A Statistical Process Monitoring Perspective on “Big Data”

Fadel M. Megahed (Email: fmegahed@auburn.edu)
Assistant Professor of Industrial and Systems Engineering

L. Allison Jones-Farmer (Email: joneall@auburn.edu)
C&E Smith Associate Professor of Analytics and Statistics
What is Big Data?

- Big data is a popular term used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹

What is Big Data?

- Big data is a **popular** term used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹

What is Big Data?

- Big data is a **popular** term used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹

---

What is Big Data? – The 3 Vs

- The term big data refers not only to the size or **volume** of data, but also to the **variety** of data and the **velocity** or speed of data accrual.\(^1\)


What is Big Data? – The 3 Vs

- The term big data refers not only to the size or **volume** of data, but also to the **variety** of data and the **velocity** or speed of data accrual.

What is Big Data? – Beyond the 3 Vs

- The **volume**, **variety** and **velocity** of data distinguish **big data** from **other types of data**.

- Other V’s important to **big data** problems are:
  - **Veracity** ➔ Trustworthiness of the data
  - **Value** ➔ Added value to creating knowledge in a topic

- Both **veracity** and **value** relate to **data quality**

---

The Fraudulent tweet from the hacked AP account

An instant drop in the DOW Jones industrial average
Some Example Applications for Big Data

- 10,000 payment card transactions/sec are made in the world.

- Wal-Mart handles ~
Deming (Out of the Crisis, P.106): “Information, no matter how complete and speedy, is not knowledge. Knowledge has temporal spread. Knowledge comes from theory. Without theory, there is no way to use the information that comes to us on the instant.”

“The Economist”

“More information ought to be useful, but only if companies can interpret it. And workers are already overloaded: 62% of them say that the quality of what they do is hampered because they cannot make sense of the data they already have ...”
How do we Analyze Big Data? – Data Analytics

1. Data cleaning – remove noise and inconsistencies
2. Data integration – combine multiple data sources
3. Data selection
4. Data transformation
5. Data mining – pattern extraction
6. Pattern evaluation
7. Knowledge presentation and visualization

Visualization of the literature on **Big Data** and **SPC** is created based on **Web of Science Citations** using the **Science of Science (Sci²) tool** [https://sci2.cns.iu.edu/user/index.php](https://sci2.cns.iu.edu/user/index.php)
Clustering Analysis of Major Journals cited in the Big Data/SPC Literature – Focusing on the Top 50 Journals

Visualization created as of **Web of Science 8/10/2013 Citation Data** Using the **VOSviewer tool**, see [http://www.vosviewer.com/](http://www.vosviewer.com/)
Big Data and SPC – Machine Learning (ML)

- Unsupervised vs. Supervised Learning

- Components of ML Algorithms:
  - Training Set – Data to which ML algorithm is applied
  - Classifier – The function/model to be discovered
  - Test Set – A set of the data that is reserved for testing

- A training set consists of a set of pairs \((x, y)\), where
  - \(x\) is a vector of values, often called a feature vector.
  - \(y\) is the label, the classification value for \(x\).

- The objective of the ML process is to discover a function \(y = f(x)\) that best predicts the value of \(y\) associated with unseen values of \(x\).
Big Data and SPC – Supervised ML

- Supervised learning algorithms in (Big Data) SPC can be classified into:
  
  • **Phase I approaches** → Supervised learning methods used to identify an in-control baseline sample, e.g.
    - Chinnam (2002, *IJPR*)
    - Hwang et al. (2007, *IIE Trans.*)
  
  • **Phase II methods** → Converting the monitoring problem into a supervised classification problem where future observations are classified as either in- or out-of-control (or several categories), e.g.
    - Li et al. (2006, *IJPR*)
    - Deng et al. (2012, *JQT*)
Common approaches include the following:

- Cluster analysis
- Mixture modeling

Applied mainly in Phase I methods:

- Outlier Detection, e.g.
  - Sullivan (2002, *JQT*)
  - Jobe and Pokajovy (2009, *JQT*)
- In-control sample, e.g.
  - Thissen et al. (2005, *J. of Chemometrics*)

