Data Mining Methods and Techniques for Monitoring and Optimizing Complex (Continuous) Processes

ASIAS Technology and Tools Symposium
MITRE Corporation, McLean, VA; July 27/28, 2009

Thomas Hill, Ph.D.
VP Analytic Solutions

StatSoft®
STATISTICA

data analysis • data mining • quality control • web-based analytics

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o.
China: StatSoft China
France:StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Topics, Summary

- The goal of this presentation is to review typical data mining workflows and applications, and to highlight what is common and what is different when applying these methods to continuous process data of different types
- Overview:
  - How is data mining different from statistical data analysis/modeling
  - Common workflows
  - “Traditional” data mining applications
  - Working with continuous process data: Unique data issues
  - Typical data mining (goals) for continuous process data
  - Example 1: Model-based monitoring of time-stamped (batch) data
  - Example 2: Optimization of continuous combustion processes
  - Summary

StatSoft Inc.
StatSoft Power Solutions
(A Subsidiary of StatSoft Inc.)

- Headquarters: Tulsa OK, USA
- 24 overseas offices on all continents

● StatSoft Offices
● StatSoft Training Centers
● Authorized Distributors

- Over 20 years of leadership in the development and application of data mining, advanced process monitoring, optimization, and advanced analytics software (*STATISTICA* data analysis and reporting solutions)
- For details see: www.StatSoft.com and www.StatSoftPower.com

Overview:
Data Mining and Statistical Modeling

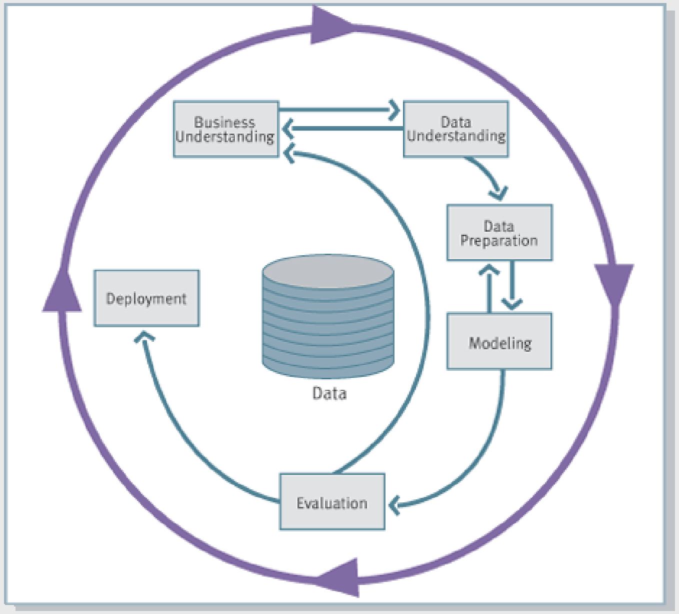
Knowledge Discovery vs. Statistical Analysis

- **Statistical Analysis**
  - Focuses on “hypothesis testing” and “parameter estimation”
  - Fits “parsimonious statistical models” with the goal to “explain” complex relationships with fewer parameters
  - Examples: Regression, nonparametric statistics, factor analysis, traditional quality control
- **Data Mining**
  - **The data are your model!**
  - Focuses on knowledge discovery, detection of patterns, clusters, and so on; we only have data and no (or few) expectations and hypotheses
  - Fits simple models or complex models (such as neural nets) to enable valid prediction
  - Examples: K-nearest-neighbor methods, recursive partitioning (trees), neural nets, stochastic gradient boosting of tree classifiers, random forests, support vector machines

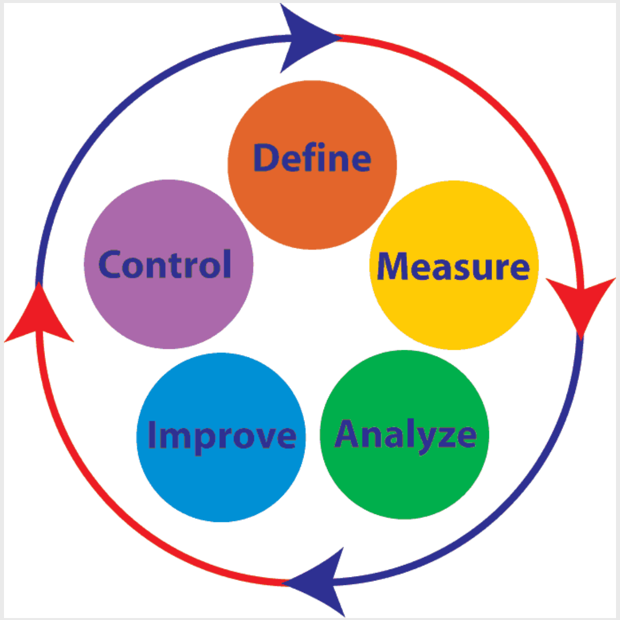
Common Workflows

- SEMMA:
  Sample → Explore → Modify → Model → Assess

- CRISP
  (Cross-Industry Standard Process for Data Mining)



- DMAIC (Six Sigma; manufacturing and engineering applications)



Common Workflows (cont.)

