MIAMI UNIVERSITY

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

An Overview of Statistical Learning in Process Monitoring

Maria Weese
Waldyn Martinez
Fadel M. Megahed
L. Allison Jones-Farmer

amazon web services

FARMER SCHOOL OF BUSINESS
Information Systems & Analytics

Example

Imagine you have a new job….

One way to understand public interest that is generated by the popular press is to consider monitoring social media (e.g. Twitter, Facebook, etc.) and/or data from web searches (e.g. Google, Yahoo, Wikipedia).

Let’s assume your new job is to monitor social media for the National Football League (NFL) to better help the organization react to certain “public relations events”, or at least separate out the typical “chatter” from an “event”.

Data

You gather data from http://dumps.wikimedia.org/other/pagecounts-raw/ which contains the hourly number of hits on all Wikipedia pages (note there are over 2 million English Language pages, about 4.8 million total pages).

Every hour contains a compressed file of approximately 100MB for the number of hits on millions of Wikipedia pages. A week of data holds over 16GB of storage.

Data

You develop a dictionary of the NFL team names, coaches, managers and all currently active players, as of 09/15/2014.

To reduce the computational burden you consider only Wikipedia hits on those pages listed in the NFL dictionary in the English language, which reduces our dimension to p = 1916 pages, including all active players, coaches, teams and managers.

Data

Your data set considers page hits per hour over a two week period beginning on 9/1/2014 and was specifically chosen to include the first two weeks of the 2014 season for a total of n=354 samples.

The first week is used to establish a baseline for monitoring, and the second week constitutes the monitored observations.

A signal to a potential out-of-control event is defined as an unusually high number of Wikipedia hits on a particular team, coach, manager, or player. And recall, the dimension is p=1916.

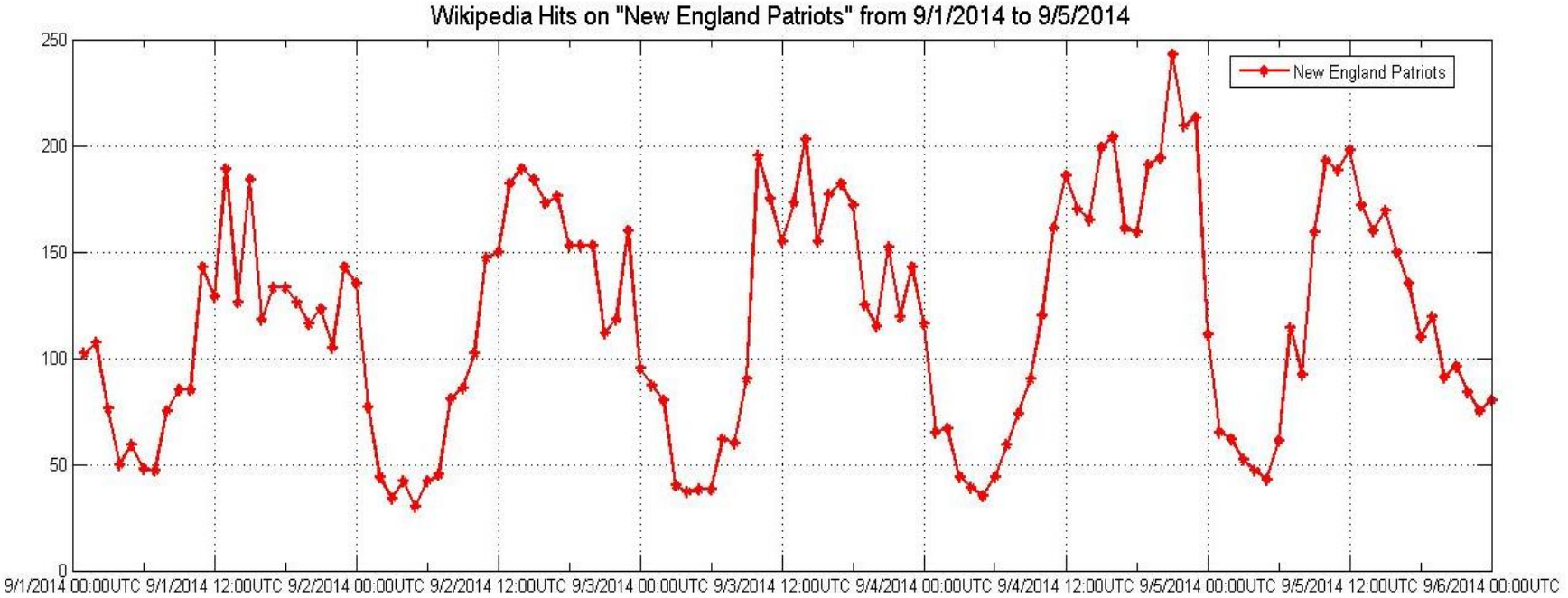
Slightly contrived, but….

(1) it represents modern data streams that would be considered *big data*

(2) the data are counts (not multivariate normal), contain many zero values, have a nested correlation structure, and contain evidence of some high-profile events that spurred intense public interest.

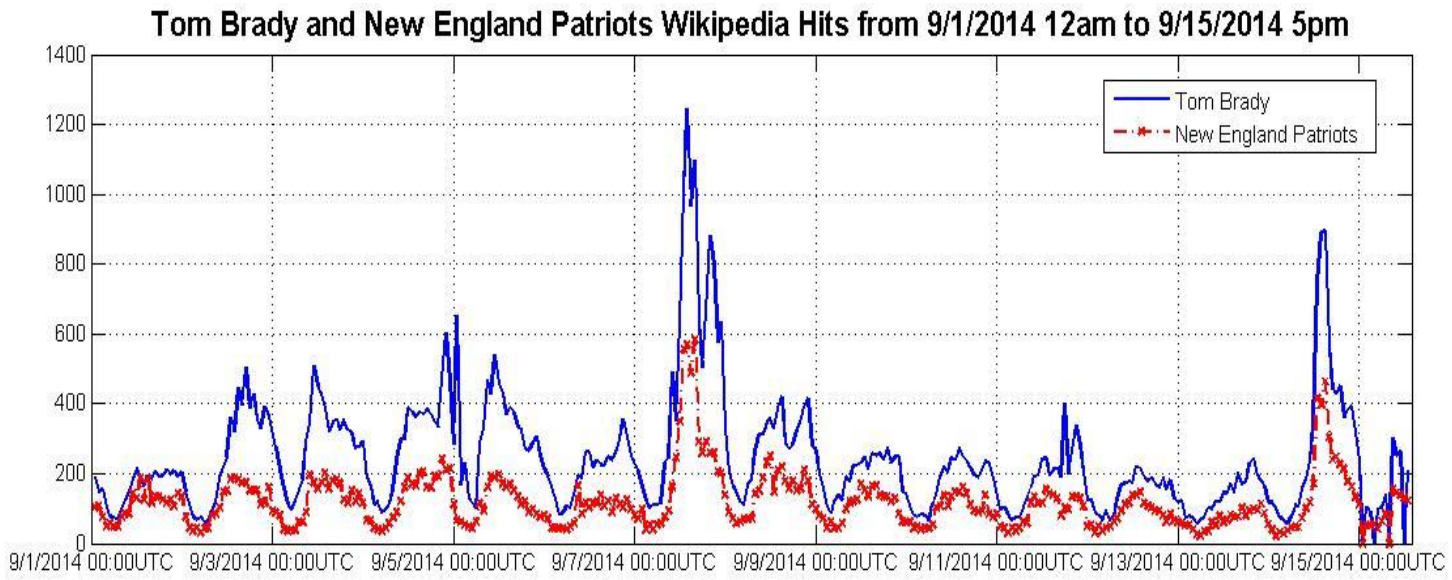
Challenges

The observed counts are cyclical, with the number of Wikipedia search hits declining late at night, and peaking at specific times, especially on game days during the season.



