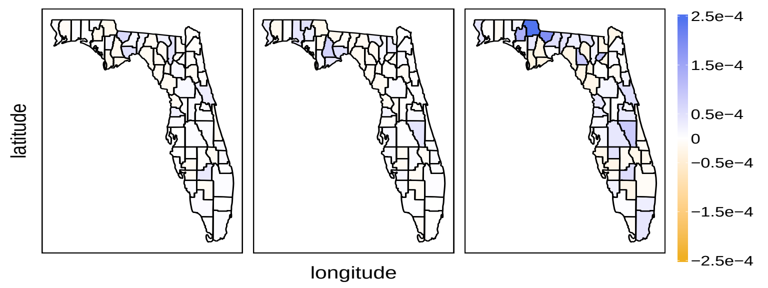


Figure 7: Residual maps on October 16 of the years 2012, 2013 and 2014.

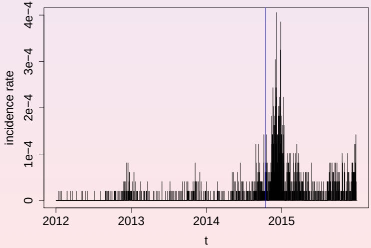
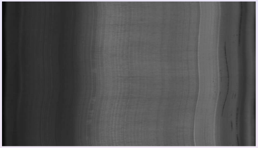


Figure 8: Disease incidence rates of the Jackson county in 2012-2015.

Image Comparison and Monitoring

(a)

(b)

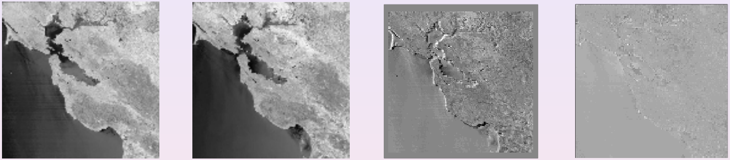


(c)

- Image comparison and monitoring are challenging
  - Images have edges and other complicated structures
  - Observed images contain noise and other contaminations
  - Conventional smoothing methods are inapplicable for estimating true images (Qiu 2005, 2007)
  - Images of different products are most likely geometrically mismatched, because the relative positions between the camera and different products are rarely the same.
  - *Image registration* (Xing and Qiu 2011, Qiu and Xing 2013a,b)

Image Registration

Satellite images of the San Francisco bay area in 1990 and 1999.



- feature-based vs intensity-based
- parametric vs nonparametric

Image Registration (con'd)

- Reference image $R(x, y)$ and moved image $M(x, y)$
- $Z_R(x_i, y_j) = R(x_i, y_j) + \varepsilon_R(x_i, y_j)$,
  $Z_M(x_i, y_j) = M(x_i, y_j) + \varepsilon_M(x_i, y_j)$
- Geometric transformation: $\mathbf{T}(x, y) = (T_1(x, y), T_2(x, y))$
- $M(T_1(x, y), T_2(x, y)) = R(x, y), \quad (x, y) \in \Omega$
- Goal: Estimate $\mathbf{T}(x, y)$ from $\{Z_R(x_i, y_j)\}$ and $\{Z_M(x_i, y_j)\}$
- Rigid-body transformation:

$$\begin{aligned} T_1(x, y) &= x\cos(\phi) + y\sin(\phi) + \Delta x, \\ T_2(x, y) &= -x\sin(\phi) + y\cos(\phi) + \Delta y \end{aligned}$$

  Euclidean distance between any two points in an image will not change after the transformation
- $\boldsymbol{\theta} = (\phi, \Delta x, \Delta y)$

Edge-Based Image Registration

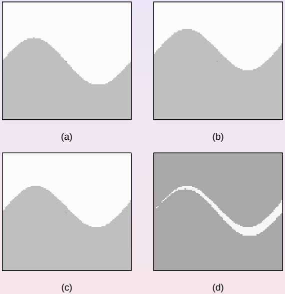
- Detected edge points: D_R and D_M
- Feature-matching: Qiu and Xing (2013, *Signal Processing*)
- When ϕ is small, we have

$$T_1(x,y) \approx x + y\phi + \Delta x, \;\; T_2(x,y) \approx -x\phi + y + \Delta y.$$

