Increasing Process Understanding through Data Analysis

Alan Brown

Syngenta

Process Studies Group

Huddersfield, UK

25th SCI Process Development Symposium

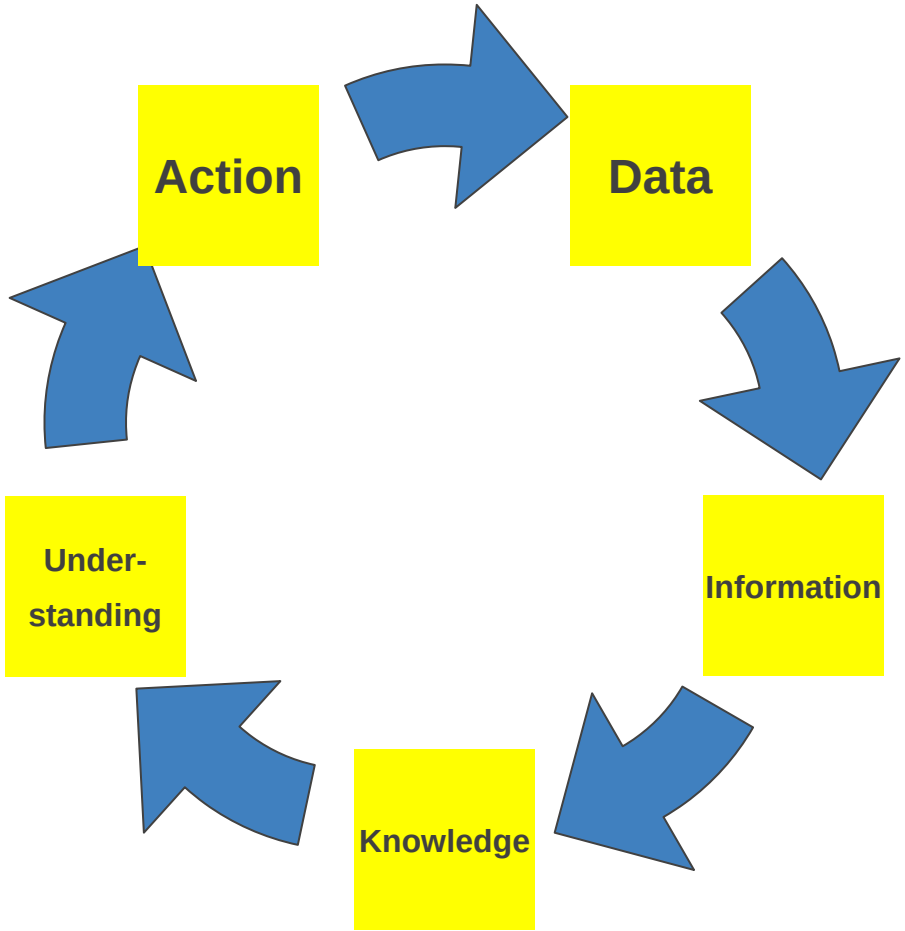
5th-7th December 2007, Cambridge,UK

Increasing Process Understanding through Data Analysis

- The Data to Action Cycle
- Process Variation
- Univariate Data Analysis
- Multivariate Data Analysis
  - Principal Component Analysis (PCA)
- Modelling (Examples)
  - Partial Least Squares (PLS) Regression
  - Recursive Partitioning (Classification Tree)
  - Artificial Neural Networks (ANN)

- Summary

Data in to Action Cycle



syngenta

Observational Data

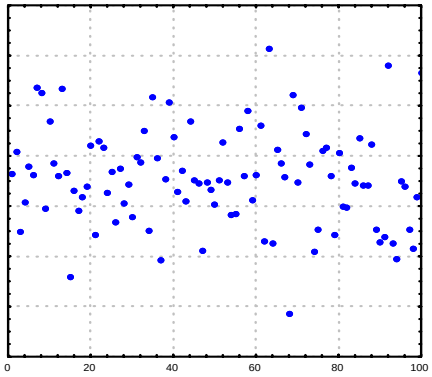
- Analytical Data
  - Assay, Impurities
- Process Summary Data
  - Input
    - Quality of raw material
    - Maximum temperature
    - pH
  - Output
    - Mass product
    - Yield
- Process Time series data
  - Flow, pressure, temperature, analyser
    - Continuous – steady state
    - Batch - dynamic

Observational Data - Variation

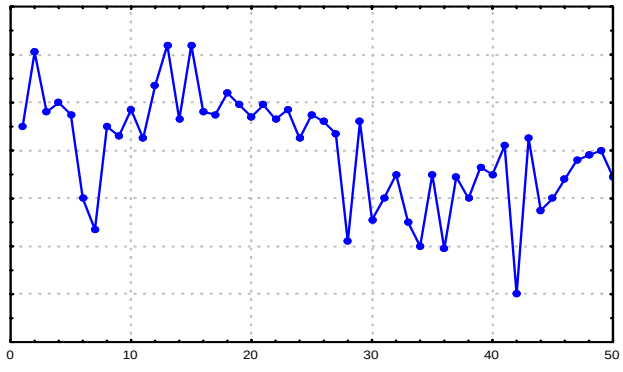
- Common theme
  - All measurements
    - Most are continuous
- Subject to variation
  - Noise – the variation observed between product manufactured under the same conditions and specifications
  - External to Process – environmental temperature, humidity
  - Process Causes – build up of waste products, ageing of catalyst, variation in loading a vessel
  - Assignable causes – quality of batches of raw material, in correct setting of equipment