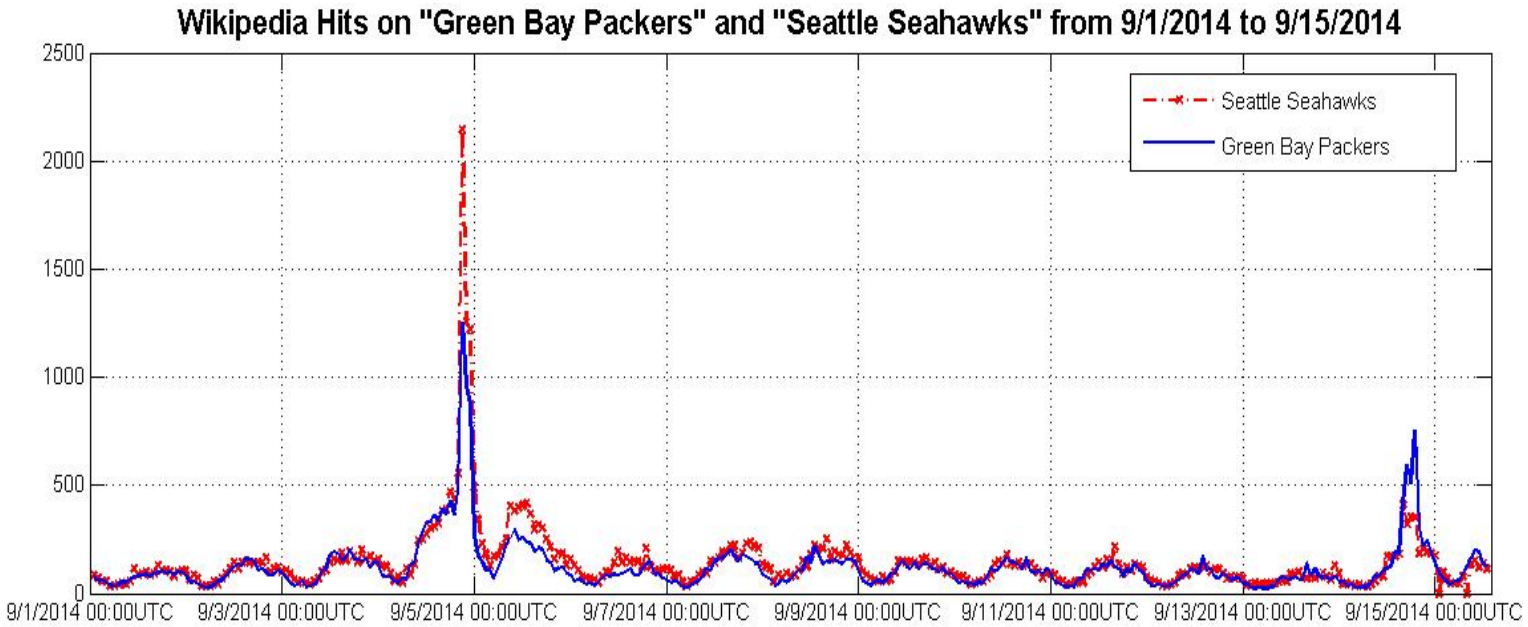
Challenges

Further, the data observed on each of the p=1916 pages are zero-inflated counts that are autocorrelated, and also cross-correlated due to the natural nesting structure of players and coaches within teams.



Challenges

Further correlations exist between teams, especially those paired as opponents during a game.



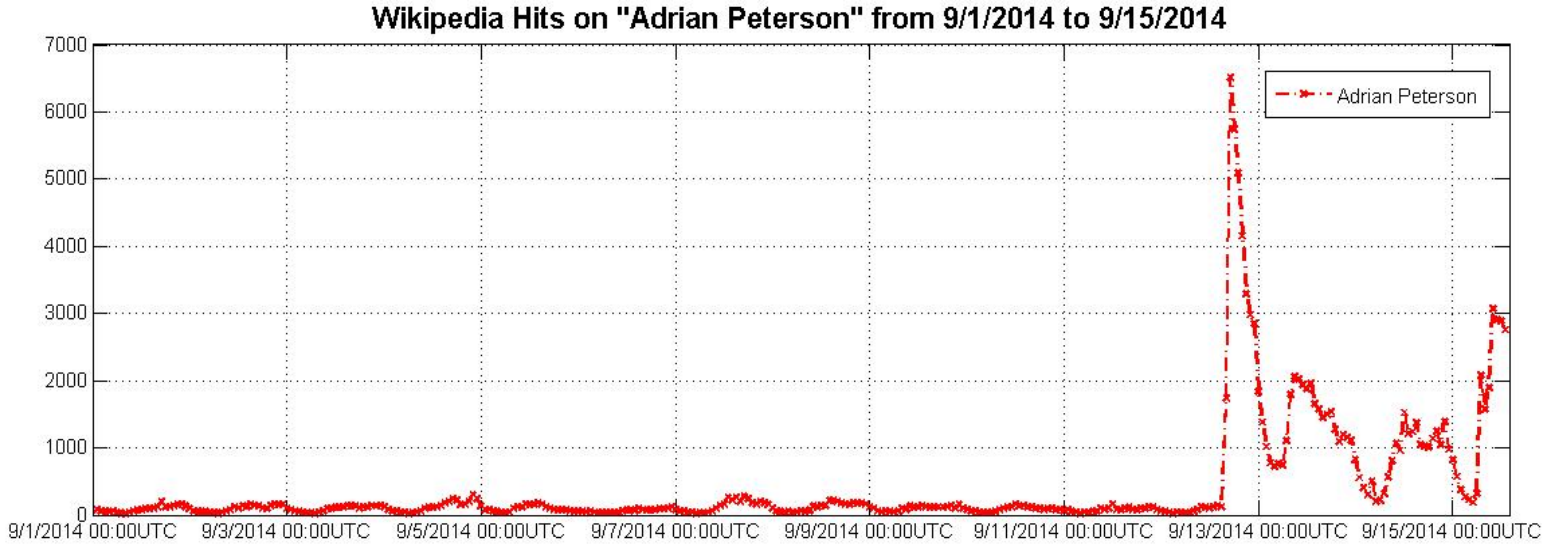
Challenges

In our example, the number of variables, *p* (team names, coaches, managers, and players) is larger than the number of observations (hourly hits) *n*.

All of these data characteristics are expected for this type of internet traffic data, but constitute a challenge in the application of statistical monitoring.

Why do you need a “statistical” chart anyway?

We readily admit that certain events that “go viral” may not need a statistical method to detect a process anomaly. For example, consider the player, Adrian Peterson, who was indicted on a child abuse charges on September 12, 2014.



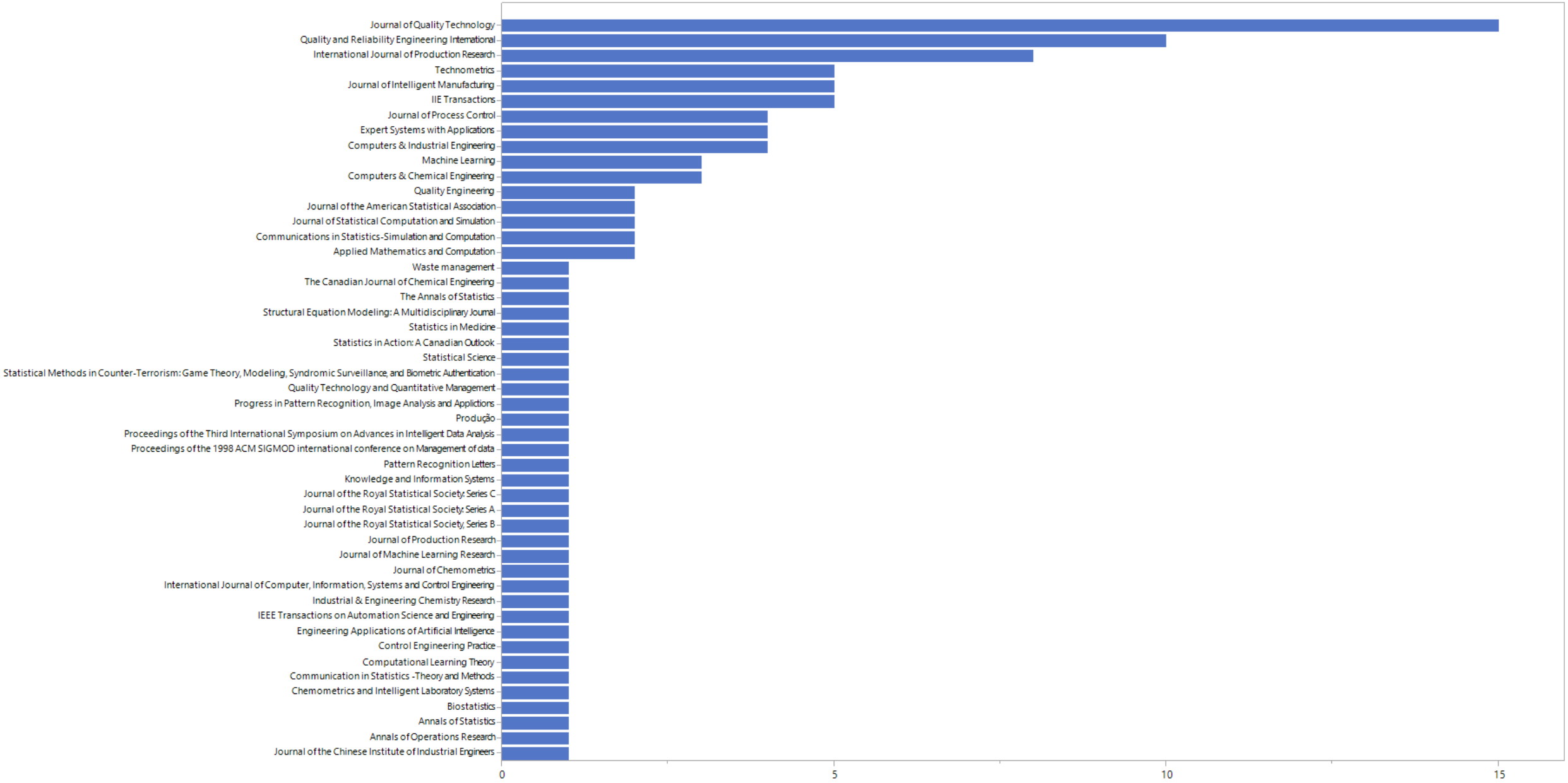
Now What?