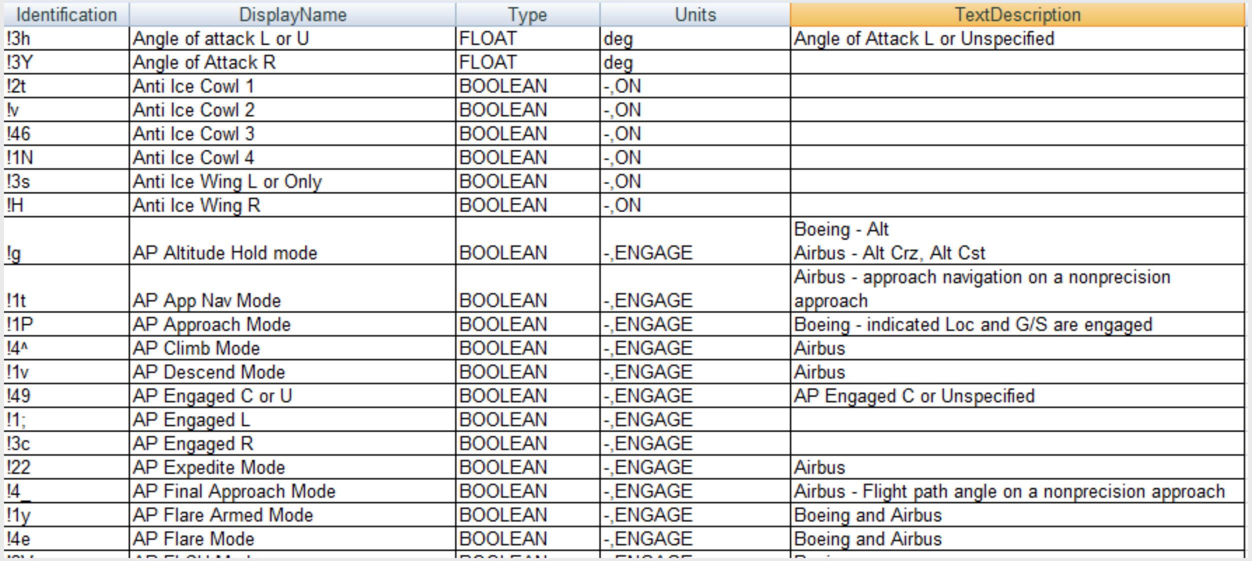
- Data Preparation
  - Missing data, outliers, bad data, too many classes (zip codes...), etc.
- Model building
  - Supervised learning (predictive modeling), unsupervised learning (clustering)
  - General approximators: Recursive partitioning (trees), neural nets, support vector machines
  - Some specialized methods, approaches, applications:
    - Text mining
    - Association, sequence analysis
    - Others: Clustering of DNA sequences, click streams, pattern recognition, anomaly detection, etc.
- Interpretation, deployment
  - Scoring (predicting, classifying) new cases
  - Inverse prediction, optimization ("what values of *X()* will optimize *Y*")
  - Simulation through the model to evaluate risk, failure rates, etc.

“Traditional” Data Mining Applications (Not a comprehensive list)

- Selecting "good cases" from large number of applicant cases
  - Credit scoring
  - Insurance rate making
- Predicting what will happen next
  - Marketing campaigns
  - Predicting churn
- Clustering, Market segmentation
  - Whom to offer what
- Identifying frequent associations, links (in “market baskets”)
  - Cross-selling, up-selling
- Some specialized applications
  - Fraud/anomaly detection
  - Intrusion detection

Working with Continuous Process Data
Unique Data Issues

- Data available from manufacturing, automated data acquisition systems, etc. are often high-dimensional and indexed against time
  - E.g., continuous flight data