Process Variation

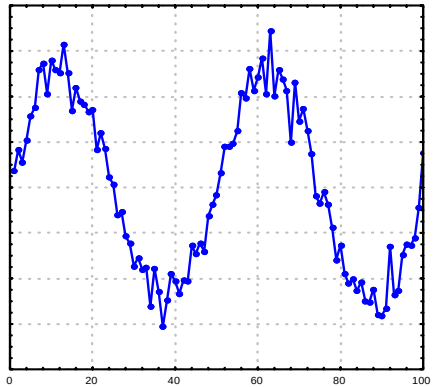
DIFFERENT TYPES OF VARIATION



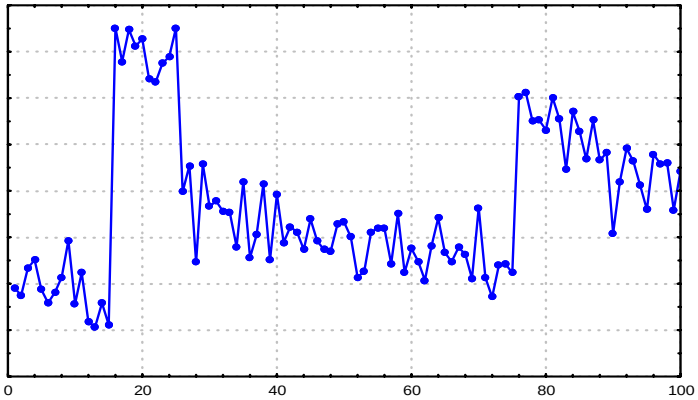
RANDOM VARIATION ONLY



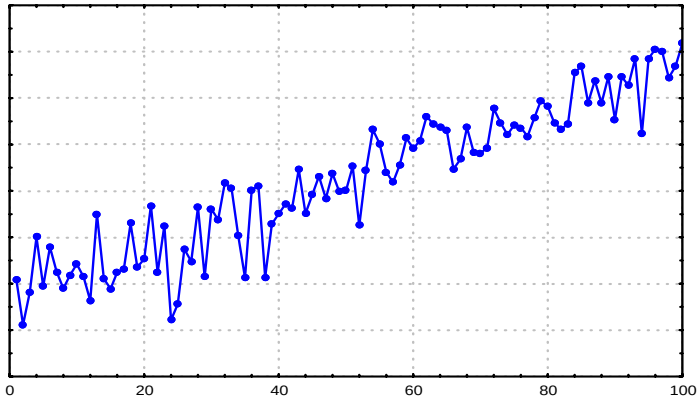
SHORT AND LONG TERM VARIATION



RECURRING CYCLES



SUDDEN JUMPS



TRENDS

syngenta

Characteristics of Random Variation

Product Assay for 200 batches

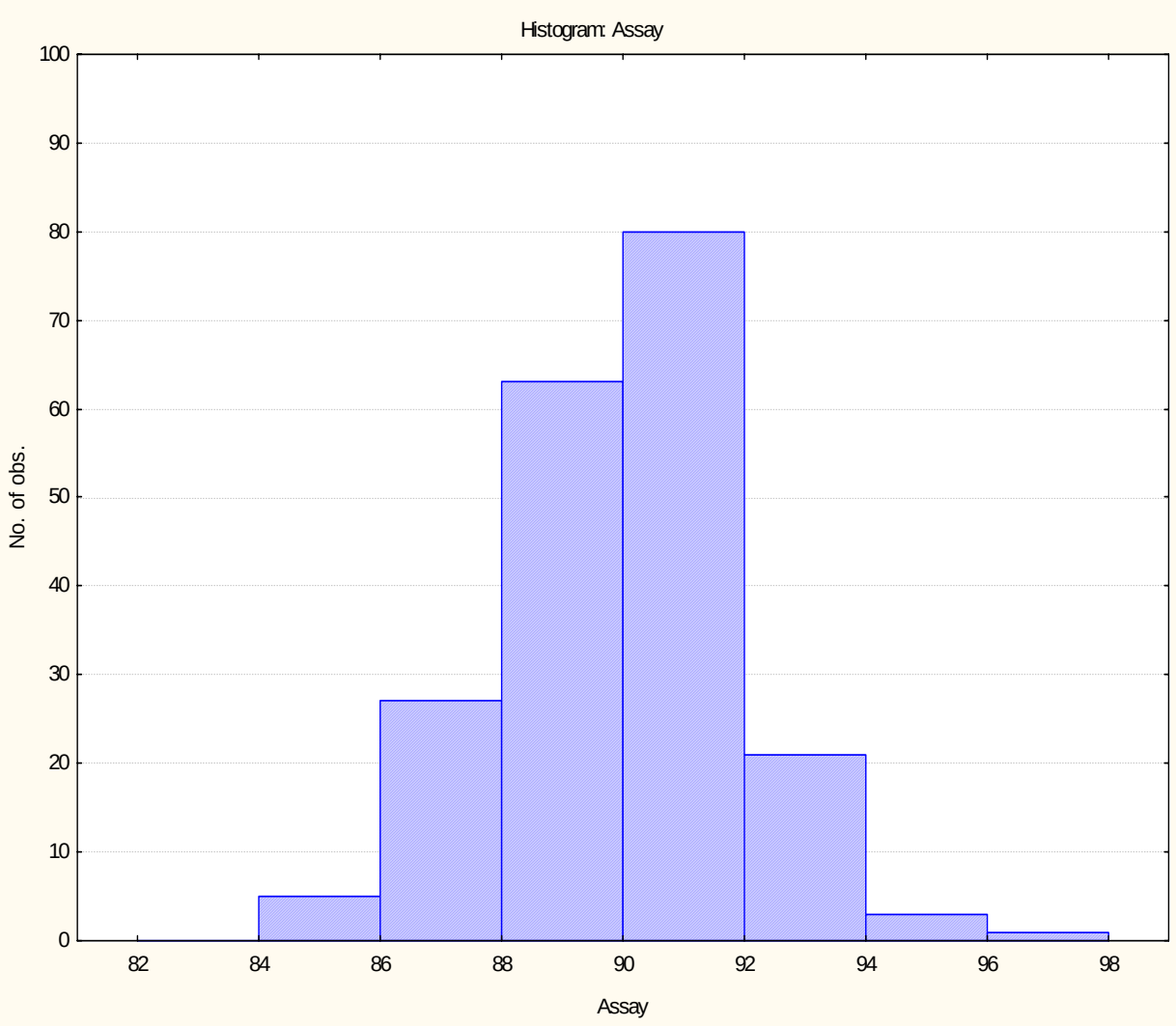
87.6	92.4	89.1	85.2	88.0	91.3	88.2	89.8	90.7	90.8
91.3	87.9	90.9	90.3	86.5	91.1	95.6	90.5	88.9	90.1
87.7	89.1	90.5	92.7	89.8	89.4	91.0	87.8	90.5	91.9
90.1	88.0	90.3	87.7	89.4	89.6	91.7	90.1	89.3	90.2
89.2	88.7	87.7	89.1	89.5	86.8	89.1	88.1	87.0	91.2
90.0	89.6	88.7	91.7	89.5	90.6	89.3	91.7	88.1	87.9
91.5	89.0	86.7	90.3	89.4	87.9	90.1	89.8	92.2	90.6
88.8	92.1	91.0	93.4	91.1	89.3	89.6	87.7	90.1	90.5
90.9	94.2	90.0	91.8	93.8	89.2	90.8	96.1	91.4	89.3
90.1	90.6	86.0	90.2	90.7	91.2	90.4	87.6	93.2	92.8
92.3	88.6	90.0	90.1	89.1	90.9	91.3	86.7	90.9	87.8
89.2	90.9	93.3	90.3	89.7	93.4	89.9	89.6	92.2	92.5
89.7	89.0	90.5	90.9	89.3	93.0	85.2	88.5	90.8	91.6
93.6	88.7	88.3	88.8	86.4	90.3	90.5	85.5	88.9	84.5
89.9	88.0	91.0	88.3	90.5	88.8	91.1	90.1	91.3	92.6
91.4	94.1	86.4	88.8	92.0	88.4	86.7	90.6	90.3	89.8
90.3	88.5	92.8	92.5	90.8	87.8	91.1	90.6	87.7	91.4
90.2	92.9	91.3	91.1	88.4	90.1	88.5	91.0	89.1	88.0
86.3	89.2	91.1	87.9	87.5	89.0	86.8	90.1	91.1	92.9
91.5	90.1	89.2	88.5	89.1	90.8	88.6	90.3	89.0	92.3

Summarise data in terms of location and spread

Mean = 89.9

Std dev = 1.9

Characteristics of Random Variation



Mean = 89.9

Std dev = 1.9

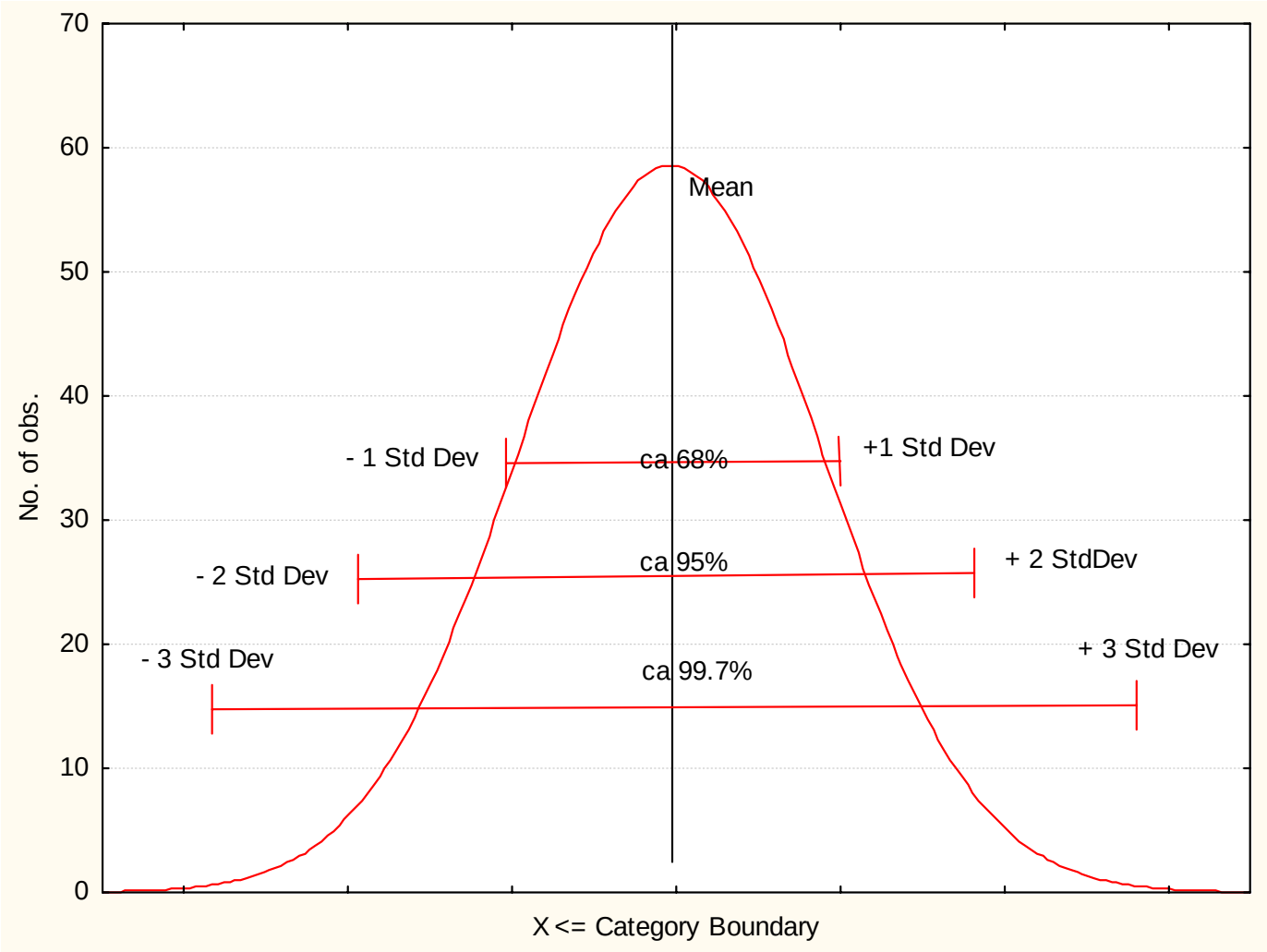
Range of observations ca 84-98

Approx 140 from the 200 (70%) are in the range 88-92
ie within ca +/- 1 std dev around the mean

Approx 190 from the 200 (95%) are in the range 86-94
ie within ca +/-2 std dev around the mean

