Update on Statistical Learning Applied to Process Monitoring

L. Allison Jones-Farmer

In Collaboration with:
Maria Weese, Waldyn Martinez,
and Fadel Megahed

MIAMI UNIVERSITY

What I want you to know…

1. Monitoring problems have evolved.
2. There are **lots of papers** that try to tackle modern monitoring problems.
3. There is a **HUGE gap between research and practice** in process monitoring.
4. Some **reasons** why this **research/practice gap** exists.
5. How **future work can be structured** to narrow the research/practice gap.

Remember when our control charts looked like this?

CONTROL CHART of the MEAN

SOP: METH 09 | Attribute: CEC (NH_4OAc pH7) | Control Sample: Nepal soil

Dates: 6-7-95, 13-7, 20-7, 27-7, 3-8, 10-8, 17-8, 25-8, 1-9, 8-9, 15-9, 22-9, 29-9, 6-10, 13-10, 20-10, 27-10, 3-11, 10-11, 17-11, 8-12, 15-12, 12-7-96, 19-7, 26-7, 23-2, 1-3, 15-3, 22-3, 5-4-96

Signatures:

cmol$_c$/kg

$\bar{x}$ = 10.7

Results: 10.4, 10.6, 10.8, 10.3, 9.8, 11.3, 11.5, 10.9, 9.5, 10.3, 10.3, 9.7, 12.2, 10.7, 11.3, 11.3, 9.9, 10.0, 9.8, 10.7, 11.5, 11.3, 11.6, 9.8, 12.0, 10.9, 11.8, 12.1, 11.6, 10.3

Warning limits : $\bar{x} \pm 2s$ = 10.7 ± 1.8 = 12.5 and 8.9

Action limits : $\bar{x} \pm 3s$ = 10.7 ± 2.7 = 13.4 and 8.0

Verified by:

Date:

Sign.:

Remarks:

Serial No.: 2

Date: 4-7-95

Data previous chart(s)

n = 10

Σx_i = 106.90

Σx_i^2 = 1150.19

$\bar{x}$ = 10.69

s = 0.91

Data this chart

n = 30

Σx_i = 323.8

Σx_i^2 = 3510.48

$\bar{x}$ = 10.79

s = 0.75

t-test: t = 0.35, df = 38

F-test: F = 1.64, df_1 = 9, df_2 = 29

F_{tab} = 1.59 (If $F<F_{tab}$: OK)

t_{tab} = 2.02 (If $t<t_{tab}$: OK)

Combine $\bar{x}$: yes/no

Combine s: yes/no

Data all charts

n = 40

Σx_i = 430.7

Σx_i^2 = 4660.67

$\bar{x}$ = 10.76

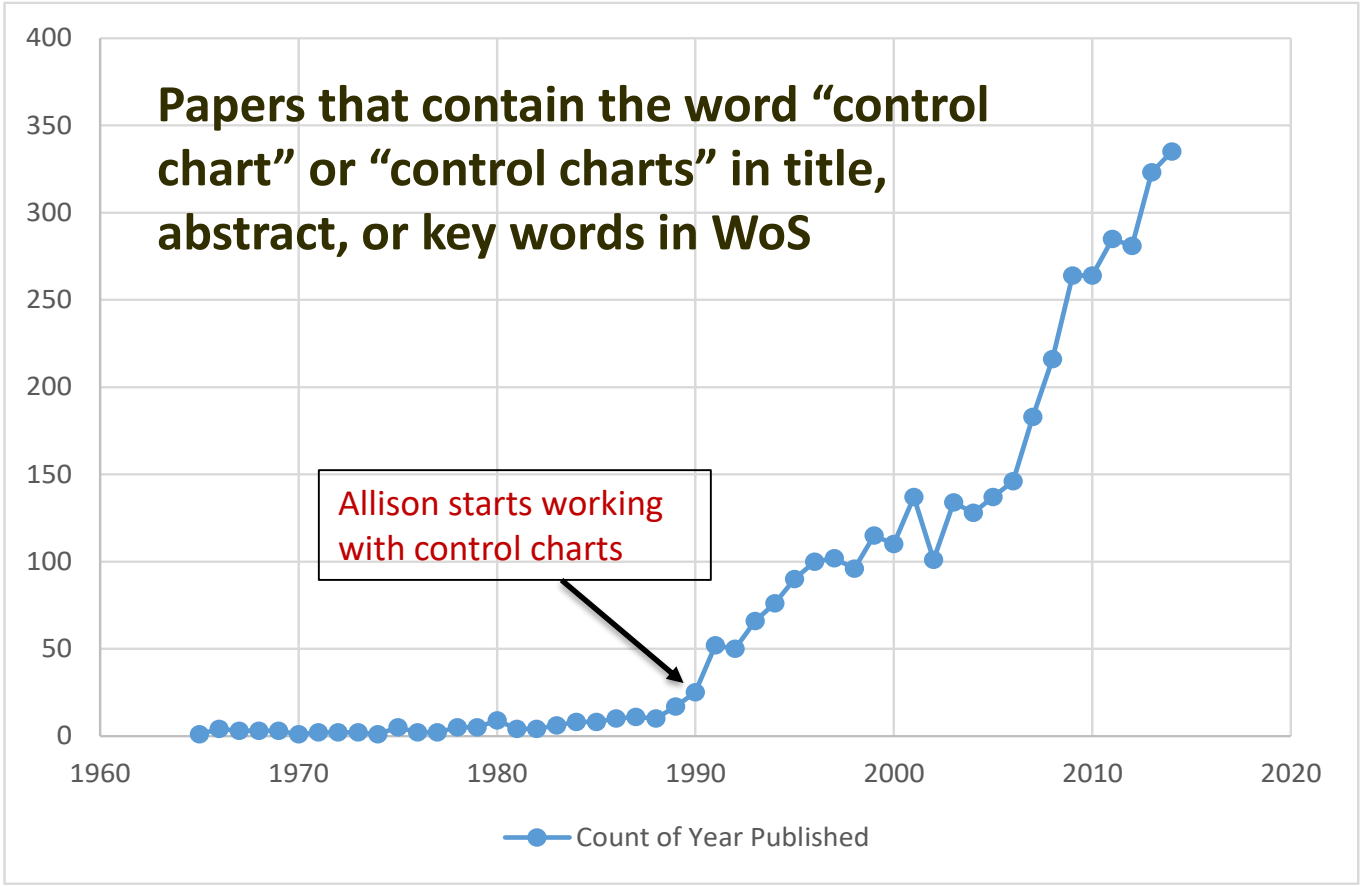
s = 0.77

MIAMI UNIVERSITY

1. Monitoring problems have evolved.

- Social media “hits”
  - High velocity multivariate counts with a nested structure
- Image quality
  - Image quality is predictive of online sales
- Semi-structured text
  - e.g. claims processing
- Wide data—p>>n
  - e.g. image, video, health monitoring
- High velocity data from multiple sensors
  - Human performance data
  - Mixed data