Identification	DisplayName	Type	Units	TextDescription
!3h	Angle of attack L or U	FLOAT	deg	Angle of Attack L or Unspecified
!3Y	Angle of Attack R	FLOAT	deg	
!2t	Anti Ice Cowl 1	BOOLEAN	-,ON	
!v	Anti Ice Cowl 2	BOOLEAN	-,ON	
!46	Anti Ice Cowl 3	BOOLEAN	-,ON	
!1N	Anti Ice Cowl 4	BOOLEAN	-,ON	
!3s	Anti Ice Wing L or Only	BOOLEAN	-,ON	
!H	Anti Ice Wing R	BOOLEAN	-,ON	
!g	AP Altitude Hold mode	BOOLEAN	-,ENGAGE	Boeing - Alt Airbus - Alt Crz, Alt Cst
!1t	AP App Nav Mode	BOOLEAN	-,ENGAGE	Airbus - approach navigation on a nonprecision approach
!1P	AP Approach Mode	BOOLEAN	-,ENGAGE	Boeing - indicated Loc and G/S are engaged
!4^	AP Climb Mode	BOOLEAN	-,ENGAGE	Airbus
!1v	AP Descend Mode	BOOLEAN	-,ENGAGE	Airbus
!49	AP Engaged C or U	BOOLEAN	-,ENGAGE	AP Engaged C or Unspecified
!1;	AP Engaged L	BOOLEAN	-,ENGAGE	
!3c	AP Engaged R	BOOLEAN	-,ENGAGE	
!22	AP Expedite Mode	BOOLEAN	-,ENGAGE	Airbus
!4_	AP Final Approach Mode	BOOLEAN	-,ENGAGE	Airbus - Flight path angle on a nonprecision approach
!1y	AP Flare Armed Mode	BOOLEAN	-,ENGAGE	Boeing and Airbus
!4e	AP Flare Mode	BOOLEAN	-,ENGAGE	Boeing and Airbus

- Observations are not independent of each other; high autocorrelation
  - Thousands of cases, but only very few independent pieces of information
  - Over-fitting is a serious issue
  - Simply applying data mining algorithms can yield *very* misleading results!

Working with Continuous Process Data Unique Data Issues (cont.)

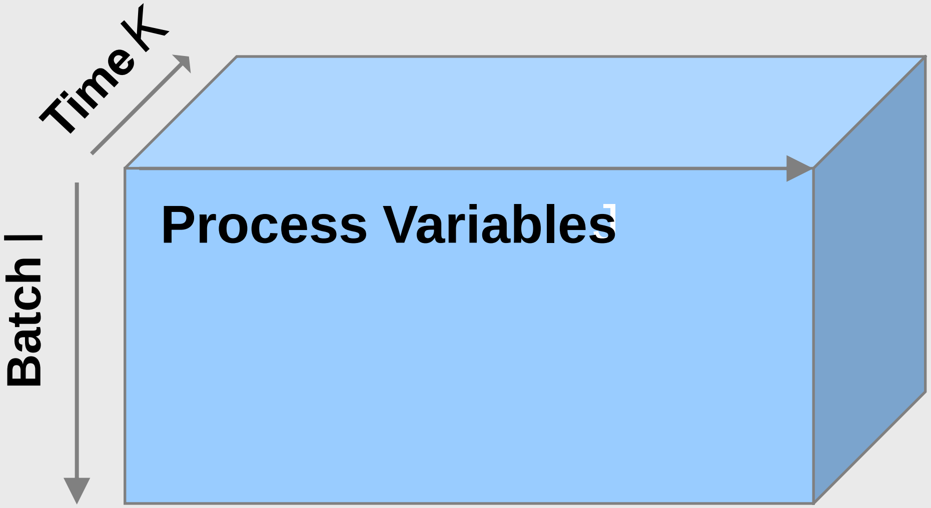
- For predictive modeling, a usually large number of data tags (variables) must be classified before analyses as:
  - Inputs, actionable: I can change them and/or do something about them
  - Inputs, non-actionable: I cannot change them, they are “noise” variables not under my control
  - Outputs: These are the results of the inputs, and the parameters of interest to predict, optimize, etc.
  - Constraints: These are variables I need to carry along and “watch”, e.g., critical temperatures, air speed limitations, etc.
- For predictive modeling in the time domain, to achieve useful results, it is usually necessary to identify “actionable time intervals”
- The majority of available data records may not be interesting
  - For example, the majority of the recorded cases may describe a single steady state (cruise flight)
  - Clustering of data, and then re-sampling from different states may be necessary; or model different states separately

Typical Data Mining Goals for Continuous Process data

- Detect problems (early, before things are about to go wrong)
  - Hundreds or thousands of parameters must be monitored simultaneously
  - How to avoid being in “perpetual alarm”?
  - To identify problems in transitional states, or “maturation”, identify deviations from typical “trajectories” (see *Example 1*)
- Use model to “optimize system,” so if I “change” some parameters *X1* and *X2*, a desirable *Y* will result (see *Example 2*)
  - This is very different from *Select-good-cases* problem; e.g., if a model predicted smokers to be a worse credit risk, will they become a better credit risk if I get them to quit smoking? Probably not…..
- Optimize for a *robust* system, i.e., one that can handle as much unreliability, randomness, and “junk” on the inputs, and still produce low-variability, reliable outputs (see *Example 2*)
  - Remember (Taguchi): Quality is a quadratic function of variability!
- Root cause (“commonality”) analysis in high dimensional data (e.g., detecting higher-order interactions)