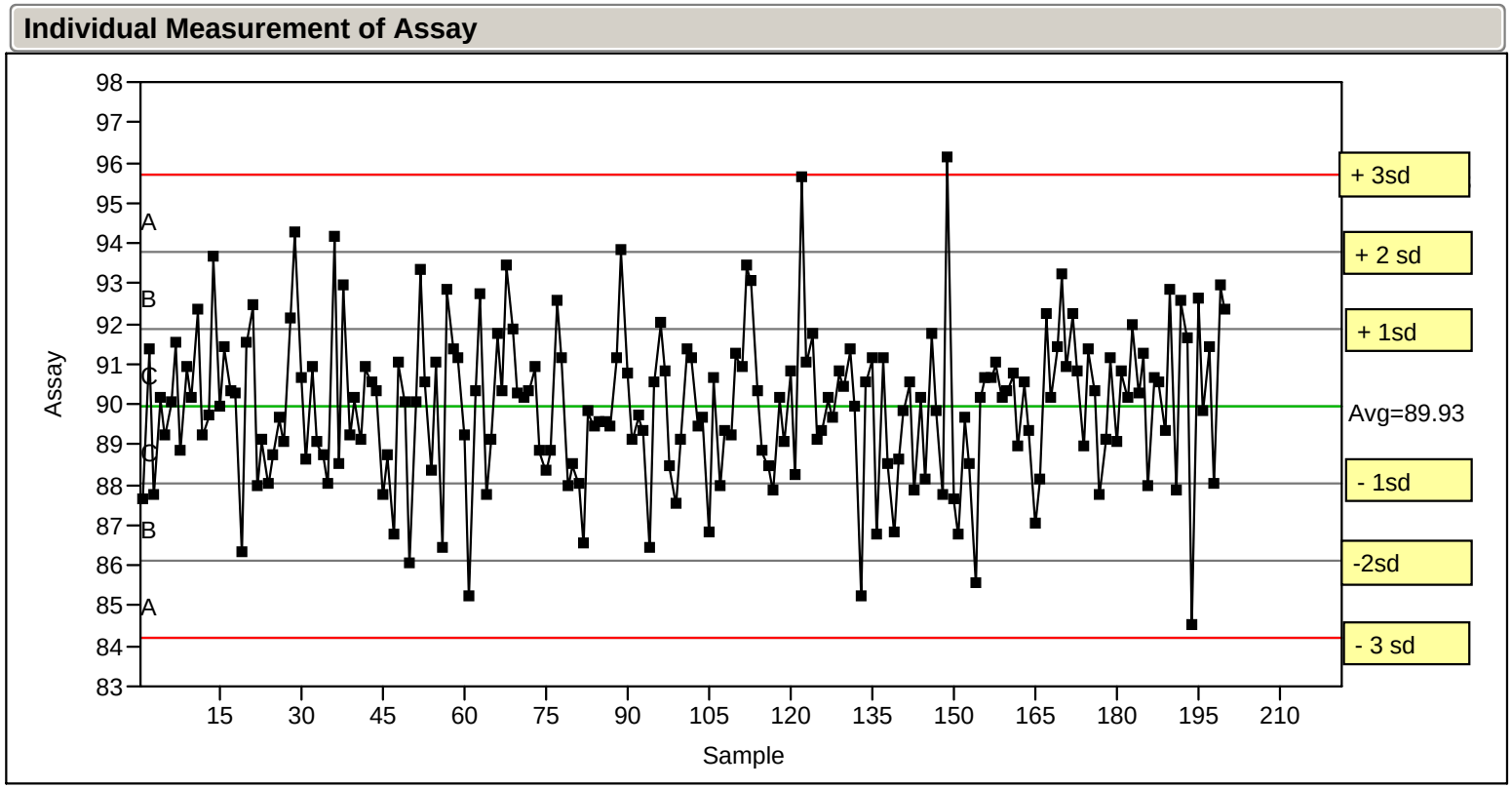
syngenta

The Normal Distribution and it's Properties



syngenta

Shewart Control Chart



+/- 3 sd : ACTION LINES, +/2 2 sd: WARNING LINES

Out of Control Signal: 1 point outside of 3 sd limits (p = 3 in 1000)

2 points outside of 2 sd limits (p = 1 in 400

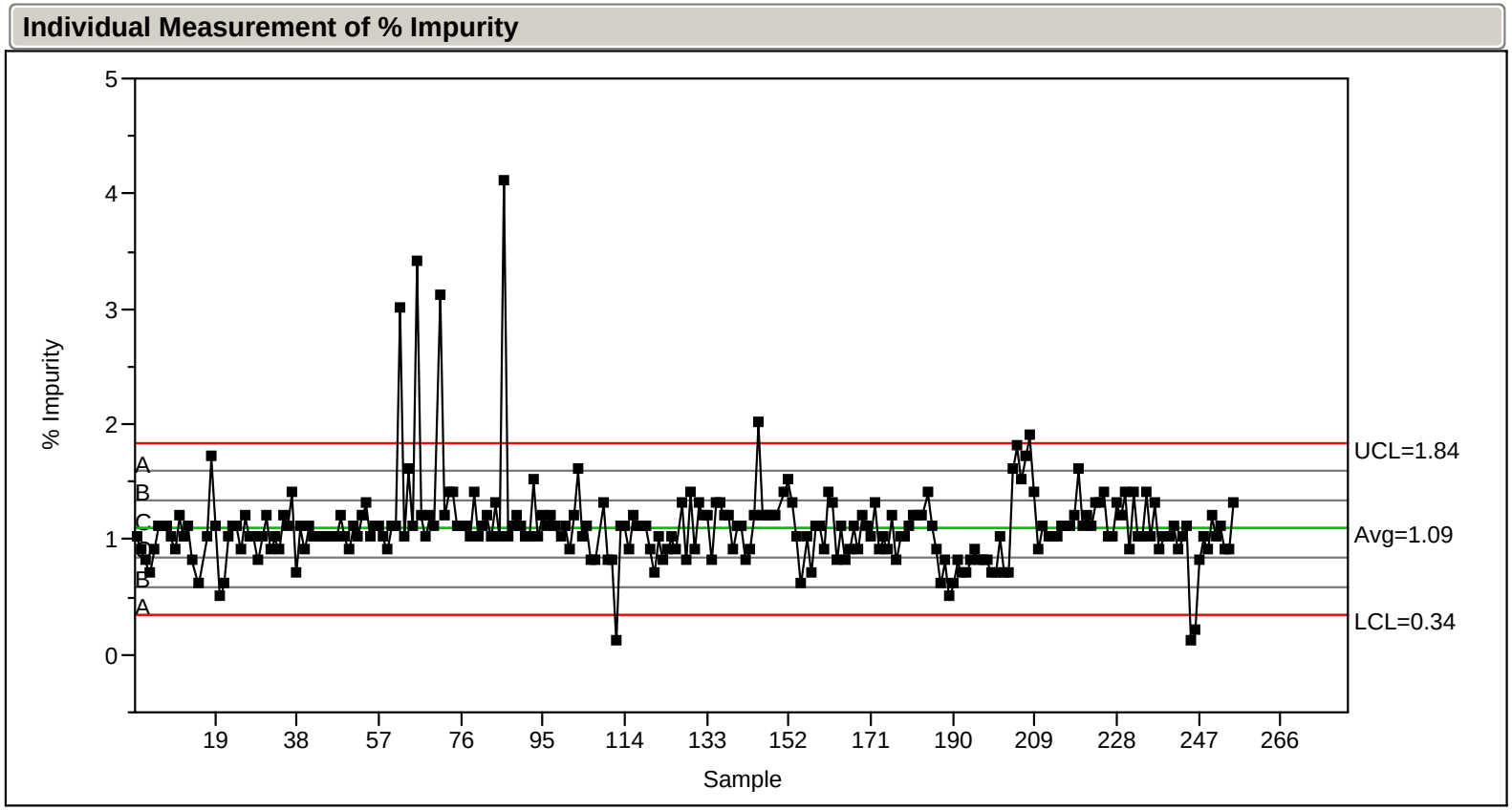
9 points in a run above or below mean (p = 1 in 512)

Control Charts

- The Shewart Chart is one of a series of charts used to detect non random behaviour in a data set
  - If no out of control signals are detected the process is said to be in Statistical Control
- Other charts include
  - Moving Average
    - Smoothes the data to allow detection of changes
  - Exponentially Weighted Moving Average
    - Similar to above – earlier points given a lower weighting
  - Moving Range
    - Detect a change in variability
  - CUSUM
    - Cumulative sum
      - Sum the difference between an observation and a target value

Shewart Chart - Example

250 batches - % of a specified impurity



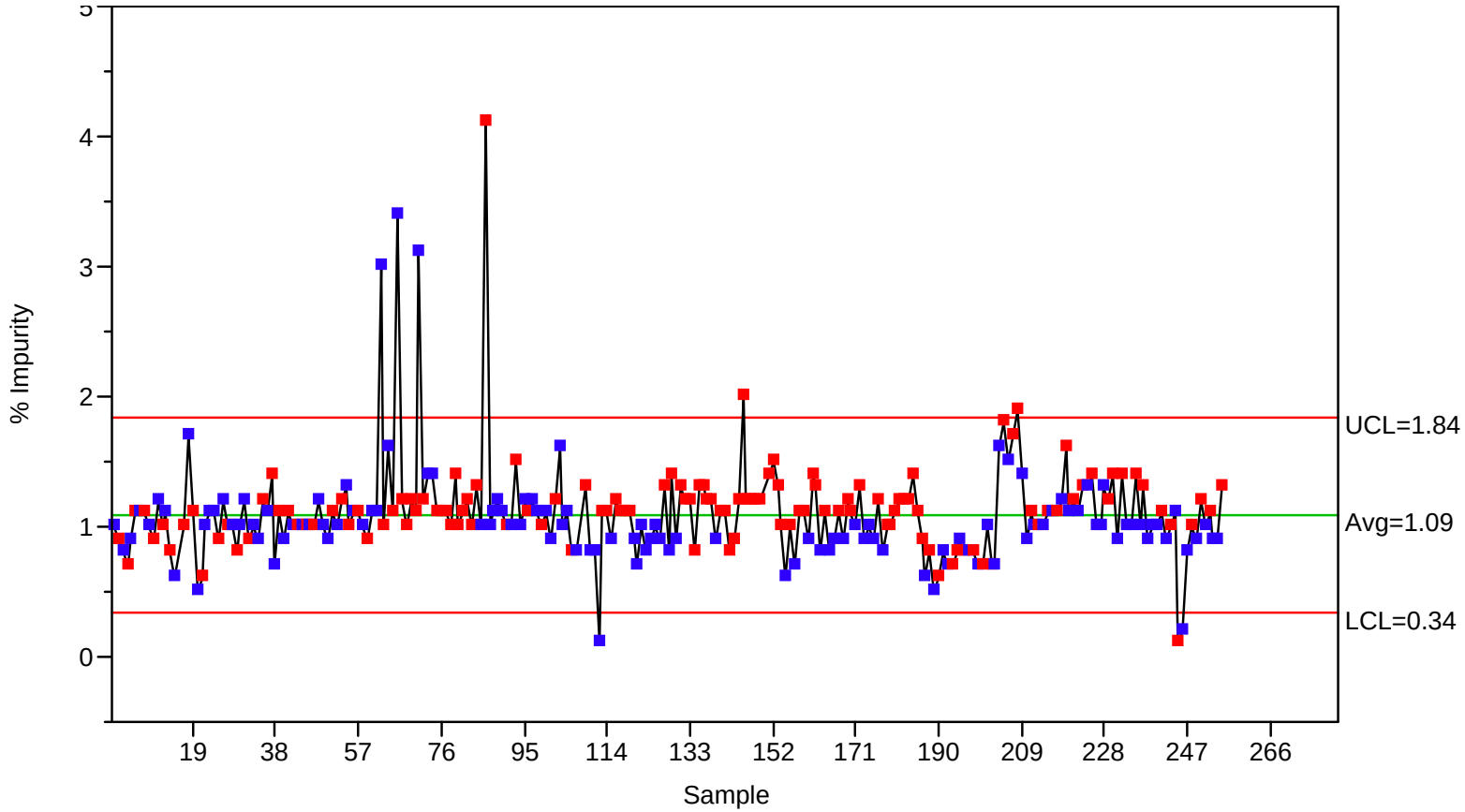
Information: Process is clearly not under Statistical Control

A number of single batches out of control

Runs above and below mean

Shewart Chart - Example

Process is run in 2 vessels (A = Blue, B=Red)



Information: After ca sample 120 the level of impurity appeared to be vessel dependent