The first step in defining a monitoring scheme for this data set is to define an appropriate method.

We need a method that can be applied to high-dimensional zero-inflated counts that are both auto-correlated and cross-correlated with a natural nesting structure.

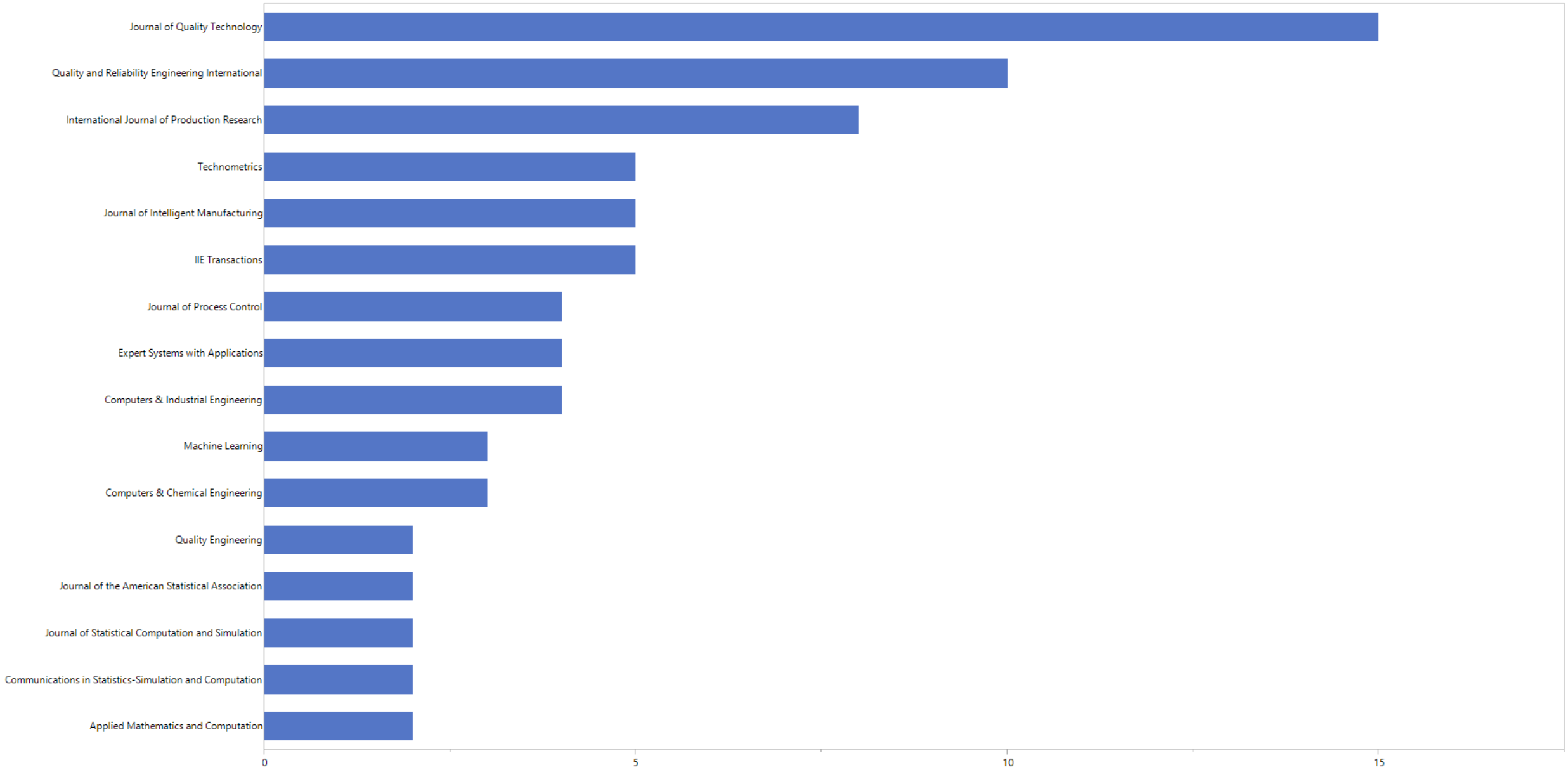
So we turned to the literature to help us solve this problem.