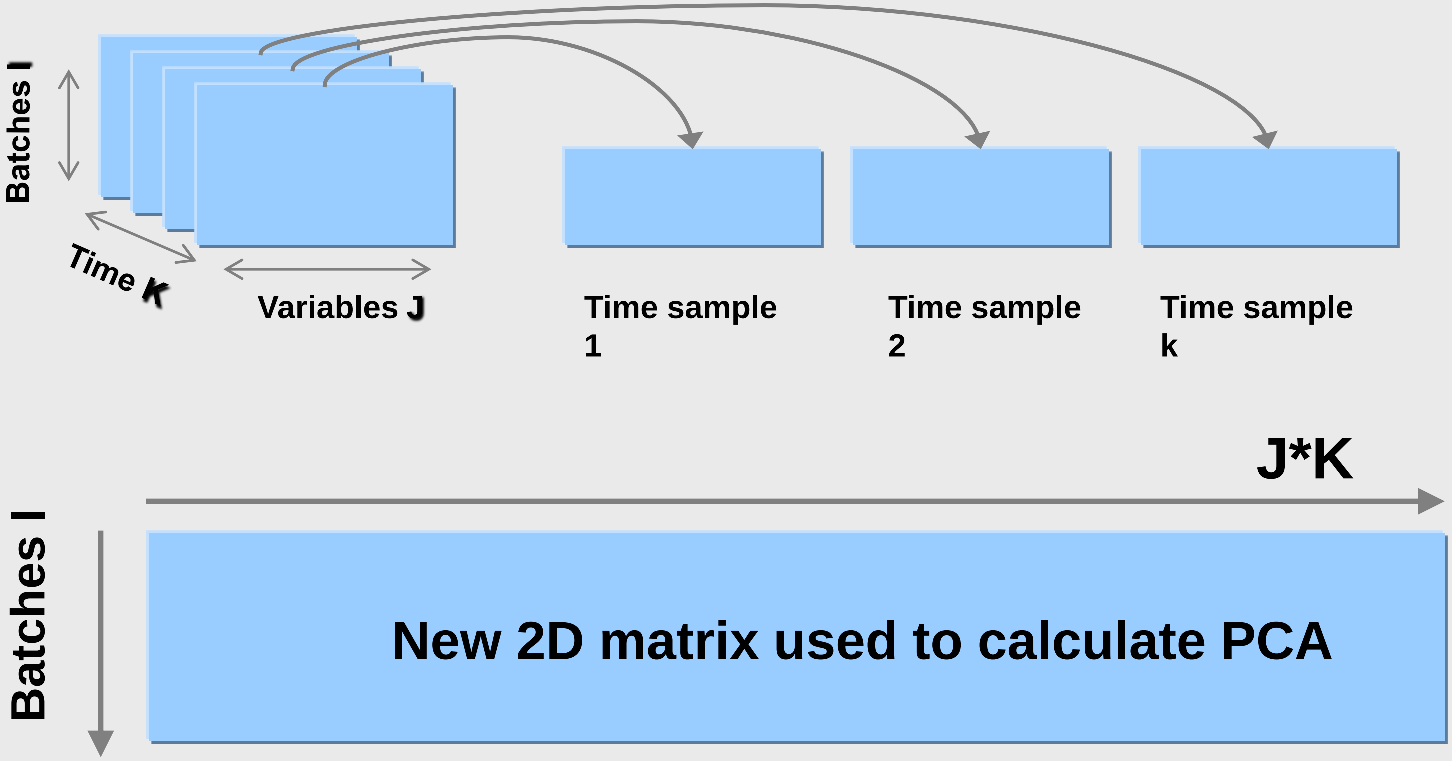
Example 1: Model-based monitoring of time-stamped (batch) data

- Commonly used workflow in chemical, pharmaceutical (batch) and process manufacturing
- E.g., data come in batches, and are time stamped; goal is to monitor maturation/progress
- Approach:
  - Identify "good" or "successful cases" ("golden batches")
  - Build a model of the process inputs, for predicting elapsed time; the model "prediction" (elapsed time) summarizes "maturation" (a latent construct)
  - Unstack data to build a lower-dimensional model summarizing the data matrix
  - Compare the lower-dimensional (derived, latent) dimensions to "good cases"; monitor "trajectories"