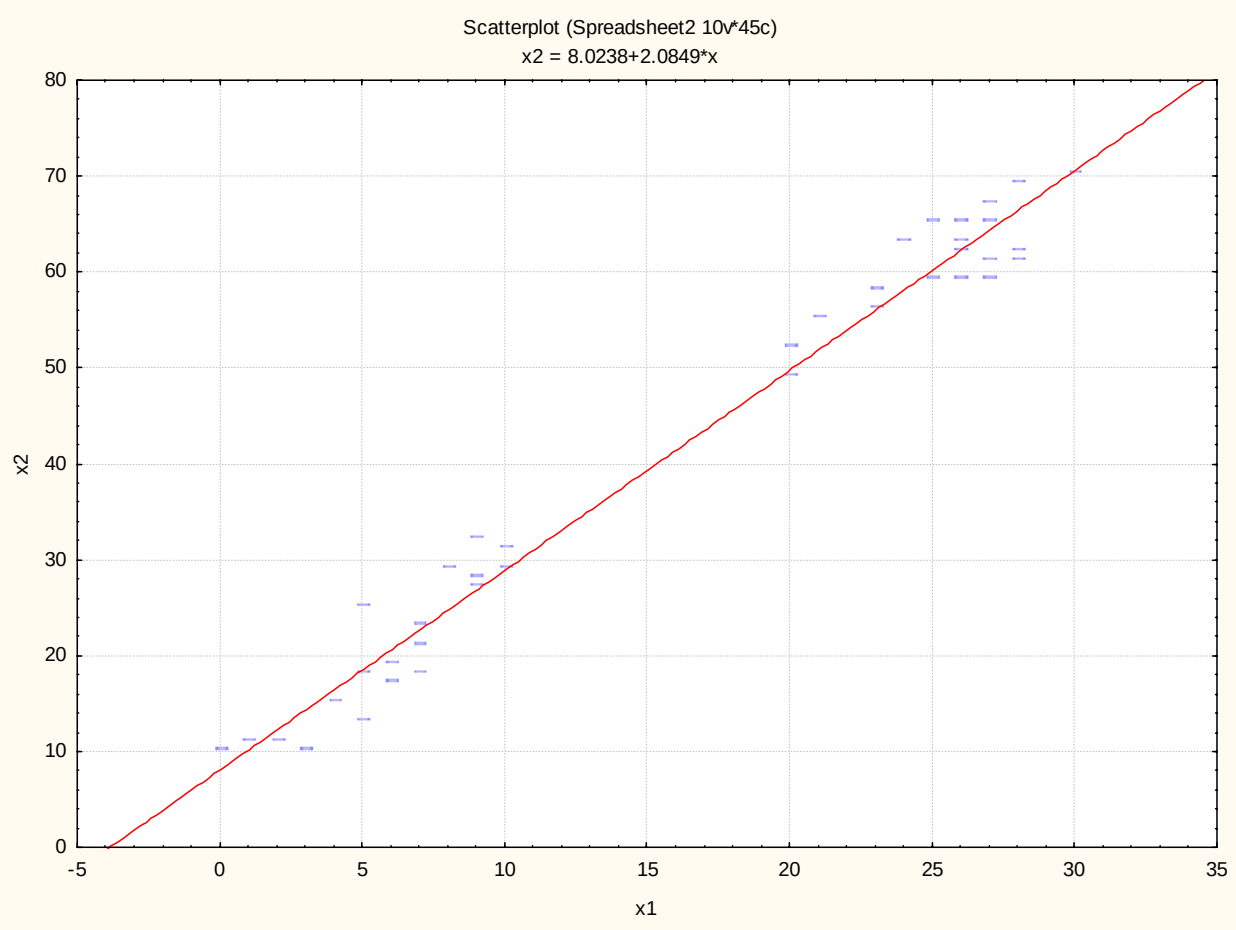
Cause?

syngenta

Example- Multivariate Analysis

- The Shewart Chart has identified a potential problem
  - Further data analysis is required to establish the likely cause
- In addition to the impurity data, observations were available for 5 further variables
  - Maximum temperature attained during the reaction
  - Time spent above 80 deg C
  - The amount of water removed from the vessel
  - The isolated yield
  - The assay of the desired product.
- Required to examine the relationships between the six variables and determine where the differences occur between the 2 vessels.

Relationship between variables - Two Variable Problem



The plot shows a clear relationship between the two variables: They are highly correlated r=0.99.

The observations appear to fall into two clusters

syngenta

Relationship between variables

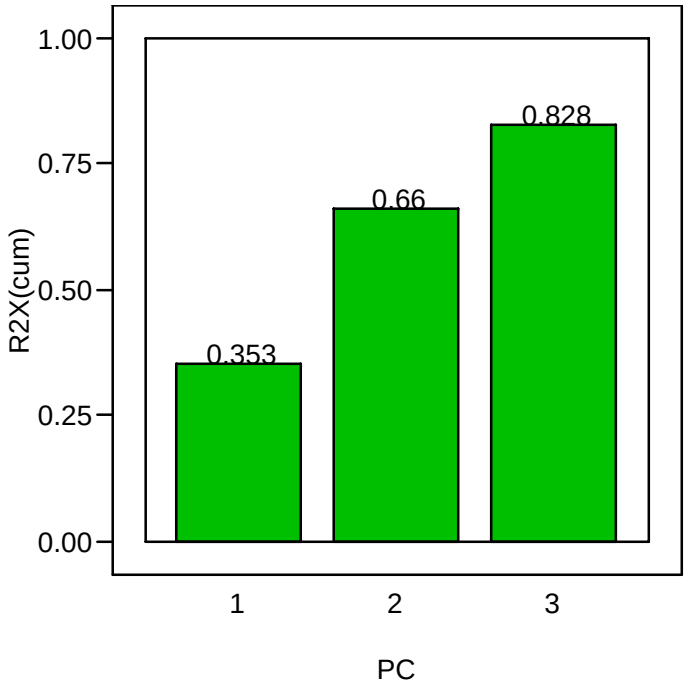
- Example requires investigation of the relationships between 6 variables
  - Require a 6x6 correlation matrix
  - 15 pair-wise plots
- An alternative approach, if the variables are correlated is that of Principal Component Analysis (PCA)
  - A principal component (PC) is a new variable constructed from a weighted sum of the original variables. The weightings are chosen such that the first PC accounts for the maximum variance in the data set
  - A second PC is then formed, with different weights, to account for the next highest variance portion
  - A total of 6 (in the example) PCs can be constructed each one explaining a decreasing amount of the variance in the data
  - The PCs are not correlated with each other
  - Most of the variance in the data is usually explained in the first few PCs

Principal Component Analysis

The output from PCA produces 2 key plots:

- The SCORES Plot shows the relationship between the observations.
  - The score for a specified PC is the value of the “New “ variable for each of the observations
- The LOADINGS Plot shows the relationship between the original variables
  - The loading of the PC is the weighting given to an individual variable when forming the PC