2. There are lots of papers…



MIAMI UNIVERSITY

4106 Number of WoS papers with “Control Chart” or “Control Charts” in title, keywords, or abstract since 1965

64,568 Number of WoS papers with “Artificial Intelligence”, “Machine Learning”, “Statistical Learning, or “Data Mining” in title, keywords, or abstract since 1965

68 Number of WoS papers with [“Control Chart” or “Control Charts”] AND [“Artificial Intelligence”, “Machine Learning”, “Statistical Learning, or “Data Mining”] in title, keywords, or abstract since 1965

2. There are lots of papers…

Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective

MARIA WEESE and WALDYN MARTINEZ
Miami University, Oxford, OH, USA
FADEL M. MEGAHED
Auburn University, Auburn, AL, USA
L. ALLISON JONES-FARMER
Miami University, Oxford, OH, USA

135 Papers Cited

Statistical Learning…

*"**Statistical learning** refers to a vast set of tools for **understanding data**…inspired by the advent of **machine learning** and other disciplines, statistical learning has emerged as **a new subfield in statistics**, focused on supervised and unsupervised modeling and prediction."*

An Introduction to Statistical Learning

Gareth James, Daniel Witten, Trevor Hastie and Robert Tibshirani

Supervised Learning…

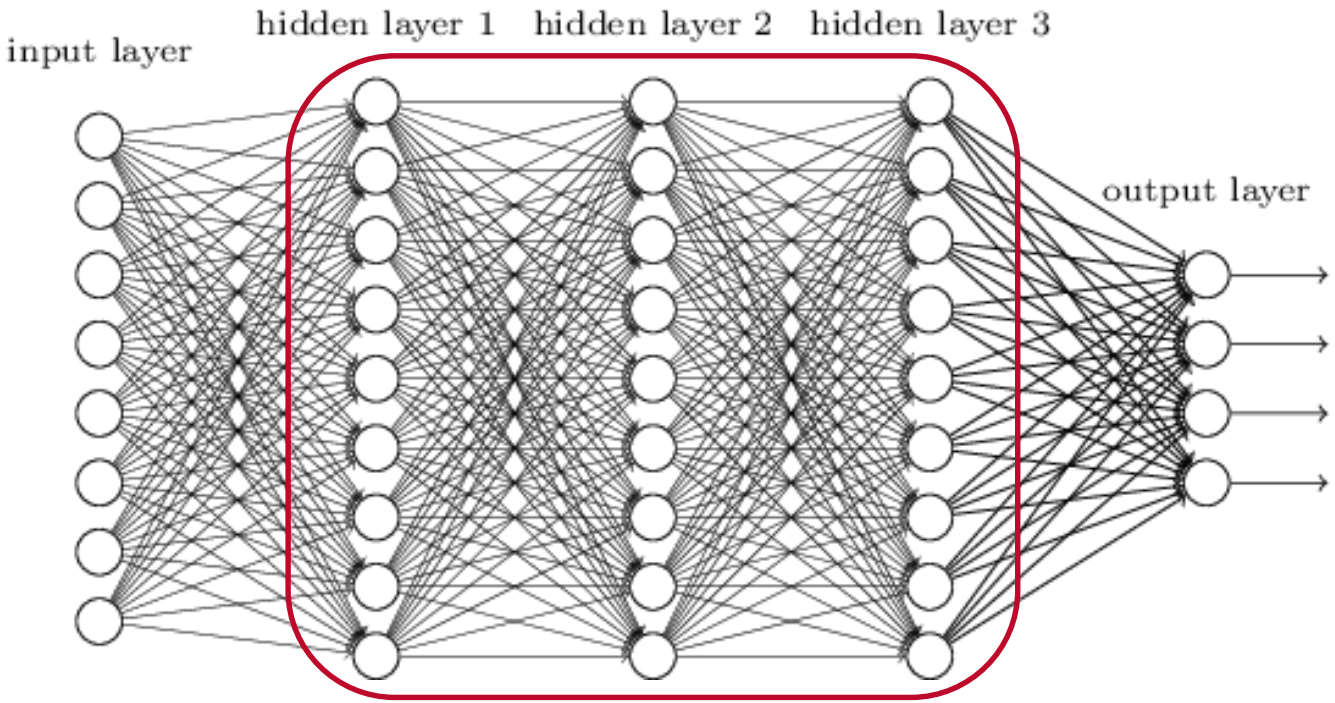
Mapping of **X→Y** using a training sample

Common supervised learning methods include

- Regression analysis
- Artificial Neural Networks (ANN)
- Support Vector Methods (SVM)
- Decision Trees (DT)

Supervised Learning (ANN)