About the Literature



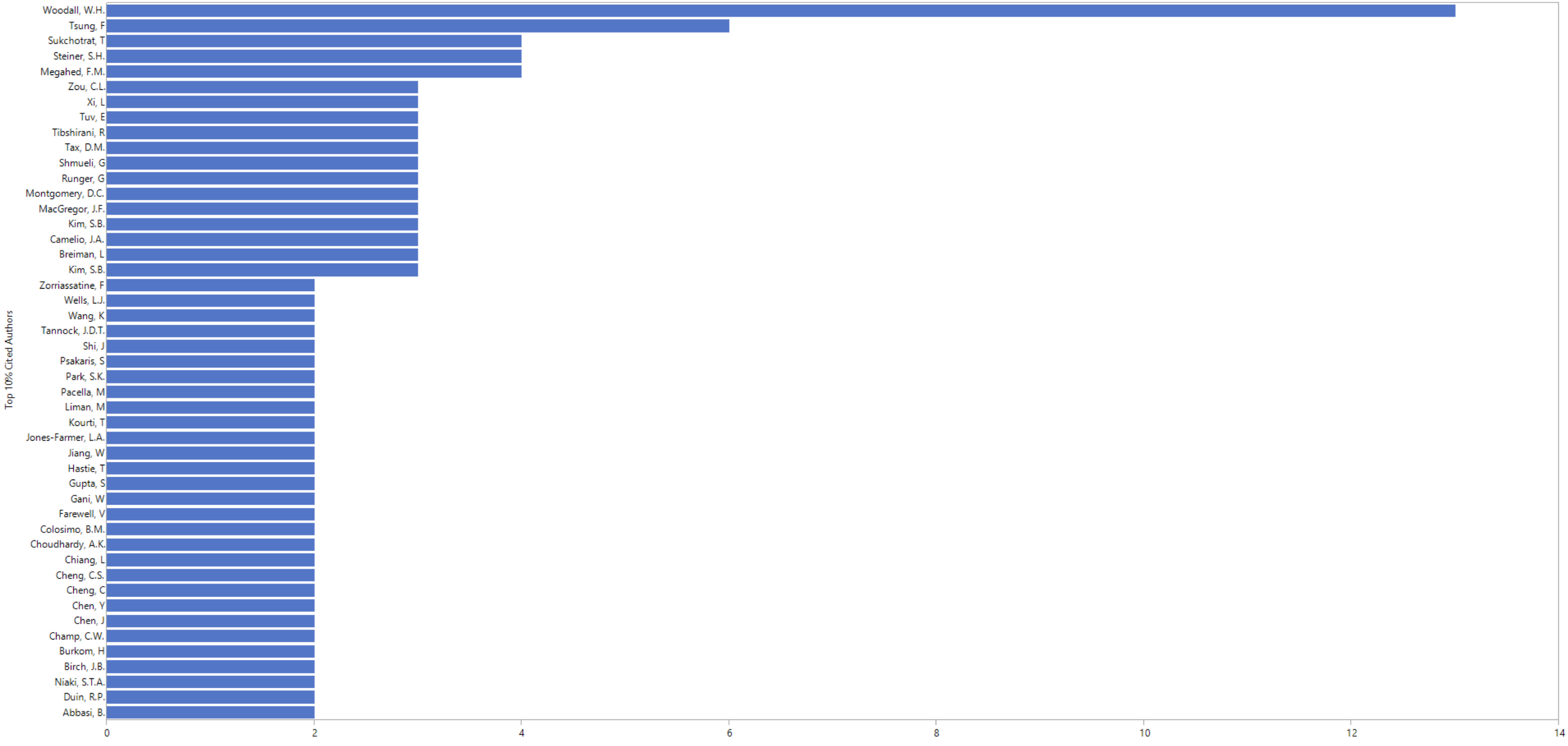
MIAMI UNIVERSITY

FARMER SCHOOL OF BUSINESS
Information Systems & Analytics



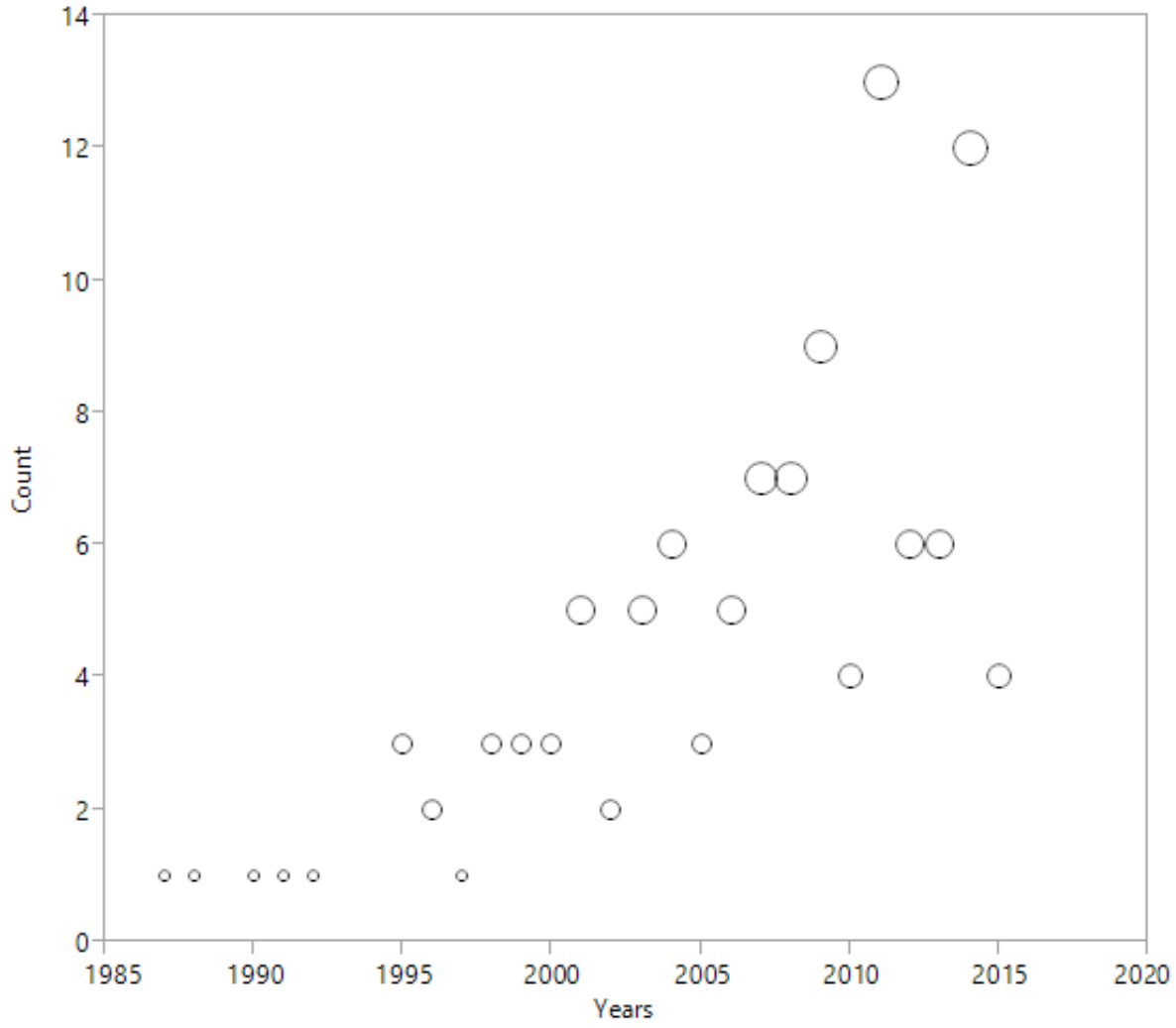
MIAMI UNIVERSITY

FARMER SCHOOL OF BUSINESS
Information Systems & Analytics



MIAMI UNIVERSITY

FARMER SCHOOL OF BUSINESS
Information Systems & Analytics


14
12
10
8
6
4
2
0
Count
1985
1990
1995
2000
2005
2010
2015
2020
Years

How big is *big data*?

The Internet in Real-Time

How Quickly Data is Generated

Like 59k Share 59k Tweet 23.6K +1 8.8k Share 5.1K Share 30.1K

Click here to watch as these internet giants accumulate wealth in real-time.

Twitter
1320 Accounts Created
684000 Tweets

YouTube
277680 Video Hours Watched
240 Video Hours Uploaded

LinkedIn
21840 User Searches

Skype
2777760 Minutes Used

Instagram
2222280 Likes
120000 Comments
83280 Uploaded

Google+
694440 +1s

WordPress
2760 Blog Posts

Google
552960 Searches
$192240 Ad Revenue

Reddit
120 Posts
1560 Comments
25440 Votes

Tumblr
55560 Posts

Pinterest
28560 Pins

Amazon
6120 Items Purchased
$283080 Money Spent

Foursquare
4200 Check-Ins

Yelp
60 Reviews

Gmail
408333360 Emails Sent

Dropbox
1388880 Files Saved

Snapchat
694440 Stories Viewed
972240 Messages Sent

Apple App Store
76080 App Downloads

Android Market
148320 App Downloads

Facebook
6263520 Likes
6597120 Posts
720 GB of Data

WhatsApp
1440 Accounts Created
26388840 Messages Sent

Netflix
46320 Hours Watched

Pandora
122280 Hours Streamed

By the way, in the 120 seconds you've been on this page, approximately 2708880 GB of data was transferred over

Statistical Learning Methods

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

Supervised learning refers to inferring a mapping between a set of input variables $\boldsymbol{x} = \{x_1, x_2, \ldots, x_p\}$ and an output variable *y*, given a training sample $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ of data pairs generated according to an unknown distribution P_{xy} with density $p(\boldsymbol{x}, y)$.

Common examples include Logistic regression, Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Decision Trees (DT)

Combining Models to Improve Performance

Several supervised learning models can be combined to obtain better predictive performance than one could obtain from fitting a single model. Algorithmically combining multiple models together to improve model performance is commonly referred to 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.

Unsupervised Learning

Unsupervised learning describes an area of statistical learning that does not benefit from the availability of an outcome variable. The goal of unsupervised learning is to develop a framework or understand a pattern in the structure of the input variables $\{x_1, x_2, \ldots, x_p\}$.

Examples of unsupervised learning methods include cluster analysis, principal components analysis (PCA), latent variable methods, and mixture modeling.

Unsupervised Learning

Unsupervised Approaches to Process Monitoring

We break this down into three applications of unsupervised methods:

1. Dimension Reduction Methods
2. Clustering
3. One-Class Classification

Dimension Reduction Methods

John Sall in his 2013 Plenary Session at the 57th Annual Fall Technical conference "Big Statistics is Different" referred to *wide data* as opposed to *tall data.* In other words, *n* is smaller than *p* (as is the case in our NFL example).