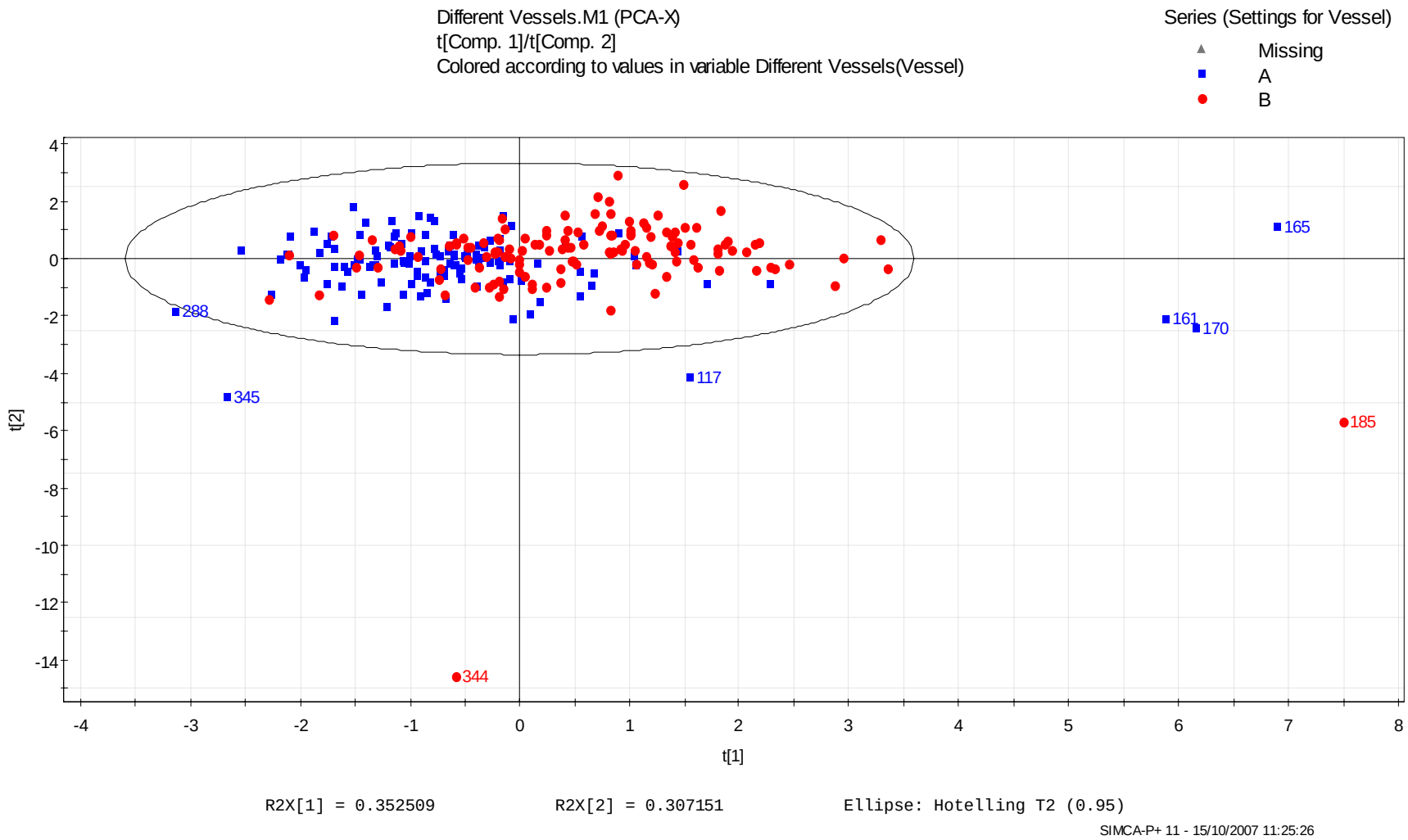
Example



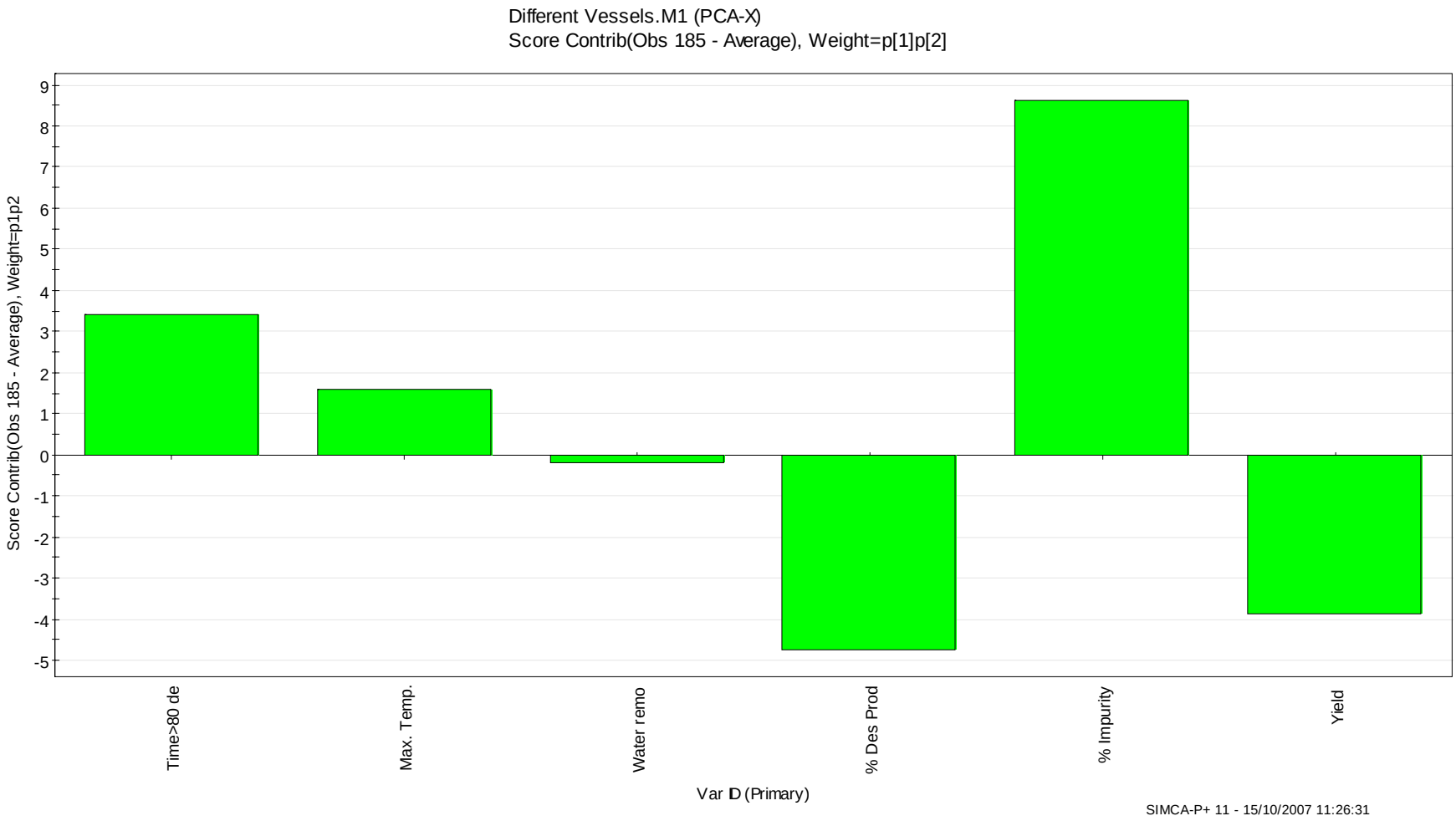
3 PCs account for ca 83% of variance

syngenta

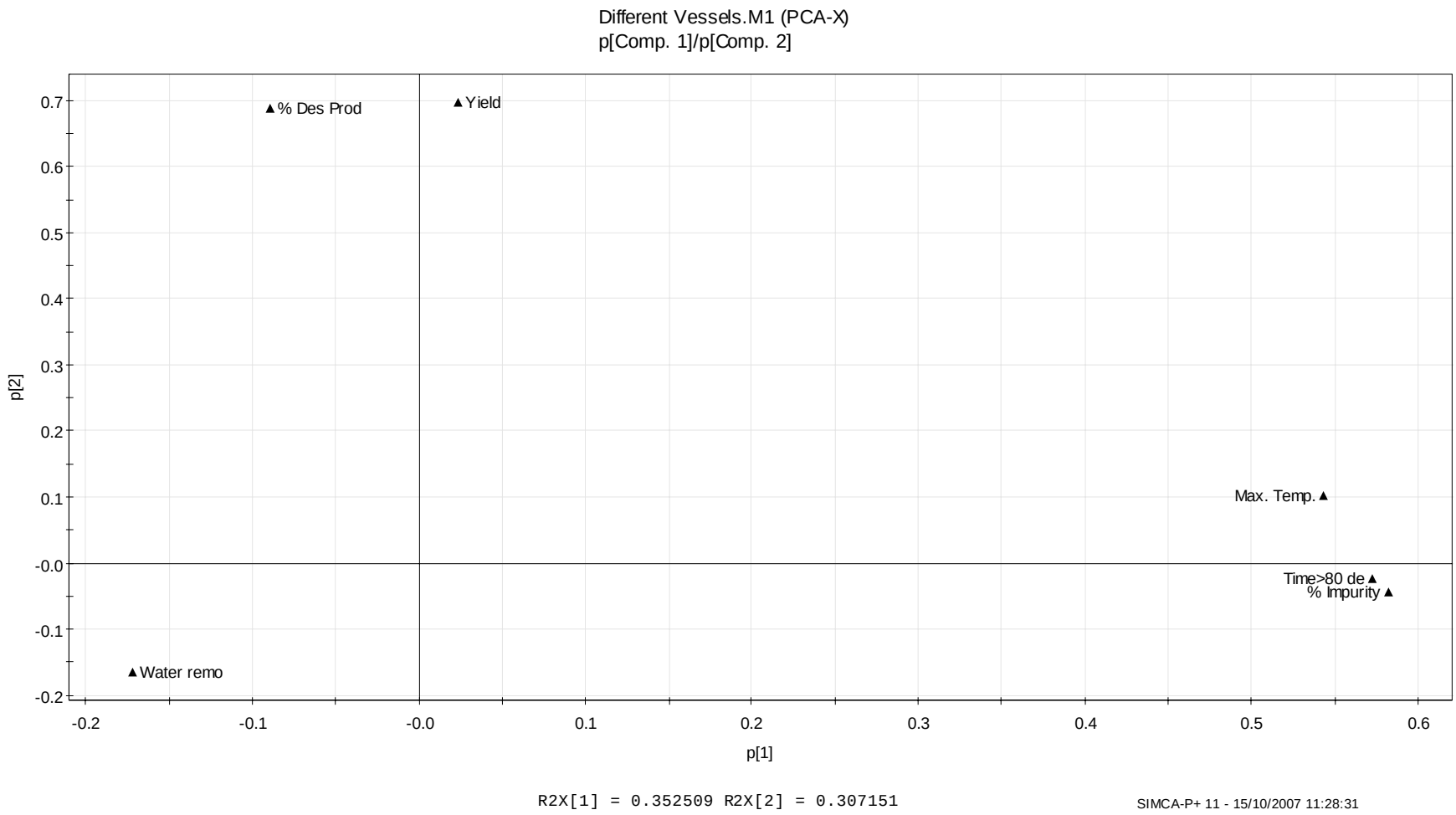
Scores Plot (Coloured according to vessel)



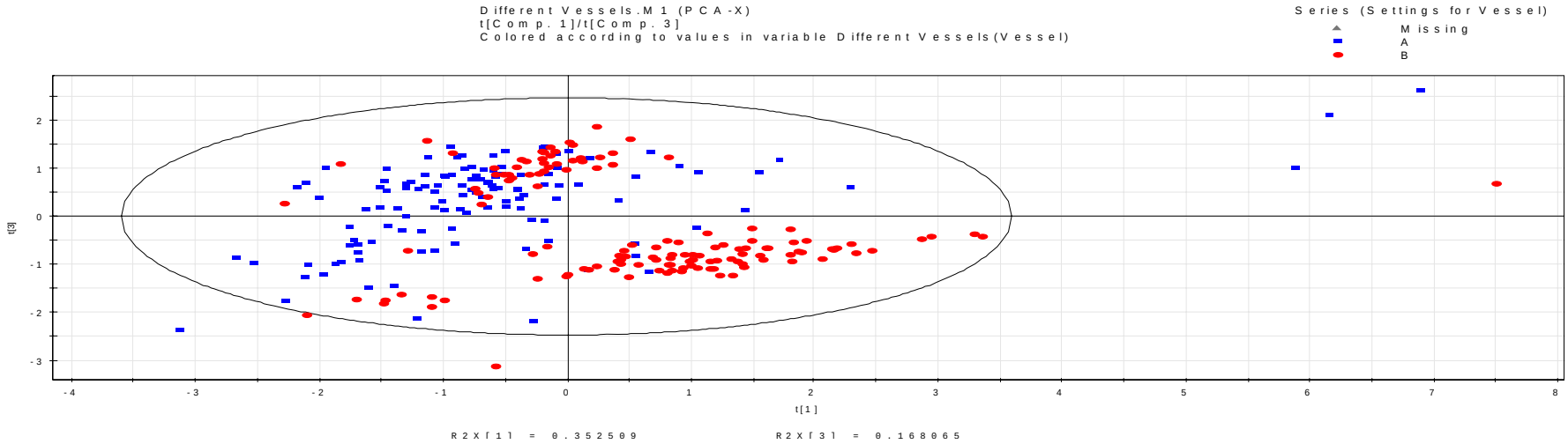
Why is Batch 185 Different?