Artificial Neural Networks



Supervised Learning (ANN) for CCPR

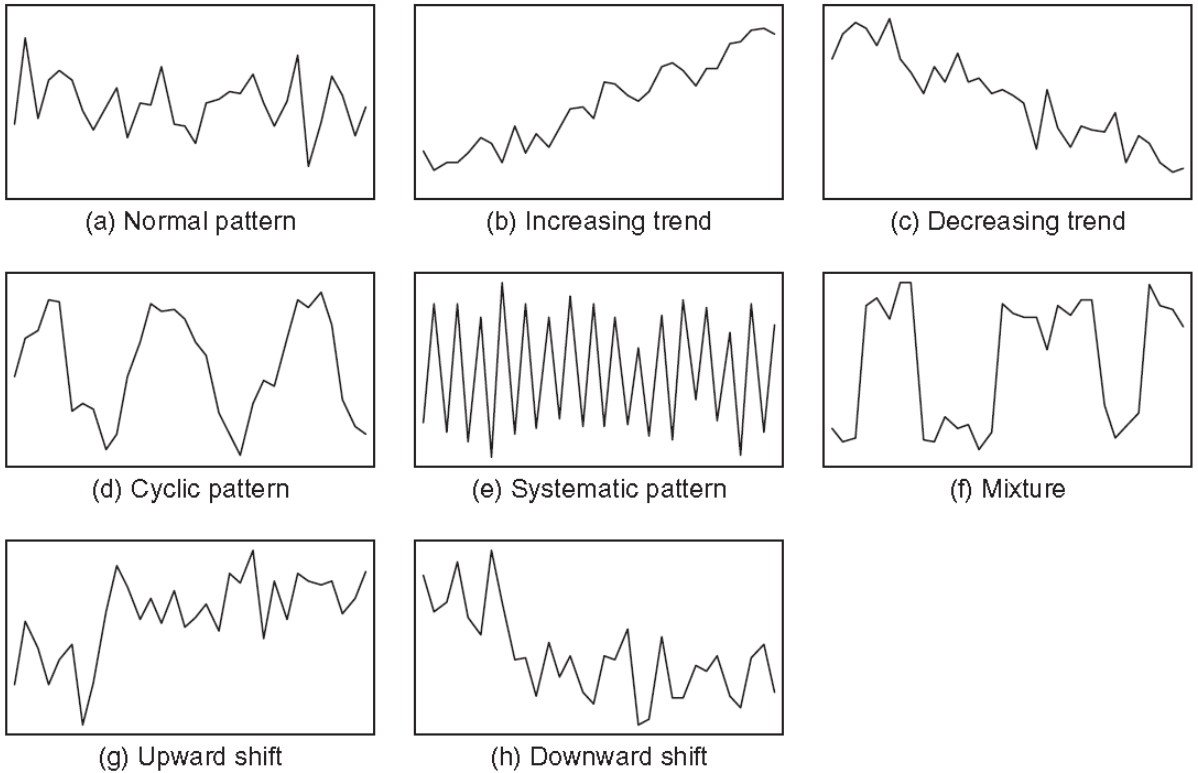


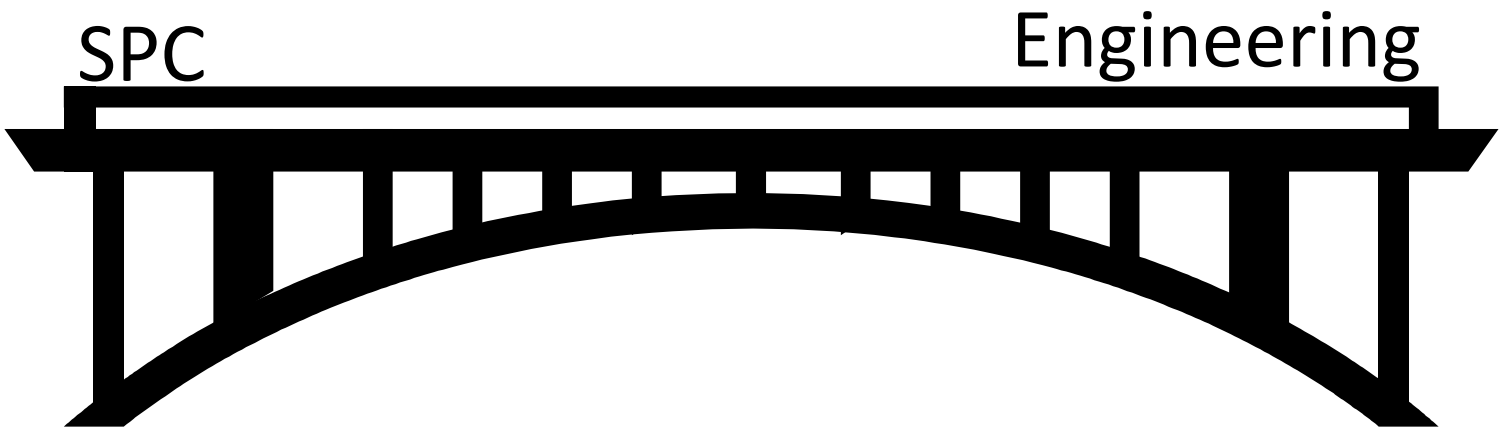
Fig. 1. Examples of typical control chart patterns.

Supervised Learning (ANN) for CCPR

- CCPR started in the 1950's with Western Electric Runs Rules.
- Recent research is dedicated to identifying trends, cyclical patterns, and specific types of process shifts.
- In most research a training sample is "infected" with patterns that should be detected, and a supervised learner (e.g. ANN, SVM) is used to recognize the patterns.
- Monitoring performance depends on how accurately the artificial patterns used to train the learner reflect reality.

Supervised Learning (ANN) for Fault Detection/Diagnosis

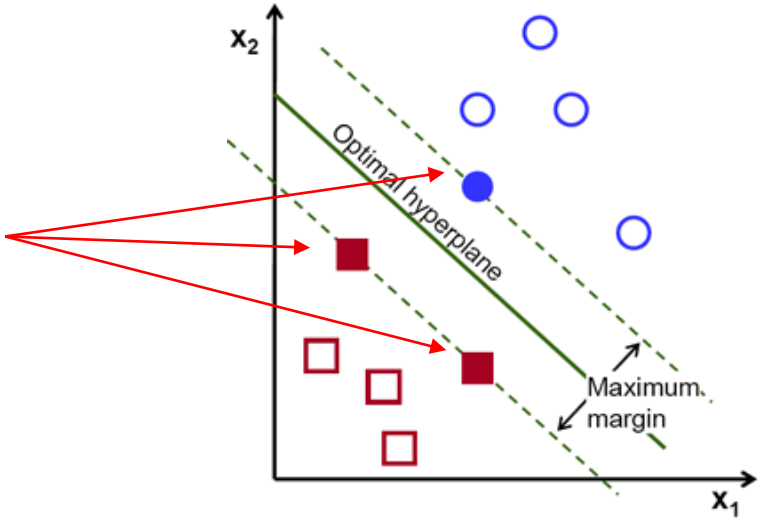
Neural network models are used in the simultaneous *detection* and *diagnosis* of process faults.



Supervised Learning (SVM)

Support Vector Machines (a supervised learning method) separates classes by maximizing the distance between the closest objects of two classes (the margin).