We can take two approaches to reducing the dimension in our NFL example:

~~(1) select a subset of the variables or~~

~~(2) project the original set of variables into a lower dimensional subspace.~~

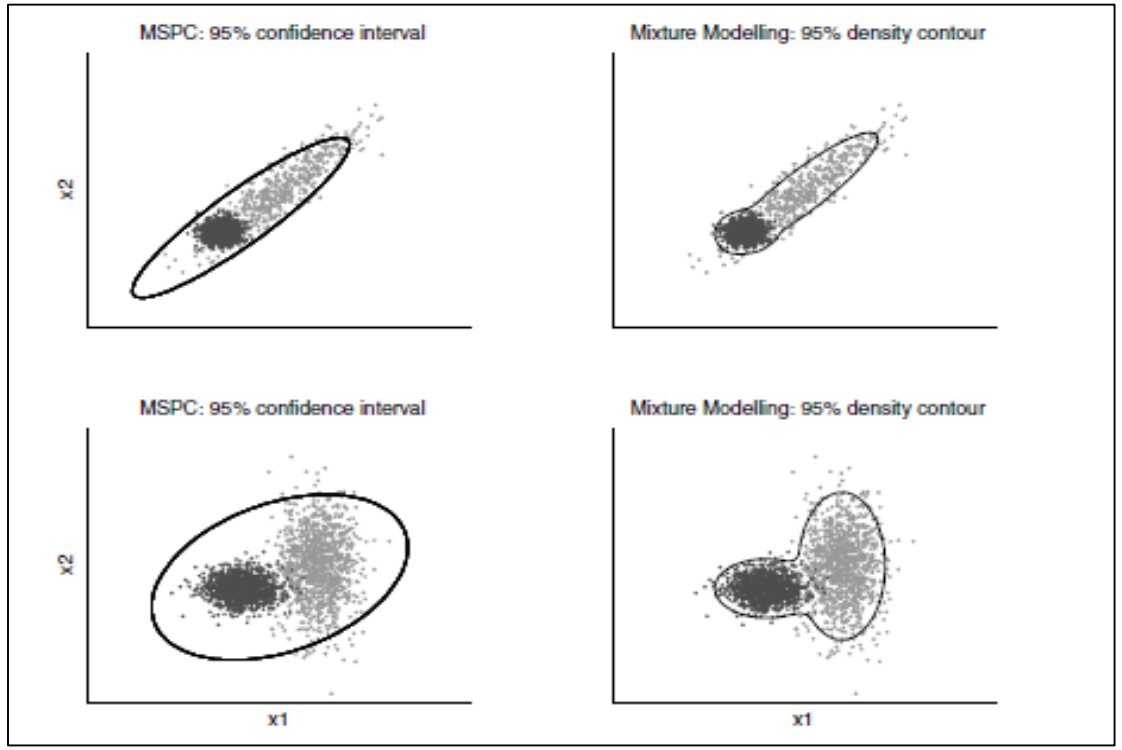
Clustering Methods

Clustering methods may be based on an a priori model, such as mixture modeling, or algorithmic methods like *k*-means or hierarchical agglomerative clustering.

There are numerous clustering methods, and many of these have been applied to control charting. For example, in the chemical industry, model-based clustering methods such as mixture modeling have been used to define the in-control state of the process, or normal operating conditions.

Mixture Modeling

Mixture modeling is a method in which the distribution of independent variables is considered a mixture of two or more distributions that may differ in location, scale or correlation structure.



Thissen et al. (2005)

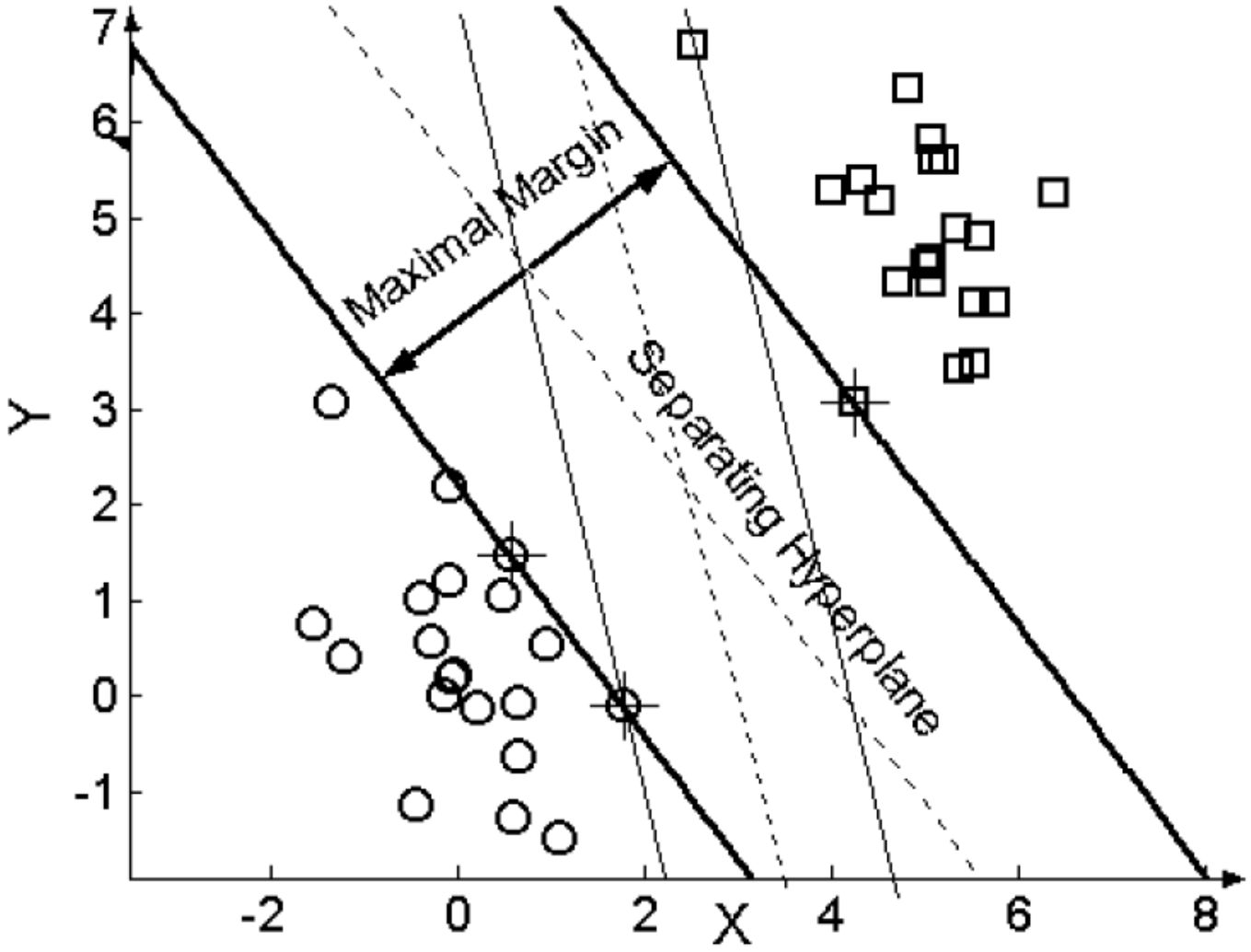
One-Class Classification

Similar to the model-based clustering methods, the one-class classification (OCC) approaches to process monitoring reframe the monitoring problem into a classification problem that classifies observations as either in- or out of control.

All OCC approaches attempt to fit a boundary to define the in-control region

This stream of research began with the introduction of the *k*-chart (Sun and Tsung 2003). The *k*-chart is a control chart based on the support vector data description (Tax and Duin 1999, 2004) and designed for non-normal process data.

Support Vector Machines (supervised)



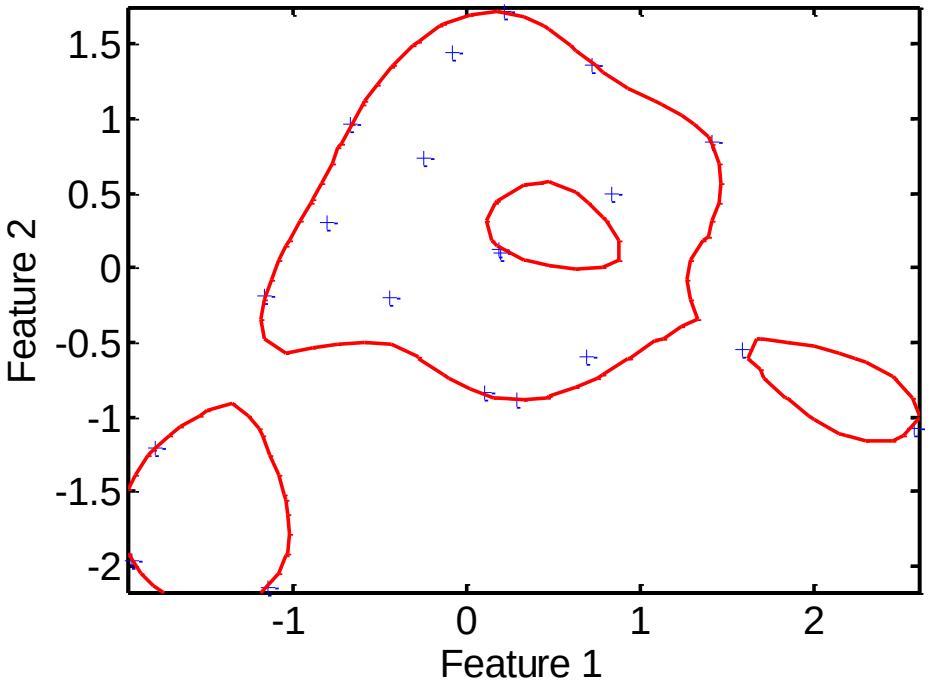
Support Vector Data Description (SVDD)

Support vector methods use hyperplanes to divide multidimensional data into groups or classes. "*The main idea of SVDD is to envelop the samples within a high-dimensional space with the volume as small as possible*" (Sun and Tsung 2003, p. 2979).