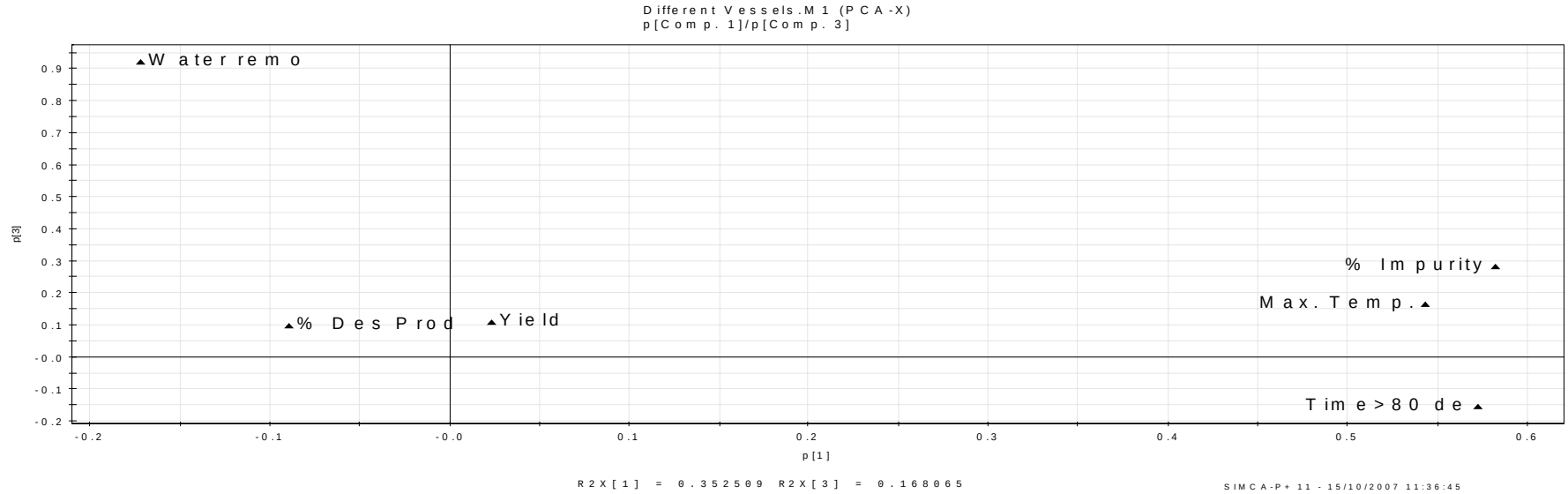


Loadings Plot



Scores and Loadings Plot t1/t3





syngenta

Example - Conclusions

- The two vessels differ in the quality of product produced
  - More impurity produced in Vessel B compared with A
  - Higher maximum temperature and time >80deg C in vessel B

- The performance in vessel B has changed with time. The amount of water removed is less for later batches
  - Level Indicator on water receiver damaged during PM and not replaced!!!!!!
  - Reaction driven to remove water led to higher max temp, higher time above 80Deg C and consequently higher impurity level

Fault Detection using PCA

- Shewart Chart used to track the performance of a single variable, eg an output variable – a univariate approach
- Many processes have measurements on multiple input variables – a multivariate problem
  - Often correlation between variables
- The important information does not, necessarily, lie in the univariate system but within the correlation structure of the data
  - Univariate charts of, potentially, limited use
- Apply Shewart Charts to PC scores

Fault Detection using PCA

Example

- A batch process is operated in two streams. Data is collected for 13 variables at 1 minute time intervals over the course of the batch. 10 batches are available for each stream. Approx 8000 rows of data

- 13 Shewart charts, for each stream, could be generated
  - Difficult to track
  - Does not take into account the correlation within the data
- Apply PCA to the data and track with a Shewart Chart for each (major) PC

Shewart Chart for PC1 (Accounts for 88% of time varying data) for Stream 2

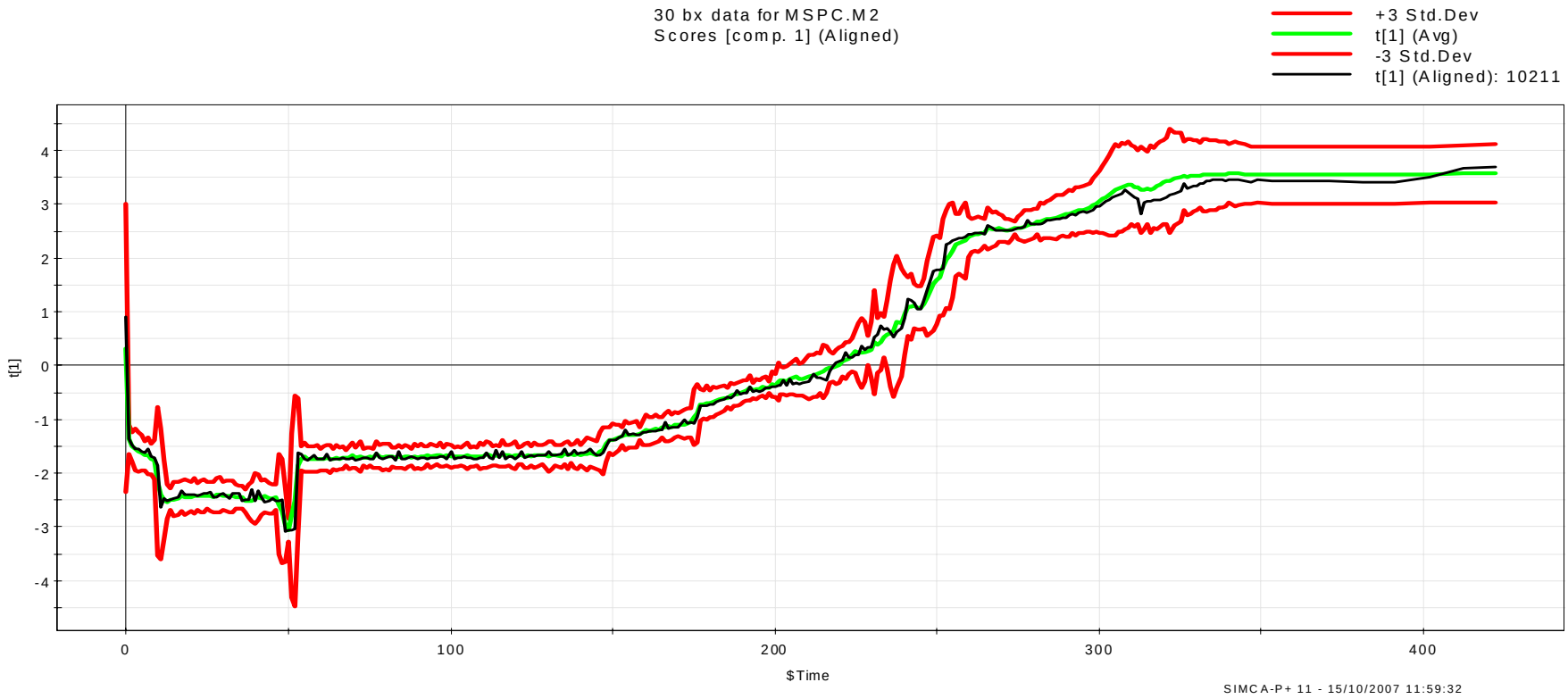
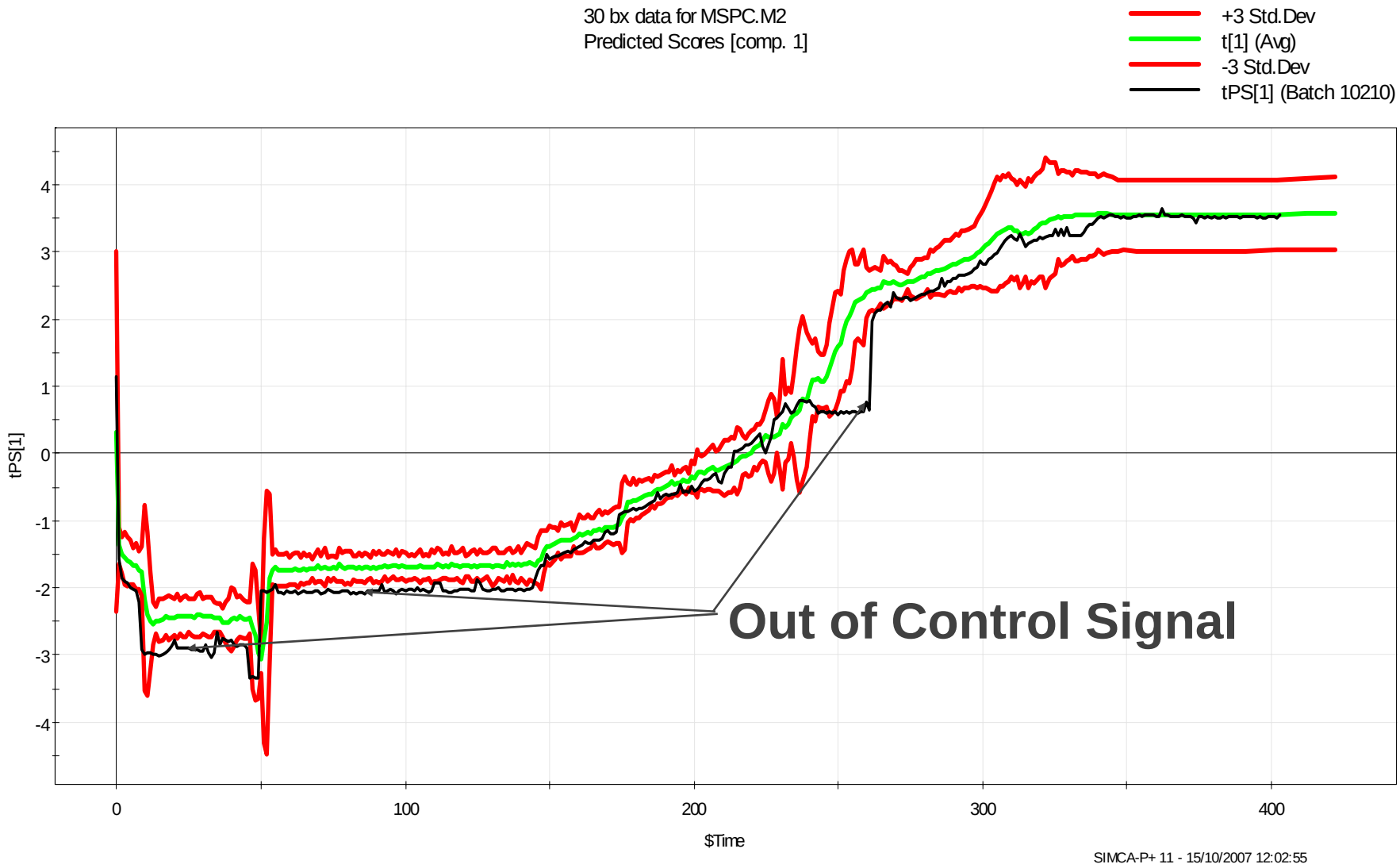