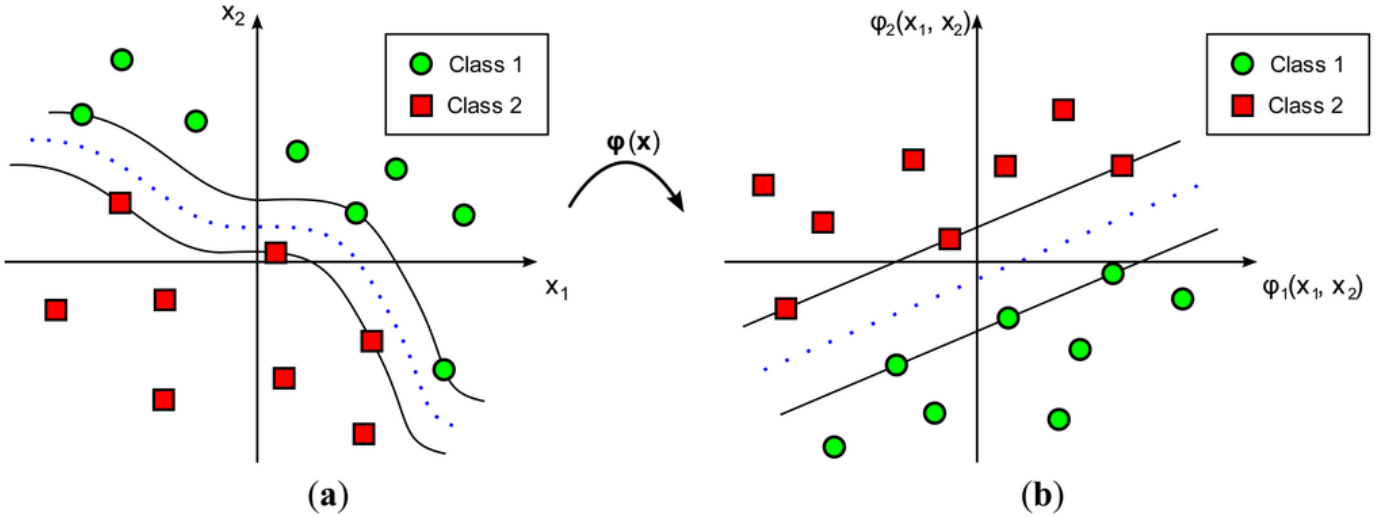
Support Vectors are the critical elements of the training set that would change the position of the separator.



MIAMI UNIVERSITY

Supervised Learning (SVM)

Not all groups can be divided with a linear separator. But a nonlinear transformation with a Kernel function can be employed such that the separator will be linear in the Kernel space.



Supervised Learning (SVM)

- Batch process monitoring (Yao et al. 2014)
- Monitoring the predicted probability of class membership (Chongfuangprinya et al., 2011)
- Fault identification (e.g. Cheng and Cheng, 2008; Mahadevan and Shah, 2009; Chiang et. al. ,2004)
- Fault identification in autocorrelated processes (Chin et. al., 2010)

Unsupervised Learning…

Unsupervised learning describes an area of statistical learning when there is ***no dependent variable***. The ***goal*** of unsupervised learning is to develop a framework or ***to understand a pattern in the structure of the data.***

Examples of unsupervised learning methods include

- principal component analysis (PCA)
- cluster analysis
- some uses of SVM
- latent variable methods such as factor analysis
- mixture modeling

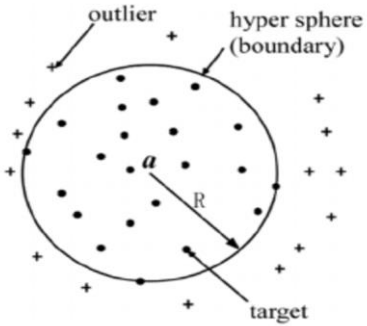
Unsupervised Learning…

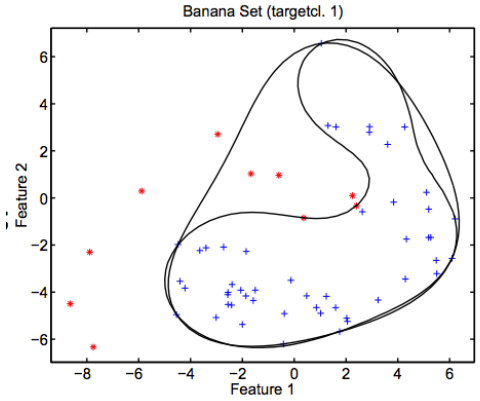
Primary uses in process monitoring

- As tools to prepare data for further analysis (e.g. dimension reduction, clustering)

- Phase 1 (clustering, mixture models, one-class classification)

Unsupervised Learning- One-Class Classification (OCC)





- Reframes the monitoring problem into a classification (in-control, out-of-control)
- Fit a "tight" boundary around the data to define the in-control region
- Motivated by the introduction of the *k*-chart (Sun and Tsung 2003)
  - *k*-chart is based on the Support Vector Data Description (Tax and Duin 1999, 2004)
  - Tax and Duin developed SVDD to retrospectively (and prospectively) **describe** multivariate data of a general (possibly non-normal) form.
  - Tax's research was an extension of Support Vector Machines developed by Vapnik in the 1990's.

MIAMI UNIVERSITY

Ensemble Models

Algorithmically ***combining multiple models*** to improve model performance is commonly referred to as an ***ensemble modeling*** approach.

Ensemble models are often used to combine learning models such as decision trees that are considered to be weak on their own, but quite powerful when multiple trees are combined into a classifier.

Ensemble Models in Process Monitoring

- Li et al. (2006) used random forests (see Breiman 2001) to find the change point and identify the “at fault variables” in a high-dimensional multivariate process.
  - their method outperformed the MEWMA control chart.
- Davila et al. (2014) illustrated the use of an ensemble of decision trees to monitor counts (or rates) of a disease.

Note: Ensembles show promise in the realm of *big data* process monitoring by strengthening/stabilizing methods that are subject to sampling fluctuations.

An Example

The Data..

- Hourly number of hits on ***all*** Wikipedia pages (4.8 Million Pages)

http://dumps.wikimedia.org/other/pagecounts-raw/

- Every hour contains a compressed file of approximately 100MB
- A week of data holds over 16GB of storage.