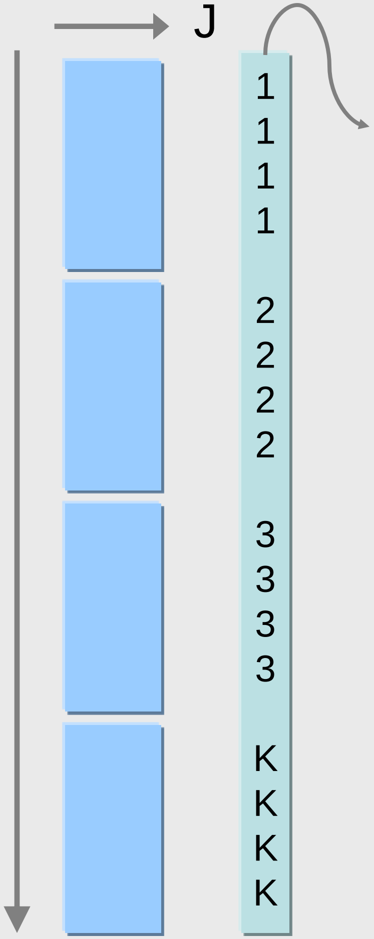
Example 1: Model-based monitoring of time-stamped (batch) data (cont..)

- Unstacking batch(phase)/time data, Timewise



Example 1: Model-based monitoring of time-stamped (batch) data (cont..)

- Unstacking batch(phase)/time data, Batchwise (phasewise)

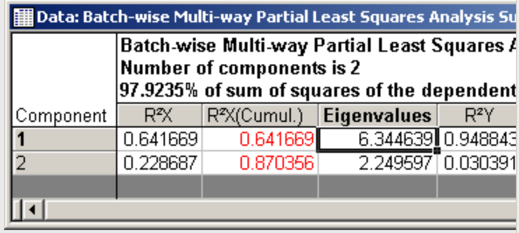


"Y" maturity vector added

- The 3D matrix of batch/phase data unfolded along the batch direction
- The resulting 2D matrix has I*K rows and J columns
- Can be modeled via Partial Least Squares or other predictive modeling methods

Example 1: Model-based monitoring of time-stamped (batch) data (cont..)

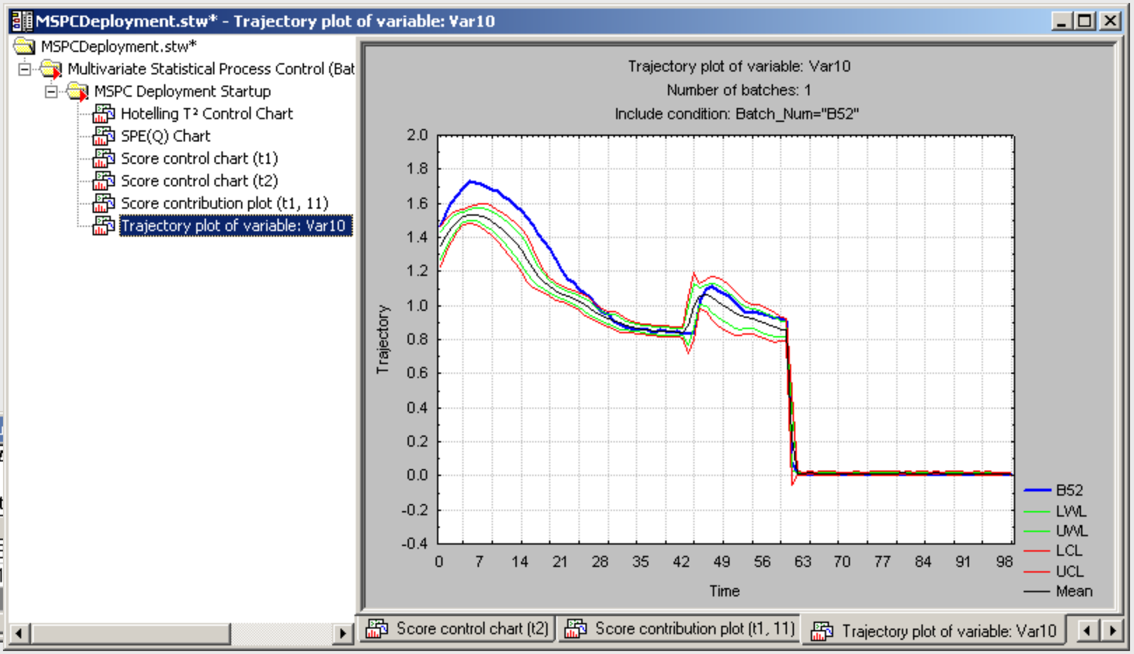
- Widely used methods are described by MacGregor, Wold: Use Principal Components Analysis (PCA and Partial Least Squares (PLS) methods; NIPALS algorithm
  - Effectively, we extract the orthogonal linear combinations of predictors that maximize "maturation", "progress“