An Example of Unsupervised Learning in a Manuf. Setting

Inferring the optimal placement for gap measurements from inherent **associations** in the data.
Opportunities for ML-based Big Data SPC

- “Much work and process understanding is often required in the transition between Phase I and Phase II analysis” Woodall (2000, *JQT*)

- **Example Applications:**
  - Classification step in a framework for intrusion detection Wenke et al. (1999, *Proc. of IEEE*)
  - “Machine learning is also helping to improve our basic understanding of cancer development and progression” Cruz and Wishart (2006, *Cancer Inform*)
Big Data and SPC – Functional Data Analysis (FDA)

Profile Monitoring
- Functional relationship between a response and explanatory variable(s)
- Received significant attention since 2003.

Image Monitoring
- Extension to profile monitoring
- Can be used for dimensional and aesthetic features
- Not well developed (with exception of chemometrics)

3D Scan Monitoring
- Also referred to as point clouds
- Millions of points depicting surface geometry
- Has potential to be more useful than CMMs
Opportunities for Big Data FDA Surveillance

- Developing methods that can help understand the data which is “… complex (correlated in time and in space in ways that are still not fully understood) and massive (a typical number might be hundreds of thousands of time series for a single subject …)” (Lazar 2008, p. vii).

- Control charting opportunities for images and point cloud data:
  - An understanding of the effect of estimation error on control chart performance
  - Evaluating the inherent assumptions in these methods
Big Data and SPC – Graph Data

- Used to extract relationships between different entities in the data. Massive graphs include:
  - **Social Networks** – Networks of Friends
  - **Cyber-Security and Computer Networks**
  - **Influence Propagation** – Rate of adoption of products/concepts is measured
  - **E-commerce and Web Advertising**

- Additional application domains include **protein-protein** networks in biology
Big Data and SPC – Graph Data Surveillance Ex.

A snapshot of the **annual communication** between members of the **al Qaeda organization** from the Yr 2001 based on **open source data**

Big Data and SPC – Graph Data Surveillance Ex.

A plot of the closeness CUSUM statistic of the al Qaeda organization from the 1994-2004 based on open source data.

For their control chart, they used the reference value \((k=0.5)\) and the control limit \((h=4)\), as these values are widely used in the literature (and corresponds to a false alarm rate every 168 years).

Opportunities in the Surveillance of Big Graphs

- There exists many opportunities in developing control charting methodologies for *(computer)* network data.

- The problem is very *complex*!!
  - State of the network is *continuously changing*
  - State of the network can be represented by *data of different types*

- In the SPC literature, we have a limited class of *multivariate control charting* that can be used to monitor *multivariate data of mixed data types*. 
Big Data and SPC – Multistream Data

- Increased availability of data $\rightarrow$ possibility of monitoring data from many, disparate sources.

- Much of the work in this area has been reviewed by Woodall and Montgomery (2013, *JQT*).
  * Manufacturing applications
  * Biosurveillance applications

- We agree with their recommendation that the adoption of the false discovery rate method can be a successful solution in managing the false alarm rate.
Some Additional Challenges in Big Data SPC

- The data is usually non-numeric and inputted by a large number of users.

- Big data surveillance problem differs somewhat from traditional SPC, as preprocessing (Phase 0?) is needed before any surveillance is done.

- Massive data can lead to not storing historical data and the analysis need to be done on the fly.

- An increased emphasis on diagnosis and not only detection.
Concluding Remarks

- **Big data analytics** is an evolving field with numerous applications, some of which can present solutions to global challenges.

- Much work is needed in how we can **preprocess massive datasets** and how do we evaluate this preprocessing phase.

- Establishing the baseline is much more complex.

- **Big data provides many opportunities for statisticians and quality engineers.**
Acknowledgments

- The authors would like to thank **Prof. William H. Woodall**, for his helpful comments that has significantly improved this paper.

- The first author was partially supported by an Amazon Web Services Higher Education grant.
A Statistical Process Monitoring Perspective on “Big Data”

Fadel M. Megahed (Email: fmegahed@auburn.edu)
Assistant Professor of Industrial and Systems Engineering

L. Allison Jones-Farmer (Email: joneall@auburn.edu)
C&E Smith Associate Professor of Analytics and Statistics