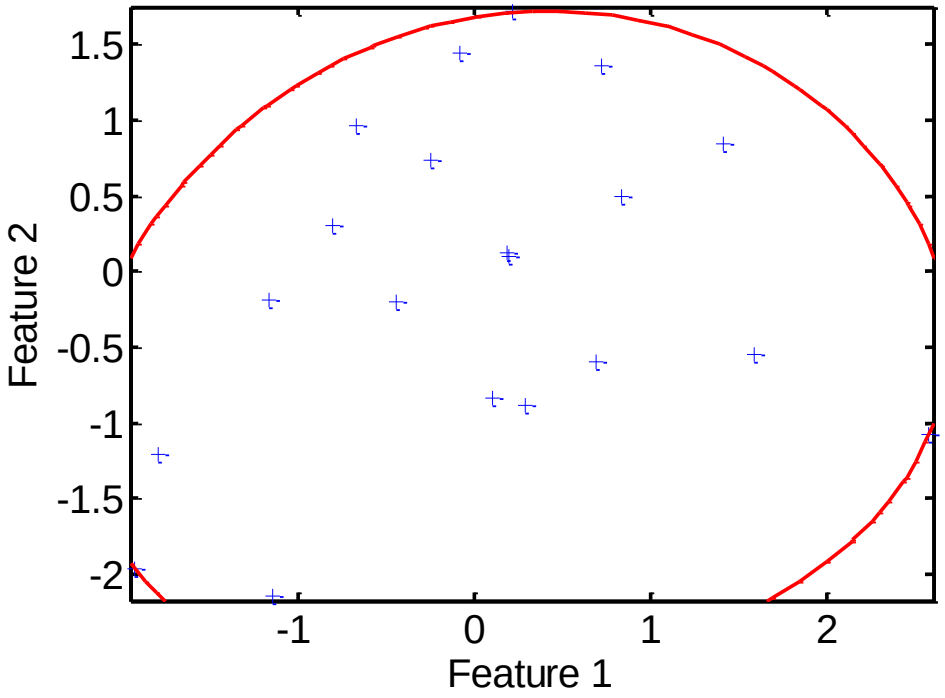
The shape of the boundary determined using the SVDD method differs based on the different types of kernel functions used. Kernel functions used in support vector machines allow the user to implement a nonlinear boundary to separate the two classes of data.

The most commonly used kernel function is the Gaussian Kernel.

Support Vector Data Description (SVDD)



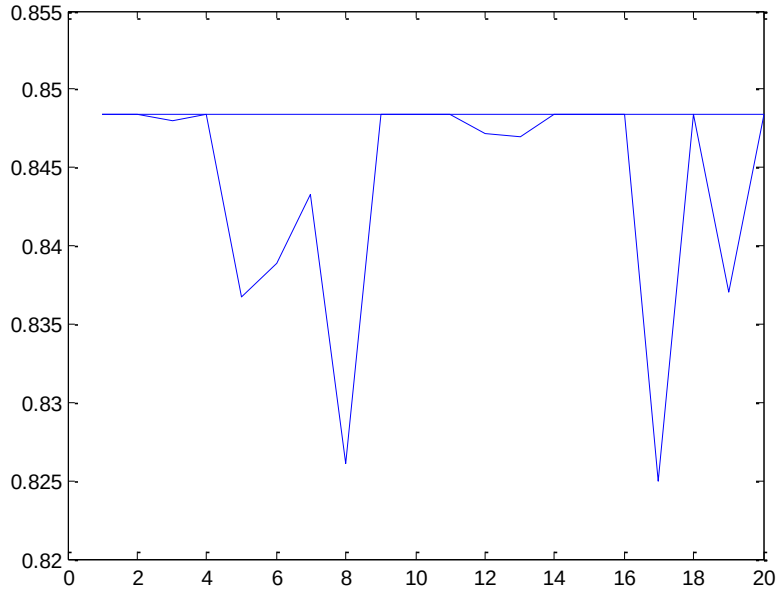
n=20, N(0, 1) random variables
σ (window width)=1

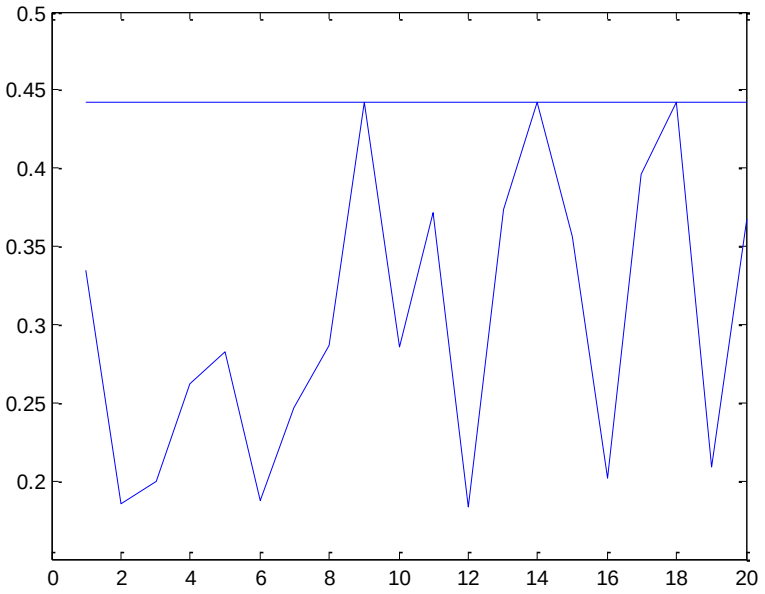


n=20, N(0, 1) random variables
σ (window width)=4

K-chart

The distance between an observation to the kernel center is the monitoring statistic used in the *k*-chart. The boundary for the data represents the control limit which distinguishes in-control observations from potential out-of-control conditions.





K-chart

Optimal determination of this boundary for use as a control limit is a topic open for future research and is discussed in Ning and Tsung (2013).

Control charts based on SVDD have the advantage of only depending on the support vectors; therefore, they are applicable to large amounts of process data and variables.

Chart Name	Motivation	Method	Control Limit	Kernel Method	Assumes i.i.d.	Misclassification Error
k-chart (Sun and Tsung 2003)	Eliminate underlying distributional assumptions for Multivariate SPC	SVDD	Kernel Radius	Gaussian radial-based function (RBF)	Yes	Gives information on changes in error rates with different number of support vectors.
Robust k-chart (Kumar et al. 2006)	Reduce the sensitivity to outliers in the reference data and to reduce potential over-fitting issues that can arise with the k-chart	RSVM	Kernel Radius	Compared 4 methods and showed Gaussian RBF performed best	Yes	Gives information on changes in error rates with different number of support vectors.
rk-chart (Camci et al. 2008)	Eliminate underlying distributional assumptions, requires only in-control data, offers methods for selecting limits based on Type I and Type II errors	SVDD and Support Vector Representation and Discrimination Machine (SVRDM)	Kernel Radius	Gaussian RBF	Yes	Employed an iterative procedure based on the data and number of support vectors to balance Type I/Type II errors.
KNNDD/KNN/K^2 chart (Sukchotrat et al. 2009)	Computationally more efficient than k-chart methods	kNNDD	Bootstrap percentile procedure	None	Shown to have better performance than a T^2 chart when data are non i.i.d. (Kim et al. 2010)	Employed a bootstrap procedure based on process data to select a control limit with a specified misclassification rate.
K-means chart (Kang and Kim 2011)	More quickly detects small shifts in the mean vector than the k-chart.	KMDD	Specified distance from the individual cluster center	None	Yes	Used an iterative procedure to determine control limit with specified misclassification error rate for differing number of clusters.
AK-chart (Liu and Wang 2014)	More quickly detects small shifts in the mean vector than the k-chart.	SVDD	Genetic algorithm to establish action and warning regions based on Variable Sampling Intervals (VSI).	Gaussian RBF	Yes	Employed a genetic algorithm based on process data and number of support vectors to determine a control limit with a specified misclassification rate.