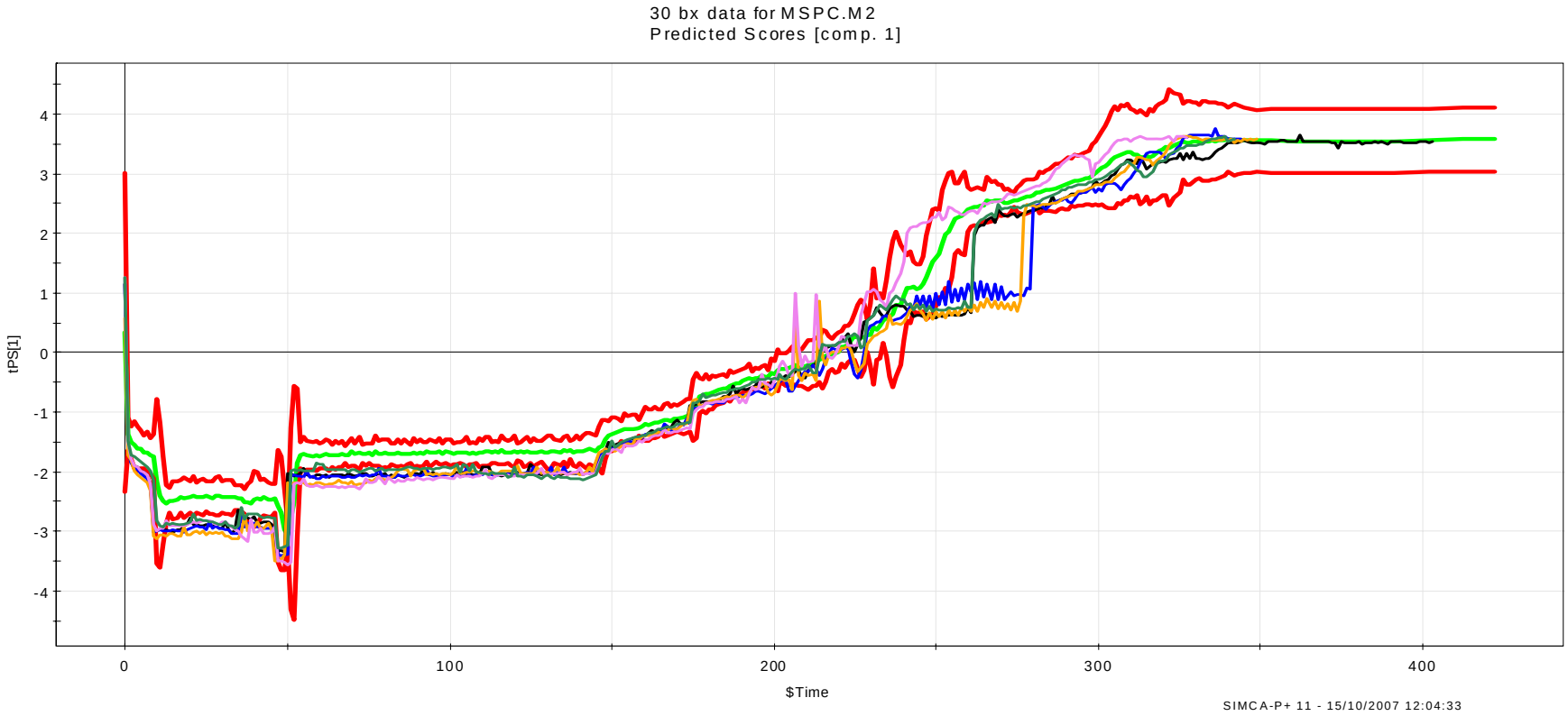
Chart shows the phases of the batch reaction: Charge/Heat, hold and distill

Data for Stream 1 projected onto Stream 2 Chart



syngenta

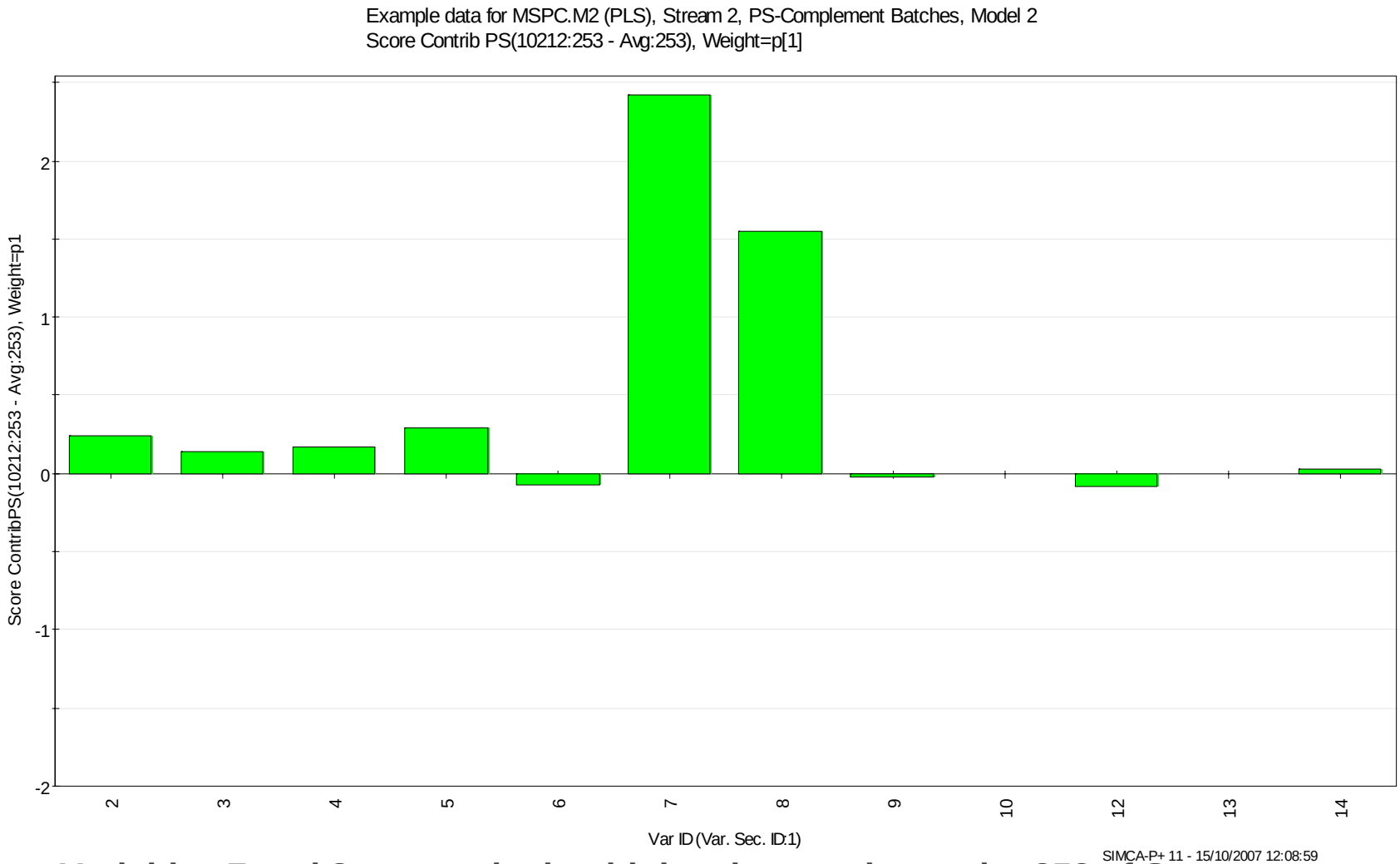
Additional Stream 1 Batches



Stream 1 is behaving different to Stream 2 for large parts of the Process.