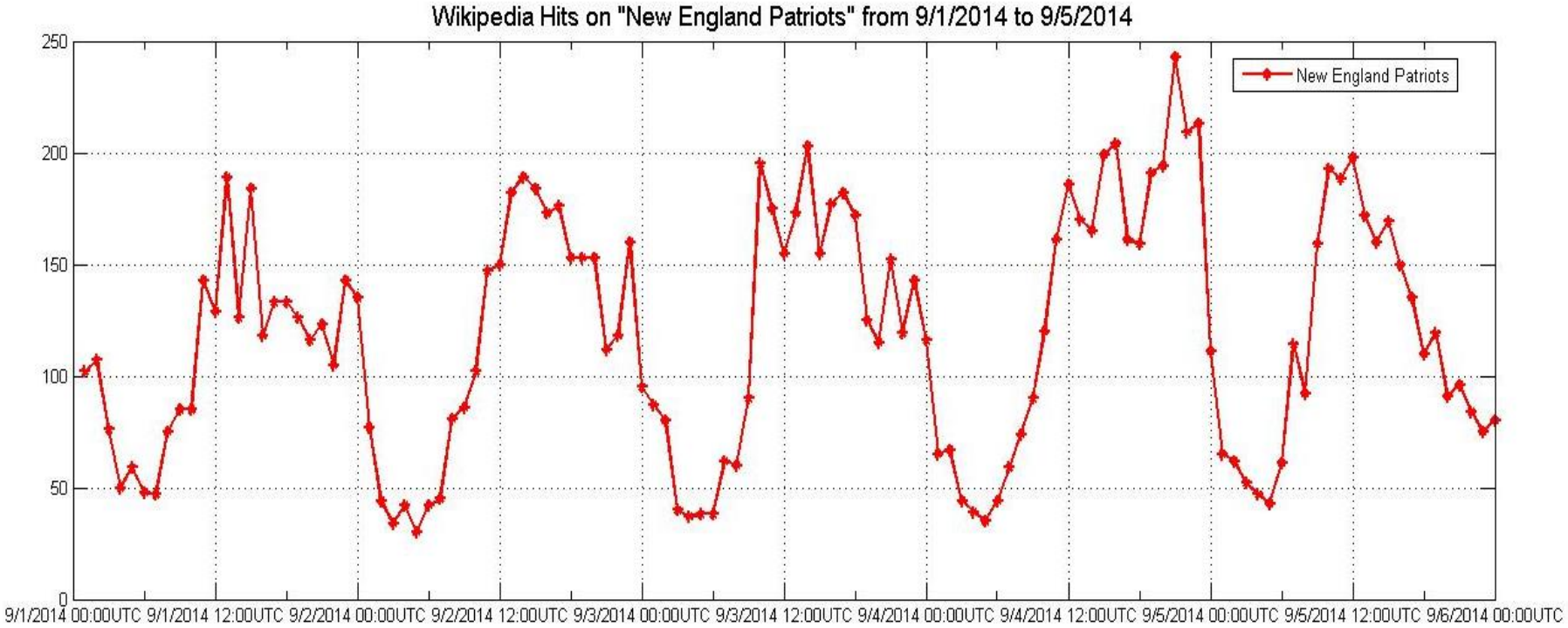
The Data..

- Wikipedia pages in English of all active players, teams, coaches as of 9/15/2014 (*p*=1916 pages)
- The number of page hits per hour from 9/1/2014-9/15/2014, the first two weeks of the NFL season (*n*=354 hours).
- Week 1 = baseline sample for monitoring
- Week 2 = monitored observations.
- Signal = Unusually high number of Wikipedia hits

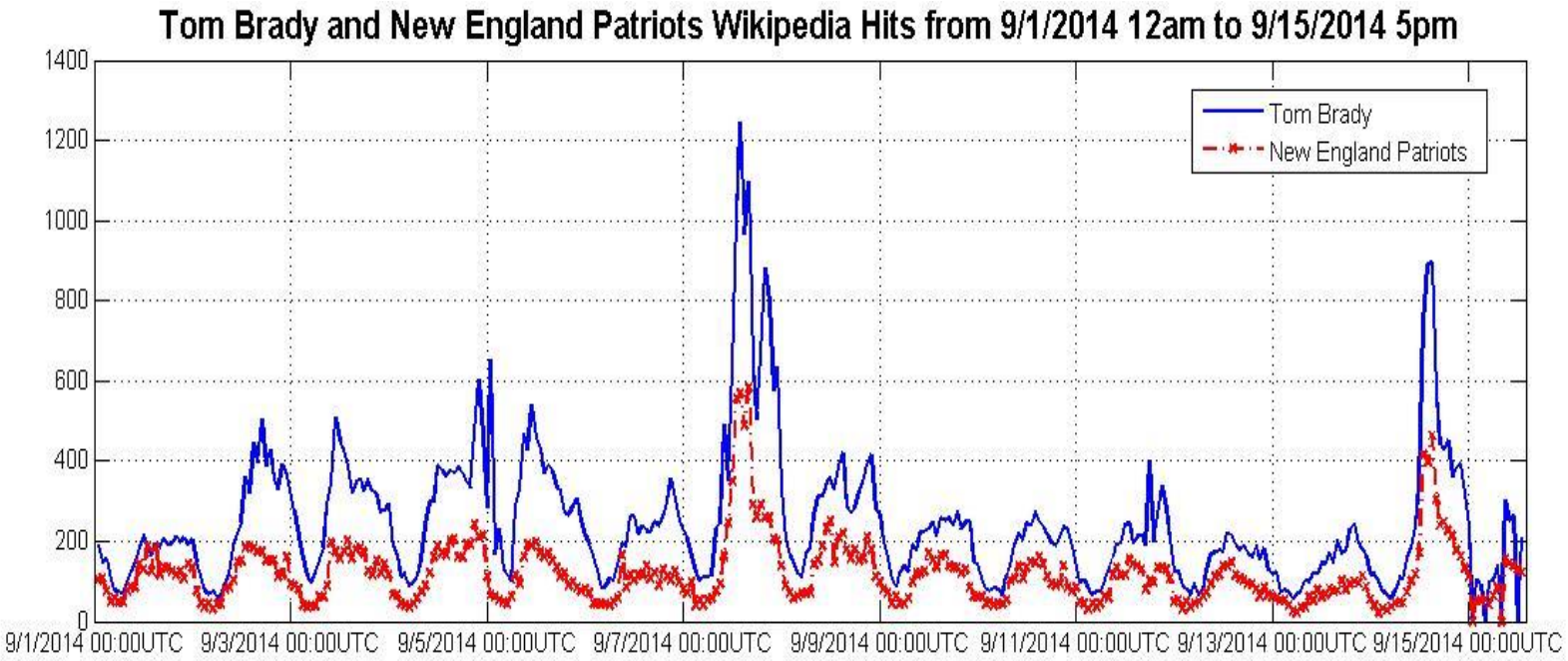
Why this data?

- The data is
  - High dimensional
  - Correlated over time within a stream and among streams
  - Contain many zeros and regular cyclical patterns
  - $p > n$ (1916 > 354)
- Analogous to many modern data stream situations
- There were a number of high profile events that took place during the study period.