Considerations for These Methods

- In quality control applications, it is generally important to maintain the time ordering of the process observations. Because many clustering/OCC methods do not preserve the time order of the data, it may be difficult to interpret signals to potential out-of-control events.
- There remain many opportunities for research in this area, particularly in Phase I applications.
- Future research should investigate the use of clustering, classification, and mixture modeling approaches for the Phase I analysis of data with multiple data types.

Supervised Learning

Control Chart Pattern Recognition (CCPR)

CCPR has its origins in the early days of SPC, starting with the Western Electric run rules in 1956.

Like the classical approach to CCPR, the recent research on this topic is dedicated to identifying and classifying out of control patterns such as trends, cyclical patterns and specific types of process shifts.

In the modern CCPR literature, a supervised learner (ANN or SVM typically) is trained to recognize specific types of process changes.

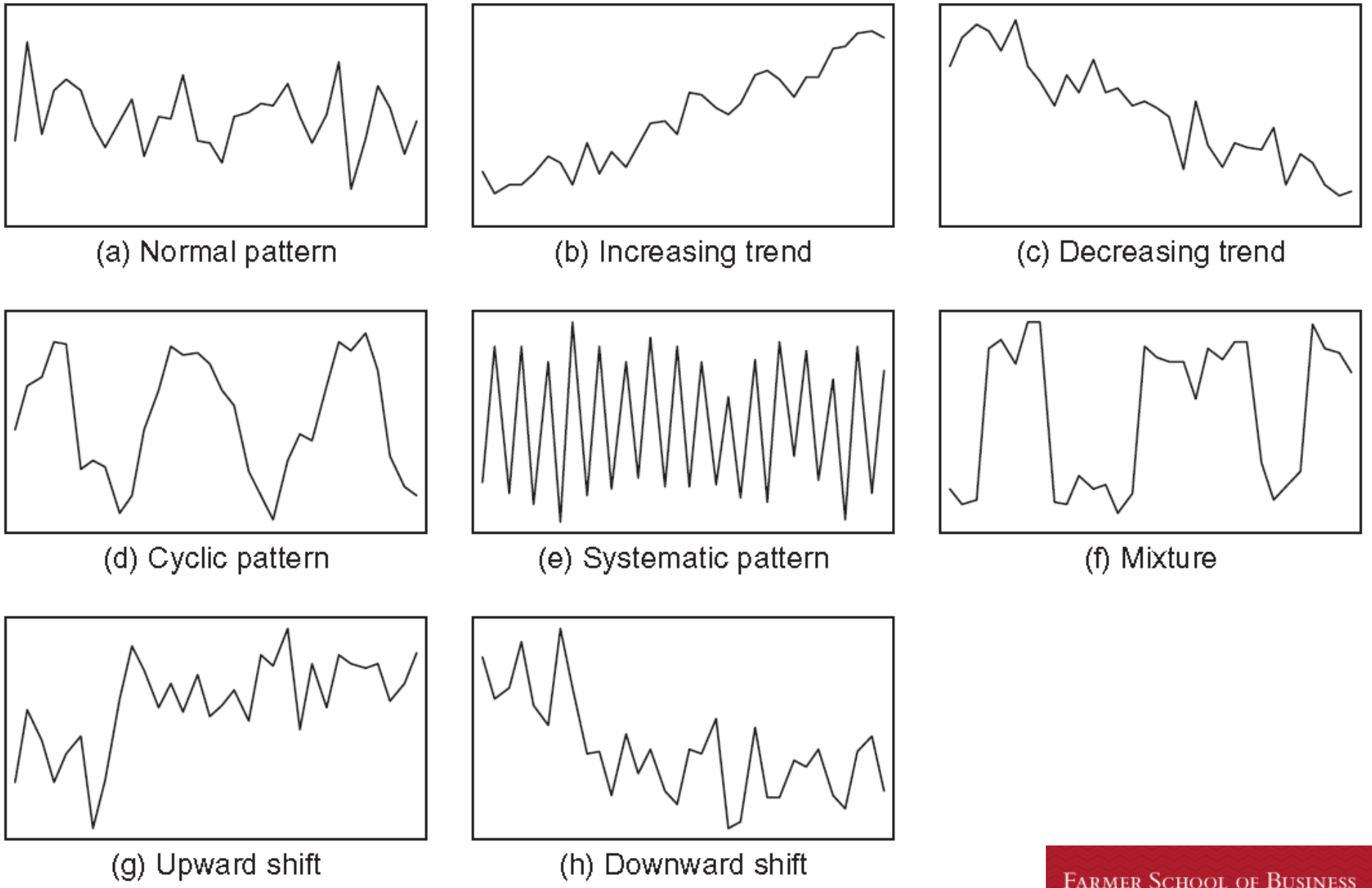


Fig. 1. Examples of typical control chart patterns.

CCRP Literature from Hachicha and Ghorbel (2012)

Interestingly, the majority of the papers they reviewed (61.47%) used an ANN approach to pattern recognition.

Their study revealed that only nine authors have published nearly half of the 122 CCPR papers reviewed.

Additionally, only 16 out of the 122 papers reviewed considered multivariate processes, and only five out of the 122 papers evaluated involved applying the proposed method on real process data (Hachicha and Ghorbel 2012, p. 210-213).

CCRP

Woodall and Montgomery (2014) stated that “Despite the large number of papers on this topic [neural network control charts including those for CCPR] we have not seen much practical impact on SPC”.

We believe that this lack of impact on the practice of SPC is due to several reasons mainly due to

(1) lack of discussion addressing the baseline operation of a process
(2) little advice given as to how to apply the methods in practice.

Interesting research, but won’t help much for our NFL example, we have no “training” data.

Other Applications of Neural Networks

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

The main motivation behind these methods lies in attempting to bridge SPC (where the focus has been on detecting an out-of-control condition) and engineering practitioners (where fault detection represents the first aspect of process monitoring).

Support Vector Machines

Other Applications of Support Vector Machines (different than OCC)

SVM has been applied

- to batch process monitoring (Yao et al. 2014)
- in the use of support vector regression (SVR) as a precursor to residual based multivariate cumulative sum (MCUSUM) chart for monitoring autocorrelated data (Issam and Mohamed 2008).
- To monitoring the predicted probability of class membership from an SVM along using bootstrap control limits, Chongfuangprinya et al. (2011)
- to estimate the magnitude of the shift in the process mean as detected by a CUSUM chart, Cheng et al. (2011)

Ensemble Models

In combined applications of fault detection-identification-diagnosis, 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 and showed that this supervised learning method outperforms a multivariate exponentially weighted moving average control chart.

Note: Not only does this method show promise in the realm of *big data,* but it forgoes the usual distributional assumptions that can be troublesome with multivariate SPC methods.

Ensemble Methods

In fact, a recent application of ensemble methods in public health surveillance by Davila et al. (2014) illustrated the use of an ensemble of decision trees to monitor counts (or rates) of a disease.

This greatly improves upon current methods of public health surveillance which typically involve only low dimensional data and cannot take into account additional data such as demographic information.

Back to the Example

So what do we do?

Unsupervised methods:

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

~~Supervised methods~~

- ~~CCRP~~
- ~~Neural Networks~~
- ~~SVM/SVR~~
- ~~Profile Monitoring (did not discuss)~~
- ~~Ensemble methods~~

K^2 or 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 nearest neighbors in a reference sample.
3. The control limit for the chart is determined by bootstrapping the average squared distances for each observation and taking the $(1-\alpha)^{th}$ quantile of bootstrapped distribution.

K^2 or KNN chart

Although we recognize several limitations to this approach, we chose to use the K^2 (kNN) chart, an OCC control chart, because

(1) it has less computational cost than the *K*-chart (see Sukchotrat et al., 2009),

(2) is more robust to the i.i.d assumption requirement (Kim, et al., 2010)

Choice of k

Breunig et al. (2000) recommended a range of *k* between 10 and 50.

In this example, the choice of *k* made little difference since there were a number of hours in the reference sample with no Wikipedia hits, and these observations formed the *k* nearest neighbors; thus, we selected *k* = 20, and α=0.01.