- Define $Q_E(\phi, \Delta x, \Delta y) =$
$\sum_{(x_i,y_j)\in D_R} \left(\widehat{T}_1(x_i,y_j) - x_i - y_j\phi - \Delta x\right)^2 +$
$\left(\widehat{T}_2(x_i,y_j) - y_j + x_i\phi - \Delta y\right)^2$
- $\widehat{\boldsymbol{\theta}}_E = \mathbf{A}^{-1}\mathbf{B}$
- Under some regularity conditions, $\widehat{\boldsymbol{\theta}}_E = \boldsymbol{\theta}_E + o_p(\delta_n)$.
- Feng and Qiu (2018, *Technometrics*): $\widehat{\boldsymbol{\theta}}_C$
- Nonparametric image registration: Qiu and Xing (2013, *Technometrics*)

How to Compare Two Images

If we use pixelwise differences to compare two images, then those around the edges will dominate the overall difference.



(a) Reference image, (b) moved image, (c) registered moved image, and (d) pixelwise difference between (a) and (c). The whiter the color, the larger the intensity value.

Image Comparison

H_0: There is a rigid-body transformation $\mathbf{T}(x, y)$ such that
$$M(\mathbf{T}(x, y)) \equiv R(x, y)$$
H_1: No rigid-body transformation exists such that H_0 is true.

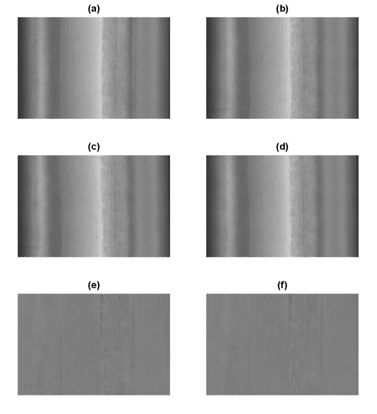
- Continuity-region-based
- Edge-based
- Hybrid
- $G = \Omega \backslash \left(D_R(h_G) \bigcup D_M(h_G)\right)$
- Test statistic

$$U = \frac{1}{\sqrt{2|G|}} \sum_{(x_i, y_j) \in G} \left(\left(Z_R(x_i, y_j) - Z_M(\widehat{\mathbf{T}}(x_i, y_j) \right)^2 / (2\widehat{\sigma}^2) - 1.0 \right)$$

- A bootstrap method to determine its critical value.
- Feng and Qiu (2018, *Technometrics*): Hybrid tests are more powerful.

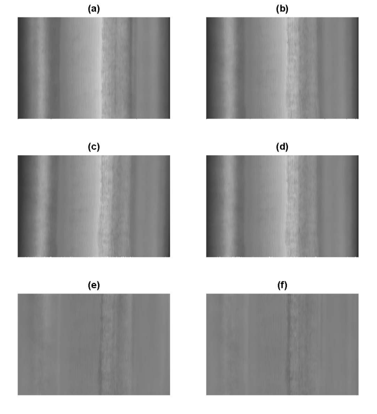
Real Data Application

- Images of a rolling bar at two consecutive times ((a),(b))
- Recovered images by $\widehat{\boldsymbol{\theta}}_E$ and $\widehat{\boldsymbol{\theta}}_C$ ((c),(d))
- Difference images (a)-(c) and (a)-(d) ((e),(f))
- $U_E = 6.95, U_{EC} = 12.31$, and $U_{CC} = 19.36$; $p-$values are $0.335, 0.116$, and 0.059



Real Data Application (con'd)

- Surface images at two times that are far away ((a),(b))
- $U_E = 4.78, U_{EC} = 175.34$, and $U_{CC} = 114.95$
- $p-$values: $0.522, 0.000$, and 0.000.



Current Research in Image Monitoring

- Comparison of multiple images
- Phase I and Phase II SPC

Concluding Remarks

- In this big data era, there are many new applications of SPC
- Data streams are common in big data problems
- SPC could be a major statistical tool for analyzing big data
- New challenges:
    - Big data volume
        - recursive computation
        - dimension reduction (accommodation of spatial structure)
        - parallel computing
    - Time-varying IC distribution
    - multiple shifts/jumps
    - IC versus OC
    - ...
- Traditional SPC methods should be modified properly and new SPC methods should be developed
- We need to learn things in the related areas/disciplines
- SPC has a bright future!