Data: Batch-wise Multi-way Partial Least Squares Analysis Su

Batch-wise Multi-way Partial Least Squares A
Number of components is 2
97.9235% of sum of squares of the dependent

Component	R^2X	R^2X(Cumul.)	Eigenvalues	R^2Y
1	0.641669	0.641669	6.344639	0.948843
2	0.228687	0.870356	2.249597	0.030391



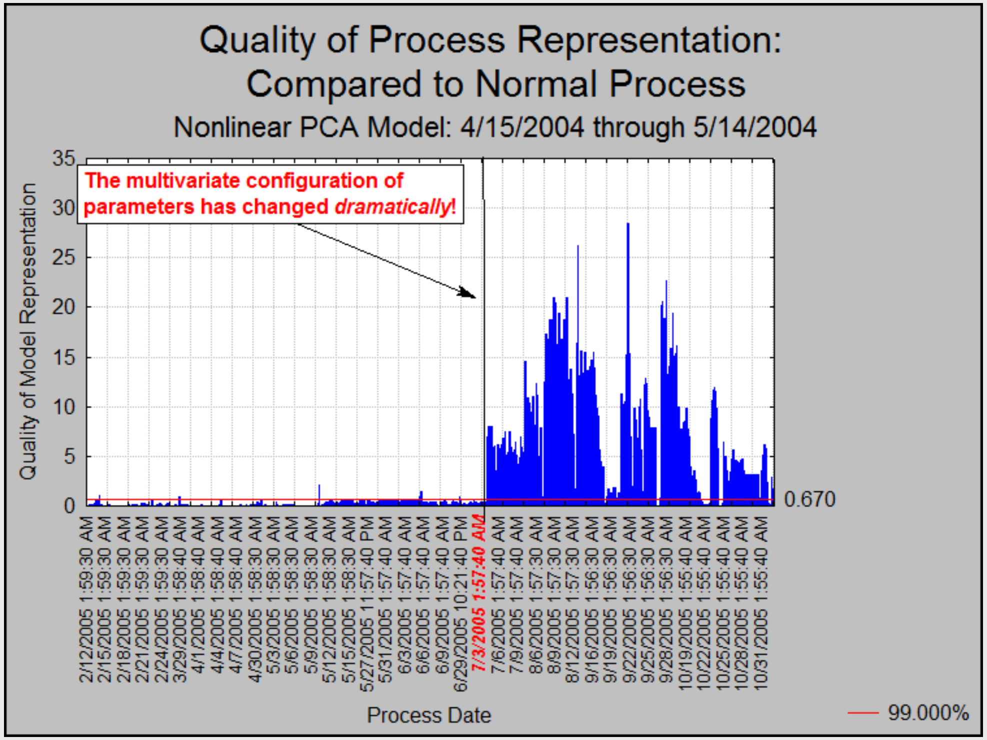
MSPCDeployment.stw* - Trajectory plot of variable: Var10

- MSPCDeployment.stw*
  - Multivariate Statistical Process Control (Bat
    - MSPC Deployment Startup
      - Hotelling T² Control Chart
      - SPE(Q) Chart
      - Score control chart (t1)
      - Score control chart (t2)
      - Score contribution plot (t1, 11)
      - Trajectory plot of variable: Var10

  - These methods are easy to use, and interpretation/workflow is well defined (number of components, T² chart, contribution plots)

Example 1: Model-based monitoring of time-stamped (batch) data (cont..)

- In other industries, "virtual sensors" are built using Neural Nets; in effect, instead of linear projections into lower-dimensional space we use nonlinear projections
  - This graph is from a mining application, monitoring the grinding process
  - More difficult to identify root causes of problems