Cyclic Autocorrelation

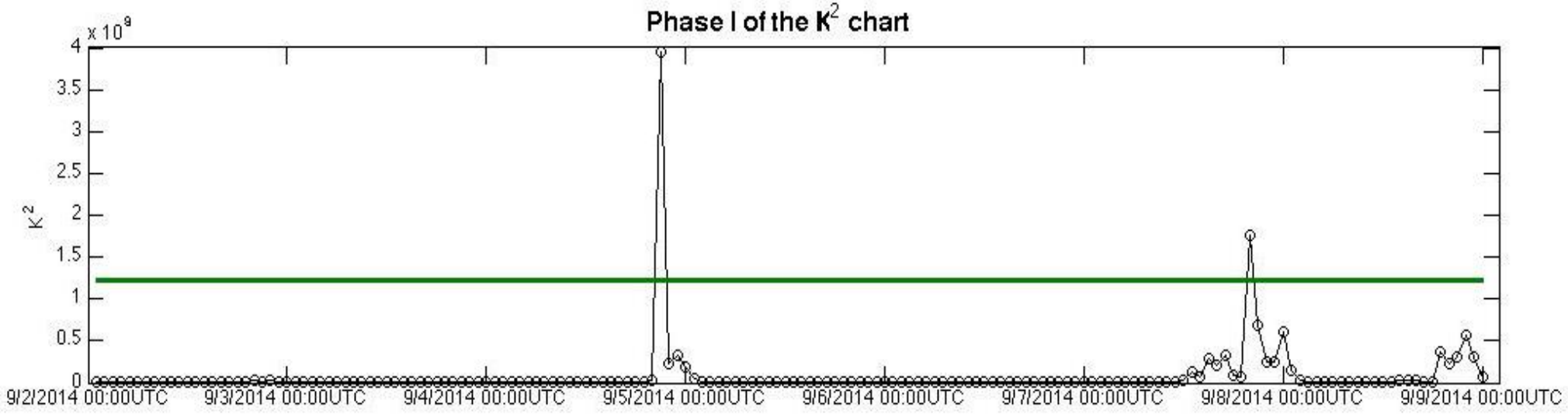
We borrowed from the bio-surveillance literature and used the residuals from Holt-Winters model lagged by 24 hours on each player with seasonal and trend component (see Shmueli and Fienberg 2006; Burkom et al. 2007).

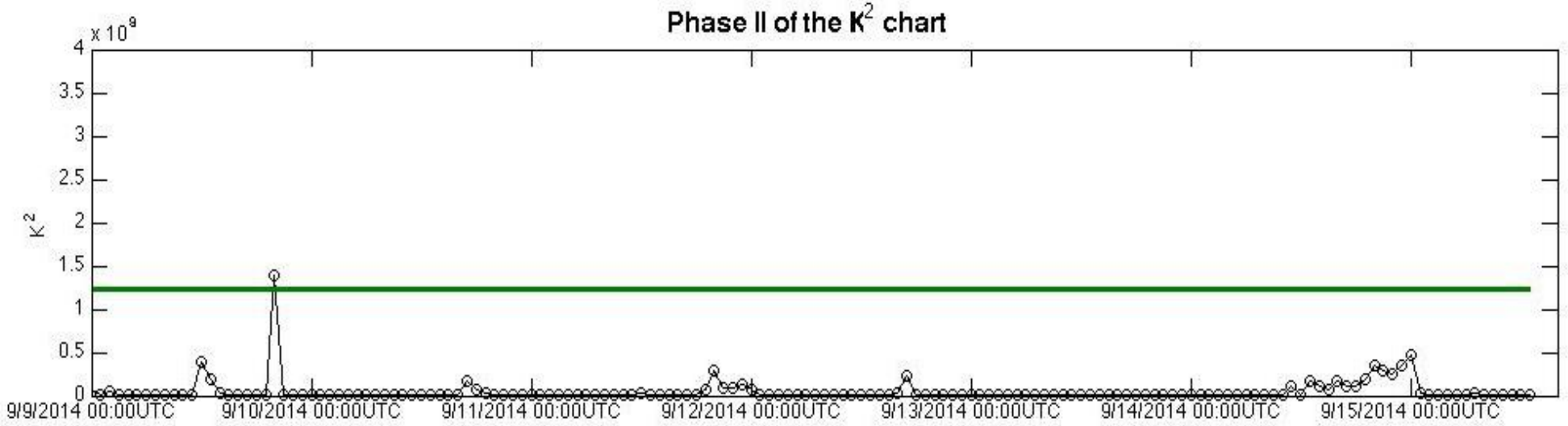
We then analyzed the multivariate data set containing the p=1916 sets of residuals using the K^2 chart.

Phase I and Phase II

The first 168 observations taken during the first week of the season were used to establish the baseline Phase I sample.

The remaining 162 observations (note the residuals for the lagged 24 observations were not used) taken during the second week of the season are using for Phase II monitoring.


Phase I of the K2 chart
x 10^9
4
3.5
3
2.5
2
1.5
1
0.5
0
K2
9/2/2014 00:00UTC
9/3/2014 00:00UTC
9/4/2014 00:00UTC
9/5/2014 00:00UTC
9/6/2014 00:00UTC
9/7/2014 00:00UTC
9/8/2014 00:00UTC
9/9/2014 00:00UTC


Phase II of the K2 chart
x 10^9
4
3.5
3
2.5
2
1.5
1
0.5
0
K2
9/9/2014 00:00UTC
9/10/2014 00:00UTC
9/11/2014 00:00UTC
9/12/2014 00:00UTC
9/13/2014 00:00UTC
9/14/2014 00:00UTC
9/15/2014 00:00UTC

Phase I

We did not remove these signals from our analysis for two reasons:

1. We are not certain that these signals are anomalous to the process, thus we chose to leave them in the sample
2. Sukchotrat et al. (2009) did not discuss an iterative approach to the K^2 chart, where assignable causes are removed and the limits recalculated.

Signals

	Date and Time of Signal	Possible Assignable Causes
Phase I	09/04/2014 20:00:00 UTC	Packers vs. Seahawks game. Aaron Rodgers had a poor performance and Russell Wilson had a particularly good game.
	09/07/2014 17:00:00 UTC	Sunday Football games
Phase II	09/09/2014 19:00:00 UTC	LeSean McCoy was called out by a restaurant owner for leaving a $0.20 tip on $61.56 meal. This incident was highly publicized. ESPN's E:60 aired an episode on Marquise Goodwin and his sister Deja, born with cerebral palsy.

Comments

While not a perfect analysis, the use of the K^2 chart in this example provides a useful example of the need for more research on data driven (as opposed to model-based) control charts (see Breiman (2001b) for an interesting discussion of model-based versus data-driven statistical models).

There are many open research questions with the OCC control charts, and they their performance has not been well-studied. This example is not intended to encompass all of the challenges present in big data monitoring, but serves as one example of a few of the complexities of this type of data.

Closing Remarks

Conclusions/Remarks

- Our view is that there is a significant need for statistical monitoring of data streams and *big data* for detection of process changes
- Unlike in traditional applications, there are often no physical or engineering principles that can be used to understand this behavior.
- We see tremendous opportunity for developments regarding how one establishes an in-control reference sample (Phase I) for multivariate processes, and especially for multivariate processes measured with mixed variable types.
- Monitoring this data with the existing techniques is challenging, and our experience suggests that the traditional model-based SPC methods are ill-suited to *big data* monitoring.

Questions?

References

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

Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000), "LOF: identifying density-based local outliers," *Proceedings of the ACM SIGMOD 2000 International Conference on Management of Data*, 29, 93–104.

Camci, F., Chinnam, R.B., and Ellis, R.D. (2008), "Robust Kernel Distance Multivariate Control Chart Using Support Vector Principles," *International Journal of Production Research*, 46, 5075-5095.

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.

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.

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

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.

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.

Kang, J.H. and Kim, S.B. (2011), "Clutering-Algorithm-based Control Charts for Inhomogeneously Distributed TFT-LCD Processes," International Journal of Production Research, 51, 5644-5657.

Kumar, S., Choudhary, A.K., Kumar, M., Shankar, R. and Tiwari, M.K. (2006), "Kernel Distance-Based Support Vector Methods and its Application in Developing a robust K-chart," *International Journal of Production Research*. 44, 77-76.

References

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

Liu, C. and Wang, T. (2014), "An AK-chart for the Non-Normal Data," *International Journal of Computer, Information, Systems and Control Engineering*, 8, 992-997.

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

Sukchotrat, T., Kim, S.B. and Tsung, F. (2009), "One-Class Classification-Based Control Charts for Multivariate Process Monitoring," *IIE Transactions*, 42, 107-120.

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.

Tax, D.M. and Duin, R.P. (1999), "Support Vector Domain Description," *Pattern Recognition Letters.* 20, 1191-1199.

Tax, D.M. and Duin, R.P. (2004), "Support Vector Data Description," *Machine Learning,* 54, 45-66.

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

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.