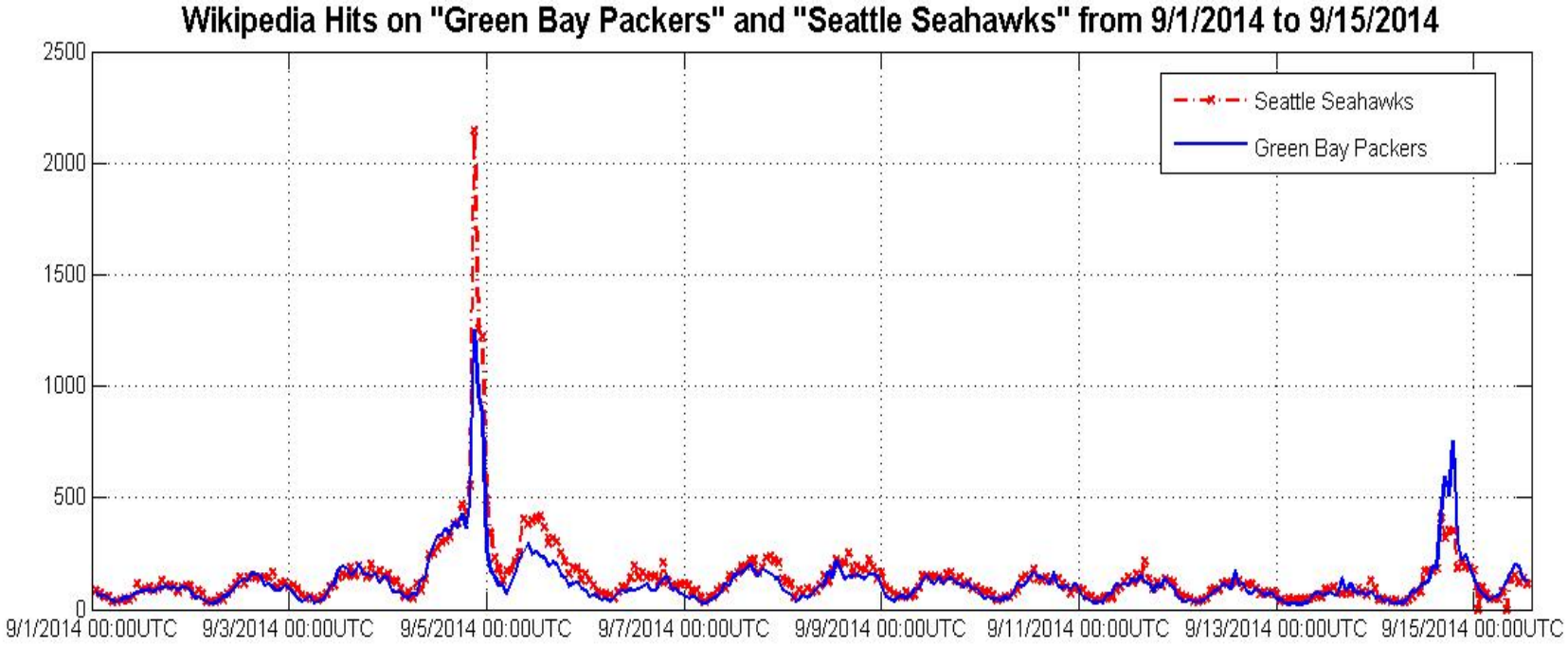
Cyclical Stream…



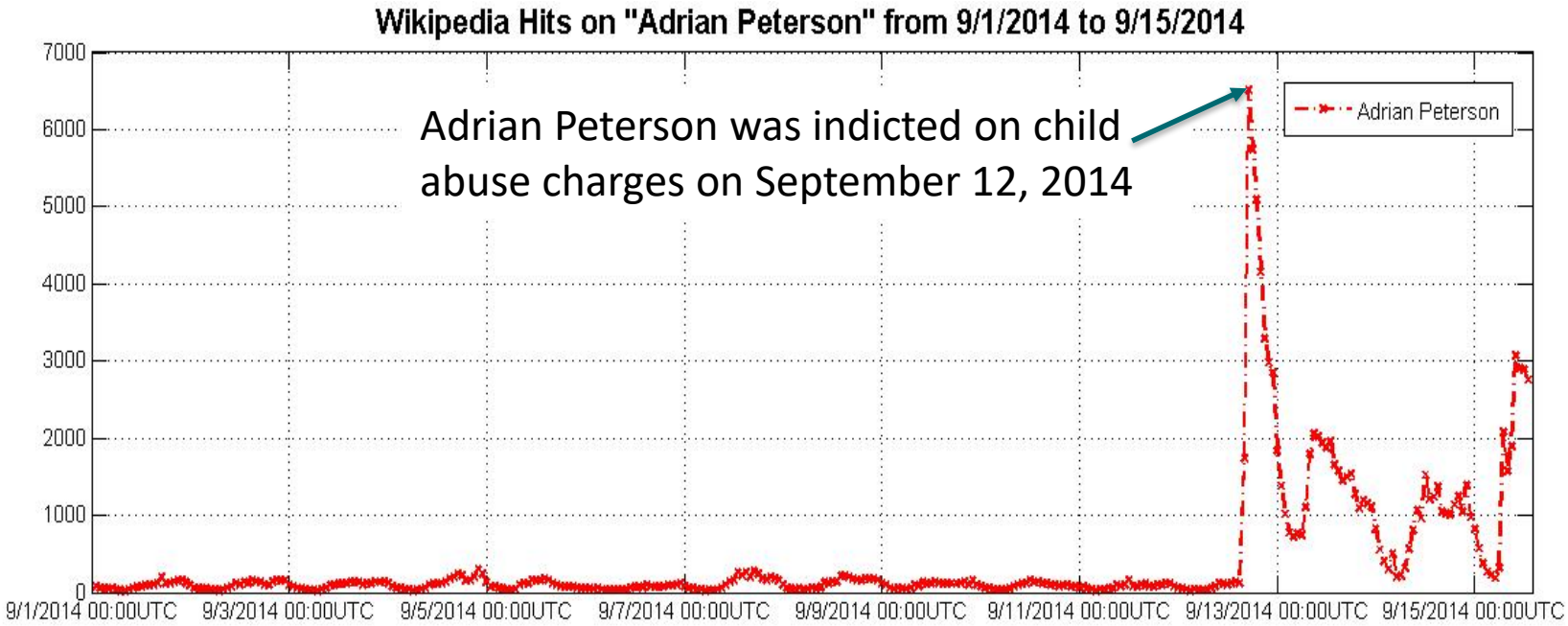
Individuals are correlated with teams...



Teams are correlated with each other…



Do we even need a chart?



Can we use statistical learning methods?

~~Supervised methods~~

- ~~CCRP~~
- ~~Neural Networks~~
- ~~SVM/SVR~~
- ~~Ensemble methods~~

Unsupervised methods:

- Dimension reduction methods
- Cluster based methods
- One-class classification methods

Prewhitening…

- We used a Holt-Winters approach lagged by 24 hours on each entity with seasonal and trend component (see Shmueli and Fienberg 2006; Burkom et al. 2007).

- We then analyzed the multivariate data containing the p=1916 sets of residuals using the K^2 chart.