Quality of Process Representation: Compared to Normal Process

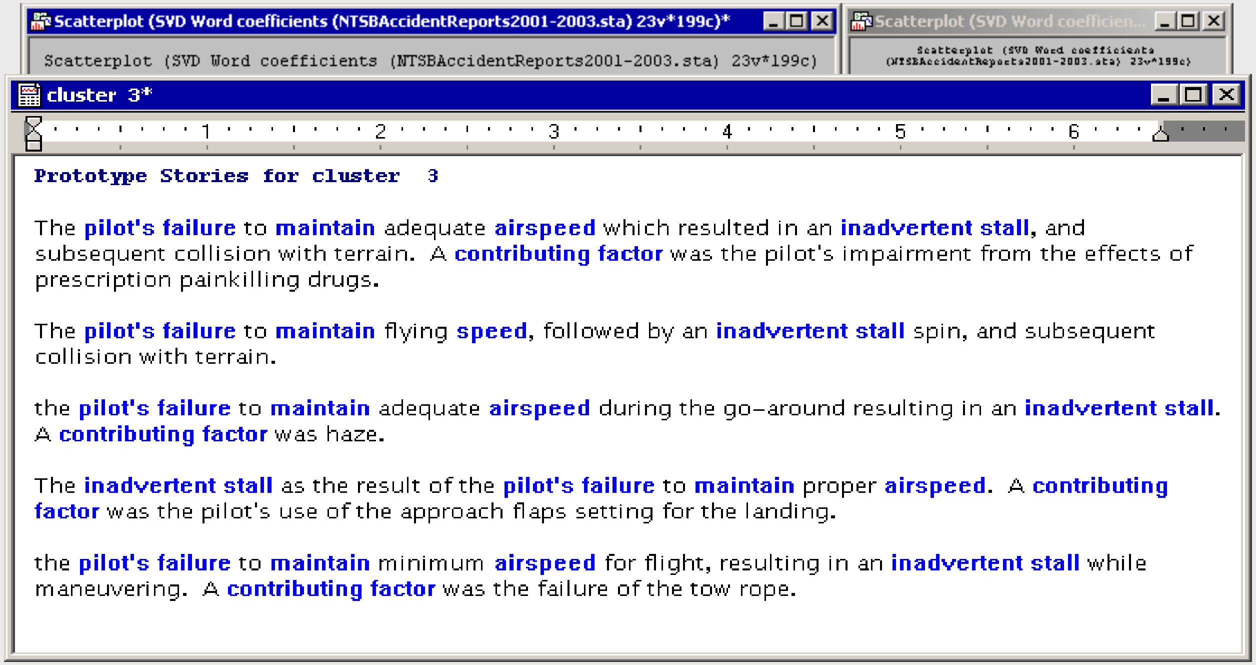
Nonlinear PCA Model: 4/15/2004 through 5/14/2004

The multivariate configuration of parameters has changed *dramatically*!

- See also *Lewicki, P., Hill, T., Qazaz, C. (2007). Multivariate quality control. Quality Magazine, April, pp 38.*

Text Mining as a Dimension Reduction Problem

- Textual or other unstructured information can also be incorporated, once it is “numericized”
  - For example, count the (stemmed) words
  - Perform singular value decomposition
  - Work with component scores (latent semantic index)
  - For example, here are some results based on *Probable Cause* narratives from the NTSB accident database


Scatterplot (SVD Word coefficients (NTSBAccidentReports2001-2003.sta) 23v*199c)*
Scatterplot (SVD Word coefficients (NTSBAccidentReports2001-2003.sta) 23v*199c)
Scatterplot (SVD Word coefficien...
Scatterplot (SVD Word coefficients (NTSBAccidentReports2001-2003.sta) 23v*199c)
cluster 3*
1
2
3
4
5
6

Prototype Stories for cluster 3

The **pilot's failure** to **maintain** adequate **airspeed** which resulted in an **inadvertent stall**, and subsequent collision with terrain. A **contributing factor** was the pilot's impairment from the effects of prescription painkilling drugs.

The **pilot's failure** to **maintain** flying **speed**, followed by an **inadvertent stall** spin, and subsequent collision with terrain.

the **pilot's failure** to **maintain** adequate **airspeed** during the go-around resulting in an **inadvertent stall**. A **contributing factor** was haze.

The **inadvertent stall** as the result of the **pilot's failure** to **maintain** proper **airspeed**. A **contributing factor** was the pilot's use of the approach flaps setting for the landing.

the **pilot's failure** to **maintain** minimum **airspeed** for flight, resulting in an **inadvertent stall** while maneuvering. A **contributing factor** was the failure of the tow rope.