K^2 (kNN) Chart

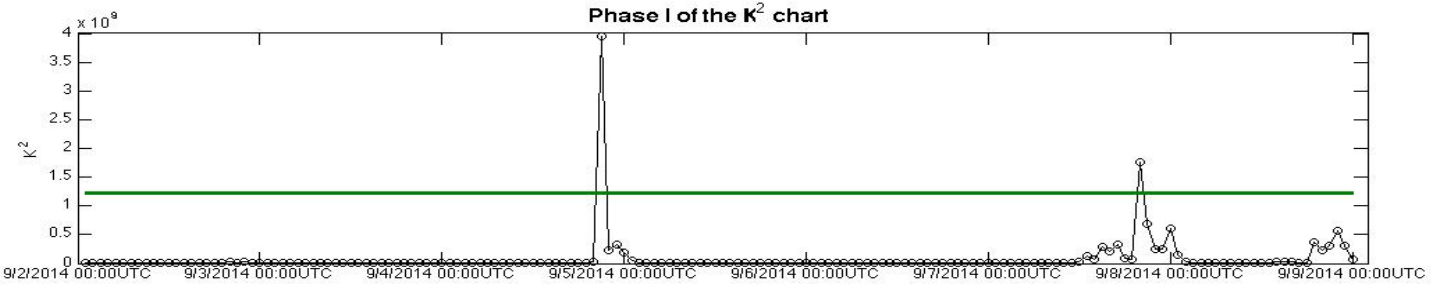
The K^2 chart is constructed as follows:

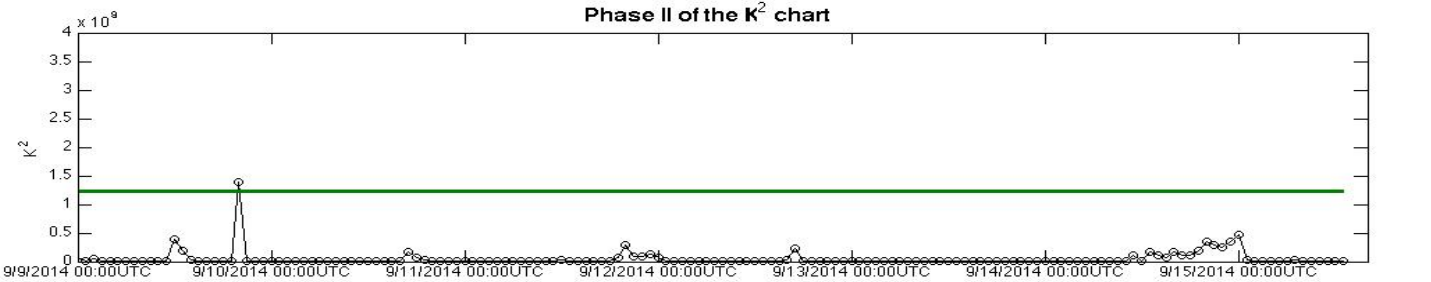
1. Determine *k*, the number of nearest neighbors.
2. Determine the mean of the squared distance between each observation and each of the *k (20)* nearest neighbors in a reference sample.
3. The control limit for the chart is determined by bootstrapping the average squared distances.

Phase I and II

Phase I: First 168 Observations

Phase II: Remaining 162 Observations (the residuals for the 24 lagged observations were not used)





MIAMI UNIVERSITY

3. There is a HUGE gap between research and practice

…there are far too many papers developing yet another charting procedure without considering whether the problem is important and whether the method can actually be used.

Vijay Nair (2008)
Youden Address

Despite the large number of papers on this topic [neural network control charts] we have not seen much practical impact on SPC

Woodall and Montgomery (2014)

3. There is a HUGE gap between research and practice

Missing Pieces…

- Almost no guidance on how to establish a baseline sample.
- Very little published code to implement the methods or replicate the results.
- Little to no guidance on how to implement the methods in practice (e.g. parameter selection).
- Very few papers apply the methods to real (or even realistic) data.

4. Some **reasons** why this **research/practice gap** exists.

Data Modeling Culture…

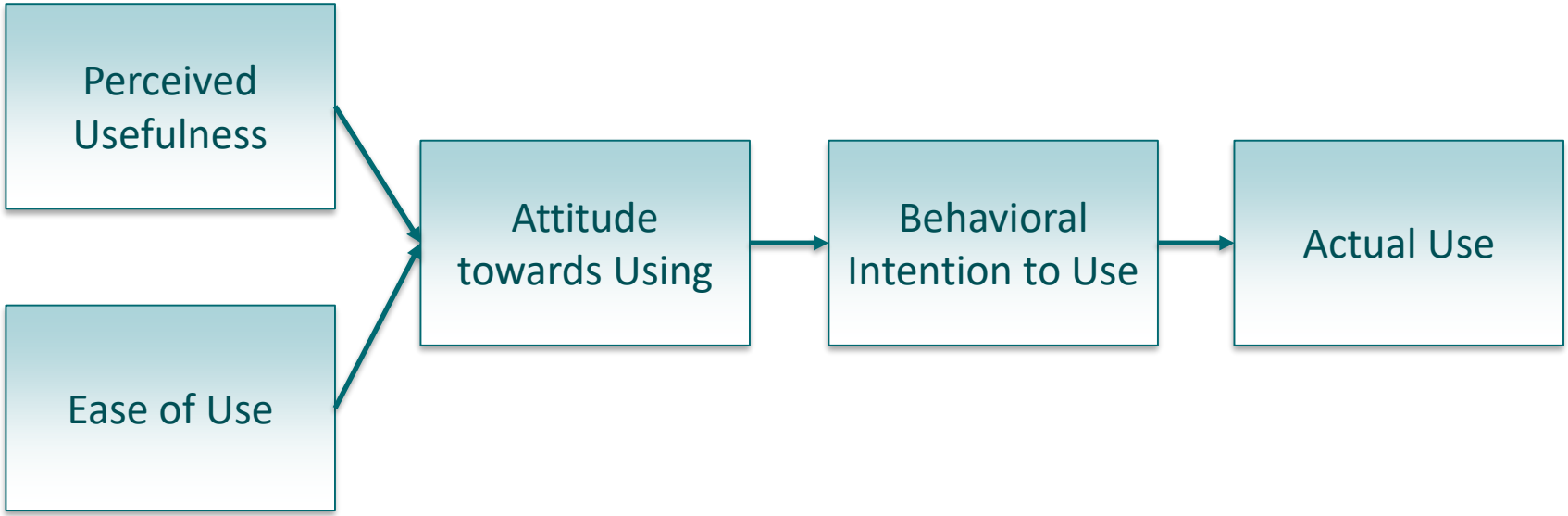
*…the focus in the statistical community on **data models** has:*

- *Led to irrelevant theory and questionable scientific conclusions*
- *Kept statisticians from using more suitable algorithmic models;*
- *Prevented statisticians from working on exciting new problems;*

Leo Breiman (2001)

Breiman, L. (2001). "Statistical Modeling: The Two Cultures", *Statistical Science* 16(3), 199-231.