Example 2: Optimization of Continuous Combustion Processes

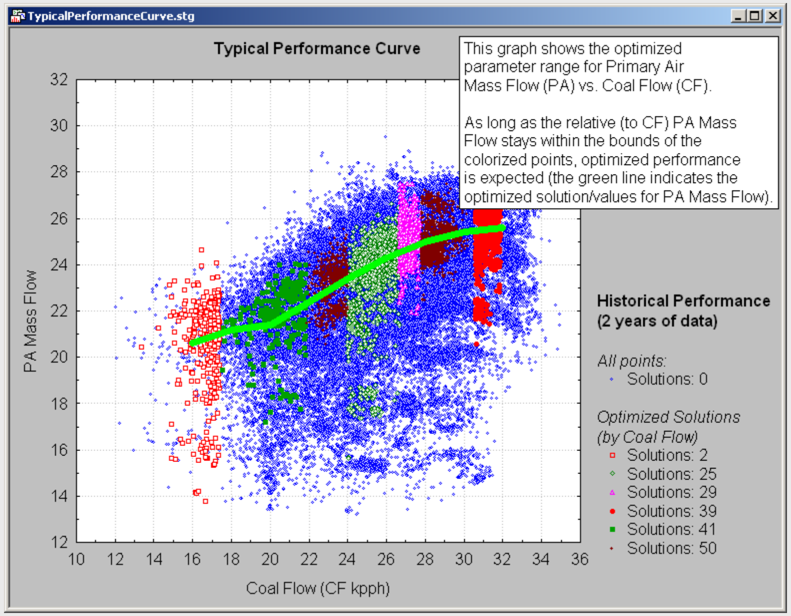
- See:
  - Electric Power Research Institute (EPRI) and StatSoft Project 44771 (2009). *Statistical Use of Existing DCS Data for Process Optimization.* EPRI: Palo Alto.
  - Hill, T. (2009). Optimization Strategies Based on Historical Data for Existing Coal Burning Furnace Technologies. *Energy and Environment Conference*; Phoenix, AZ.
- One-minute interval data for hundreds (thousands) of parameters
- Goal is to optimize and stabilize combustion processes using existing control systems and strategies, to achieve
  - Less NOx, CO
  - Better efficiency (lower CO2)
  - Stable, sustainable operational conditions (temperatures etc.)

Example 2: Optimization of Continuous Combustion Processes (cont.)

- Approach:
  - Clean and aggregate data to meaningful intervals (e.g., Autocorrelations < .8)
  - Identify in the data “steady states”, and transitions between states; usually in these applications we are interested in steady states
  - Build a model of the process inputs predicting important outcomes (e.g., NOx, CO emissions)
    - Also build models as necessary for constraints (e.g., FEGT: furnace exit gas temperatures)
  - Optimize for robust performance (stochastic optimization; consistent low-variability NOx, CO, and sustainable FEGT)
    - Carefully review degree of extrapolation: The models are likely not valid in “areas” where there are no empirical data!

Example 2: Optimization of Continuous Combustion Processes (cont.)

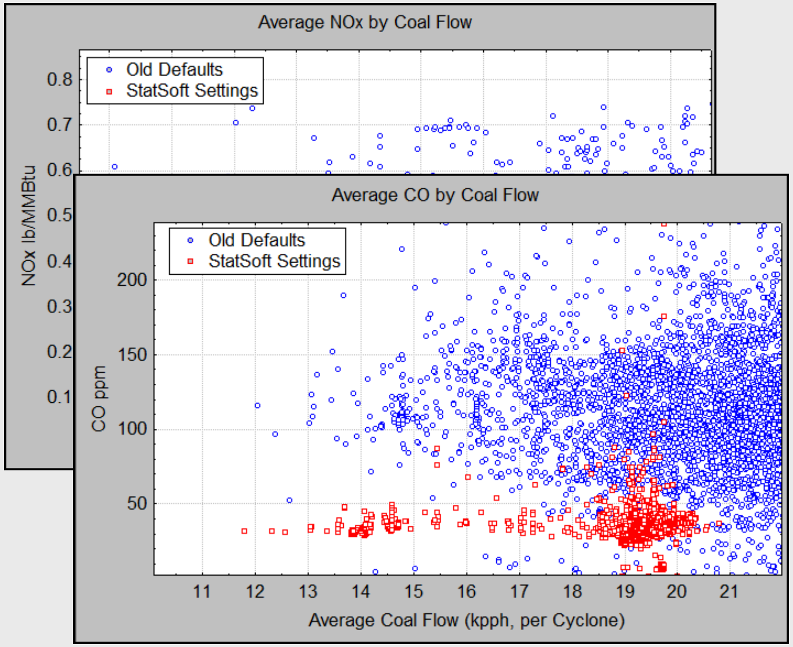
- New parameter "curves" implemented into the control system



This graph shows the optimized parameter range for Primary Air Mass Flow (PA) vs. Coal Flow (CF).

As long as the relative (to CF) PA Mass Flow stays within the bounds of the colorized points, optimized performance is expected (the green line indicates the optimized solution/values for PA Mass Flow).

- Results (after implementation of optimized parameters into automated control system) show improved sustained performance over the entire load range; e.g.:



Summary

- Statistical modeling involves the fitting “models” to data; in data mining for continuous data, the best approach often is to consider the historical data as your “model (“the data are your model”)
- Successful applications of data mining in discrete and continuous process manufacturing are different from “typical” data mining applications, e.g.,
    - Model-based process monitoring for high-dimensional (continuous) data, to track distance from desirable states, processes
    - Robust optimization for to identify stable states with desirable process characteristics (e.g., low emissions)
- Questions?