4. Some **reasons** why this **research/practice gap** exists.



The Technology Acceptance Model

4. Some **reasons** why this **research/practice gap** exists.

- Publishing with “real data” is hard.
- Industry/Academic collaborations are extremely difficult to negotiate.
- Papers that are promote “easy to use” methods may be at odds with the peer review process.

5. How future work can narrow the research/practice gap.

Publish Useful Methods

- that are motivated by real problems (industry/academic collaborations)
- that can be applied to a broad class of problems
- that are benchmarked on real (not just simulated) data (see, e.g. Campos et al. 2016).

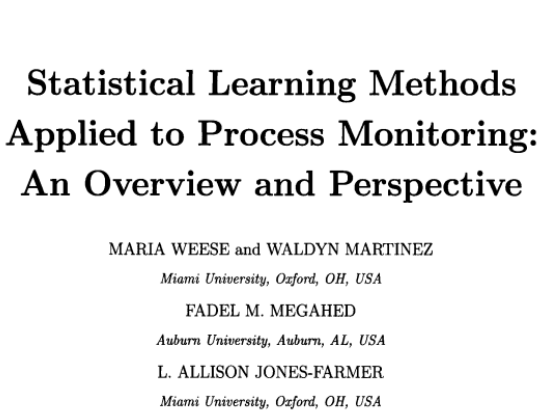
http://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/

5. How future work can narrow the research/practice gap.

Make Methods Easy(ier) to Use

- Publish (documented, executable, open source) code.
- Include specific instructions on how to implement the method.
- Include a detailed example WITH data.
- Make sure your results can be replicated.

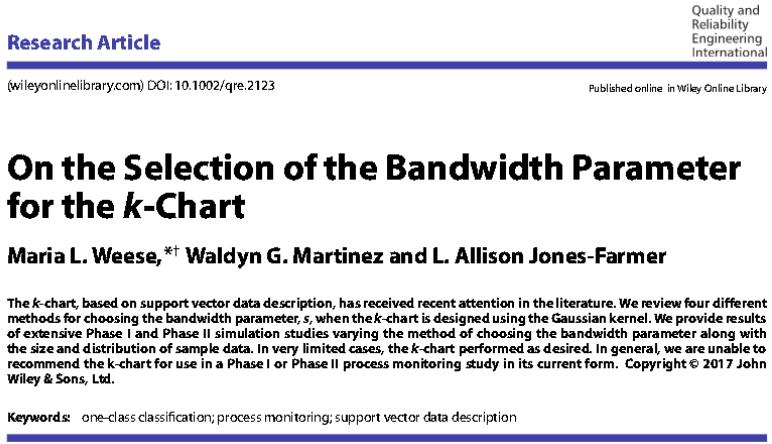
What we learned…



Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective

MARIA WEESE and WALDYN MARTINEZ
Miami University, Oxford, OH, USA
FADEL M. MEGAHED
Auburn University, Auburn, AL, USA
L. ALLISON JONES-FARMER
Miami University, Oxford, OH, USA

JQT, Vol. 48, No. 1, January 2016



Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.2123

Published online in Wiley Online Library

On the Selection of the Bandwidth Parameter for the *k*-Chart

Maria L. Weese,*† Waldyn G. Martinez and L. Allison Jones-Farmer

The *k*-chart, based on support vector data description, has received recent attention in the literature. We review four different methods for choosing the bandwidth parameter, *s*, when the *k*-chart is designed using the Gaussian kernel. We provide results of extensive Phase I and Phase II simulation studies varying the method of choosing the bandwidth parameter along with the size and distribution of sample data. In very limited cases, the *k*-chart performed as desired. In general, we are unable to recommend the k-chart for use in a Phase I or Phase II process monitoring study in its current form. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: one-class classification; process monitoring; support vector data description

Forthcoming in QREI

One-Class Peeling for Outlier Detection in High Dimensions

Submitted for Publication

Opportunities to Innovate..

An encouragement for innovation is to put oneself in a position in which, in order to solve a particular scientific problem, one is forced to learn and discover new things.

Box and Woodall (2012)

References

Box, George EP, and William H. Woodall (2012). "Innovation, quality engineering, and statistics." *Quality Engineering* 24.1, 20-29.

Breiman, Leo (2001). "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical Science* 16.3, 199-231.

Gareth, J., et al (2013). *An introduction to statistical learning*. New York: Springer.

Thissen, U., Swierenga, H., de Weijer, A.P. (2005), "Multivariate Statistical Process Control Using Mixture Modeling," *Journal of Chemometrics*, 19, 23-31.

Sun, R. and Tsung, F. (2003), "A Kernel-Distance-based Multivariate Control Chart using Support Vector Methods," *International Journal of Production Research,* 41(13), 2975-2989.

Ning, X., and Tsung, F. (2013), "Improved design of Kernel-Distance-Based charts using Support Vector Methods", *IIE Transactions*, 45, 464-476

Hachicha, W., and Ghorbel, A. (2012), "A Survey of Control-Chart Pattern-Recognition Literature (1991-2010) Based on a New Conceptual Classification Scheme," *Computers & Industrial Engineering*, 63, 204-222.

Woodall, W. H., and Montgomery, D. C. (2014), "Some Current Directions in the Theory and Application of Statistical Process Monitoring," *Journal of Quality Technology*, 46, 78-94.

Yao, M., Wang, H. and Xu, W. (2014), "Batch Process Monitoring based on Functional Data Analysis and Support Vector Data Description," *Journal of Process Control*. 24, 1083-1097.

References

Issam, B.K. and Mohamed, L. (2008), “Support Vector Regression Based Residual MCUSUM Control Chart for Autocorrelated Process,” *Applied Mathematics and Computation*. 201, 565-574.

Chongfuangprinya, P., Kim, S.B., Park, S.K., and Sukchotrat, T. (2011), “Integration of Support Vector Machines and Control Charts for Multivariate Process Monitoring,” *Journal of Statistical Computation and Simulation*, 81, 1157-1173.

Cheng, C., Chen, P., and Huang, K. (2011), “Estimating the Shift Size in the Process Mean with Support Vector Regression and Neural Networks,” *Expert Systems with Applications* 38, 10624-10630.

Li, F., Runger, G.C., and Tuv, E. (2006), “Supervised learning for change-point detection,” *International Journal of Production Research*. 15, 2853-2868.

Breiman, L. (2001a), “Random Forests,” *Machine Learning*, 45, 5-32.

Dávila, S., Runger, G., and Tuv, E. (2014), “Public Health Surveillance with Ensemble-Based Supervised Learning,” *IIE Transactions*, 46, 770-789.