syngenta

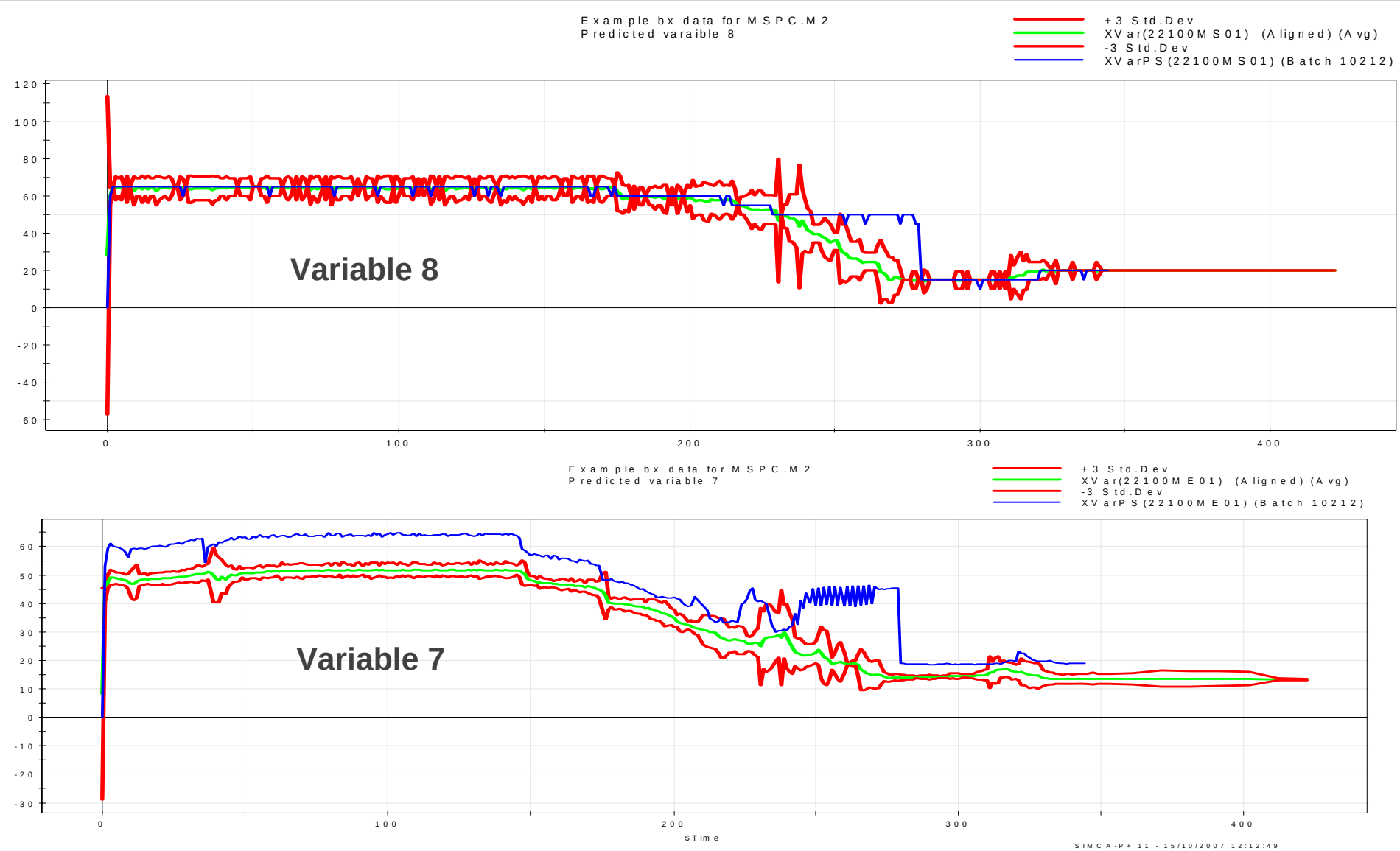
Out of Control Signal time point 253



Variables 7 and 8 are producing high values at time point 253 cf Stream two average batch

syngenta

Out of Control Signal time point 253



Other Stream 1 batches behaved in a similar way

Fault Detection using PCA

- Stream 1 batches exhibit different behaviour cf Stream 2 for PC1 scores
- Variable 8 is similar in both streams except after time point 253
- Variable 7 is higher all through the batch

- Var 7 = Agitator Power; Var 8 = Agitator Speed
  - Stream 1 agitator requires more power to maintain the same speed
    - Agitator out of alignment!!

Statistical Modelling of Process Data

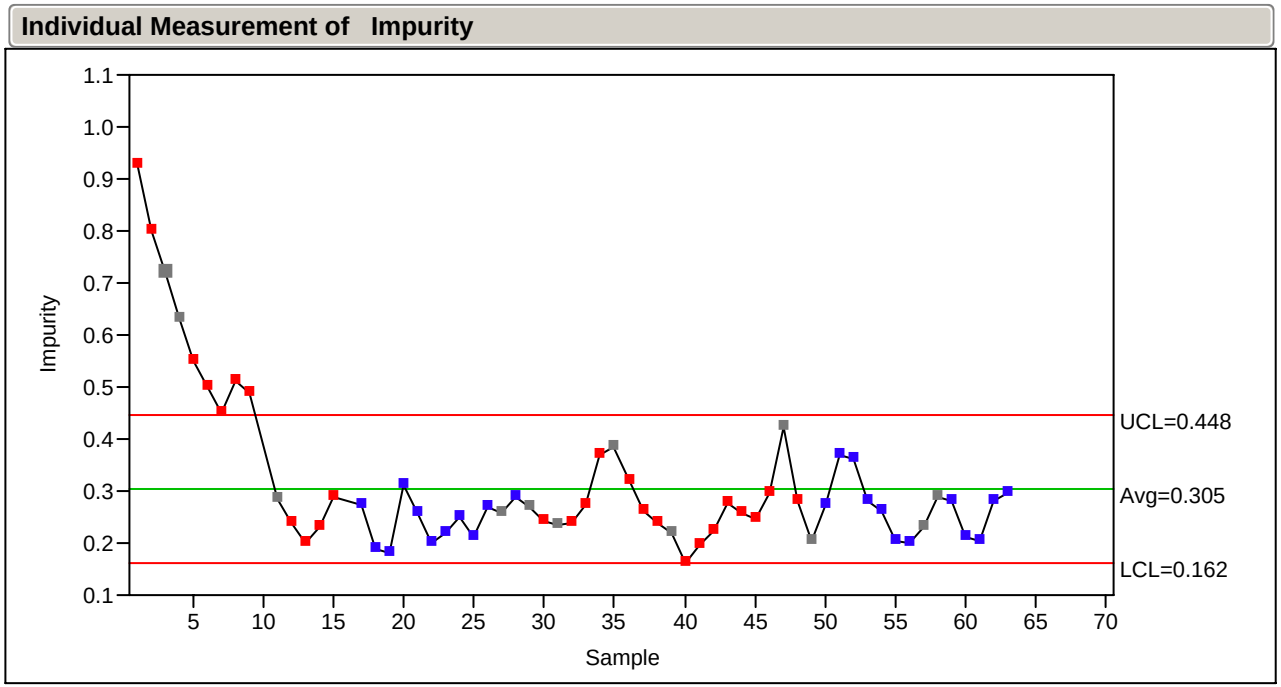
- Examples involved fault detection
- Data Analysis can enable much more – eg how can we relate process variables to a desirable product property eg level of an impurity, ability to formulate.
- The standard way of relating a response to a series of predictor variables is by way of multiple linear regression, eg in fitting data form SED
  - however, assumes the predictor variables are independent of each other
  - Much of the data generated from process operations is correlated, therefore need a regression technique that does not make the assumption of independence of the predictors
  - Partial Least Squares (PLS) Regression is such a technique

Partial Least Squares Regression

- PLS is a latent variable method with a number of similarities to PCA
  - In PCA components are extracted from a data matrix such that they maximise the variance explained
  - In PLS components are extracted from a matrix of predictor variables such that they maximise the covariance with the response variable
    - The components are weighted combinations of the original variables
    - A series of regression coefficients is determined which can be used to understand the magnitude and direction of the effect on a response of a particular predictor

PLS Example: Prediction of impurity in a product

- A product is made by a continuous process. The amount of a specified impurity in the product varies with time.



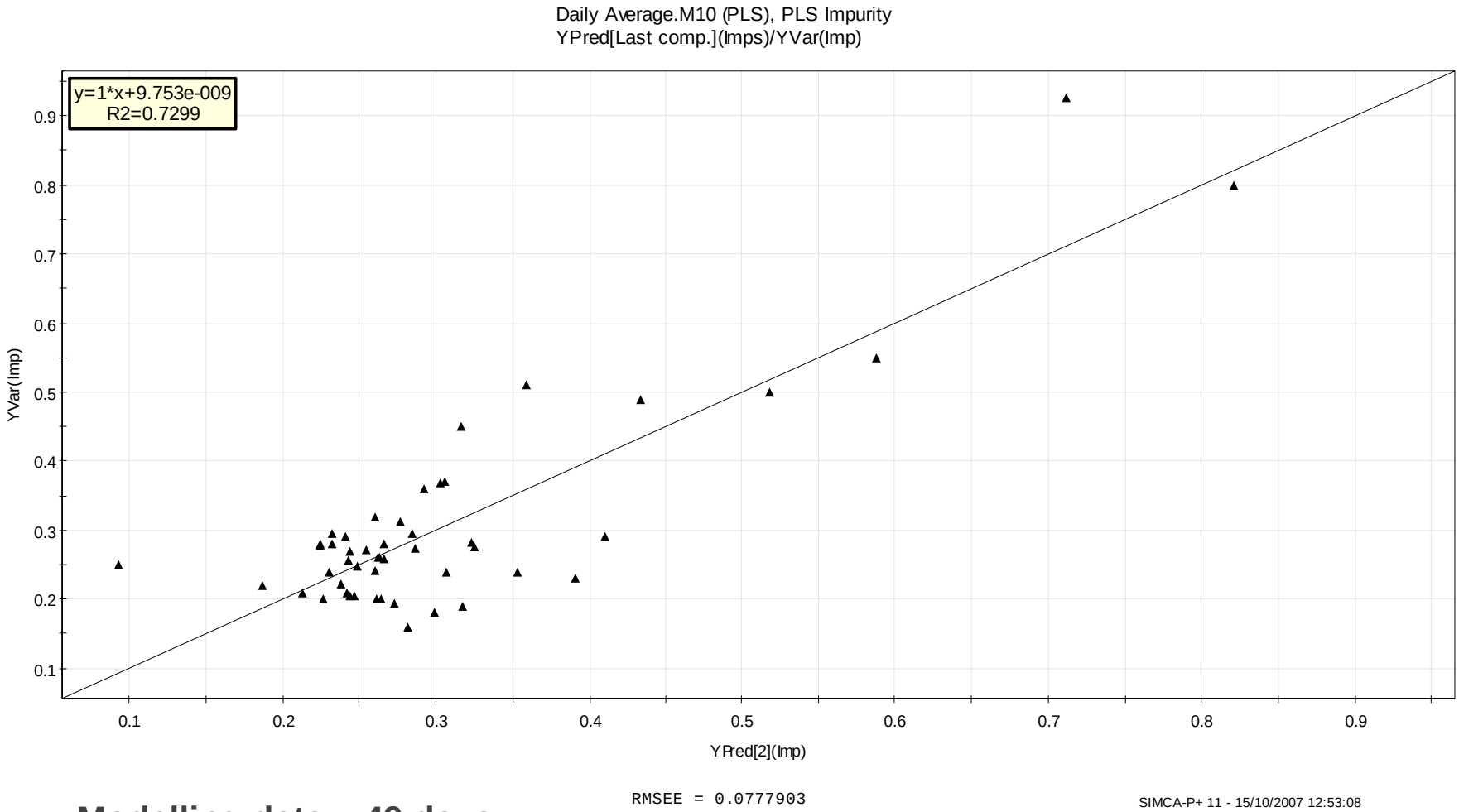
Note: 12 samples were excluded.

- 31 process variables are measured on the plant

PLS Example: Prediction of level of impurity in a product

- Data for 61 days were collected. The daily average for each of the 31 process variables was used for the modelling. Data for 12 of the days (chosen at random) were excluded. (These were later used to validate the model)
- A 2 component PLS model accounted for 73% of the variance associated with the level of impurity

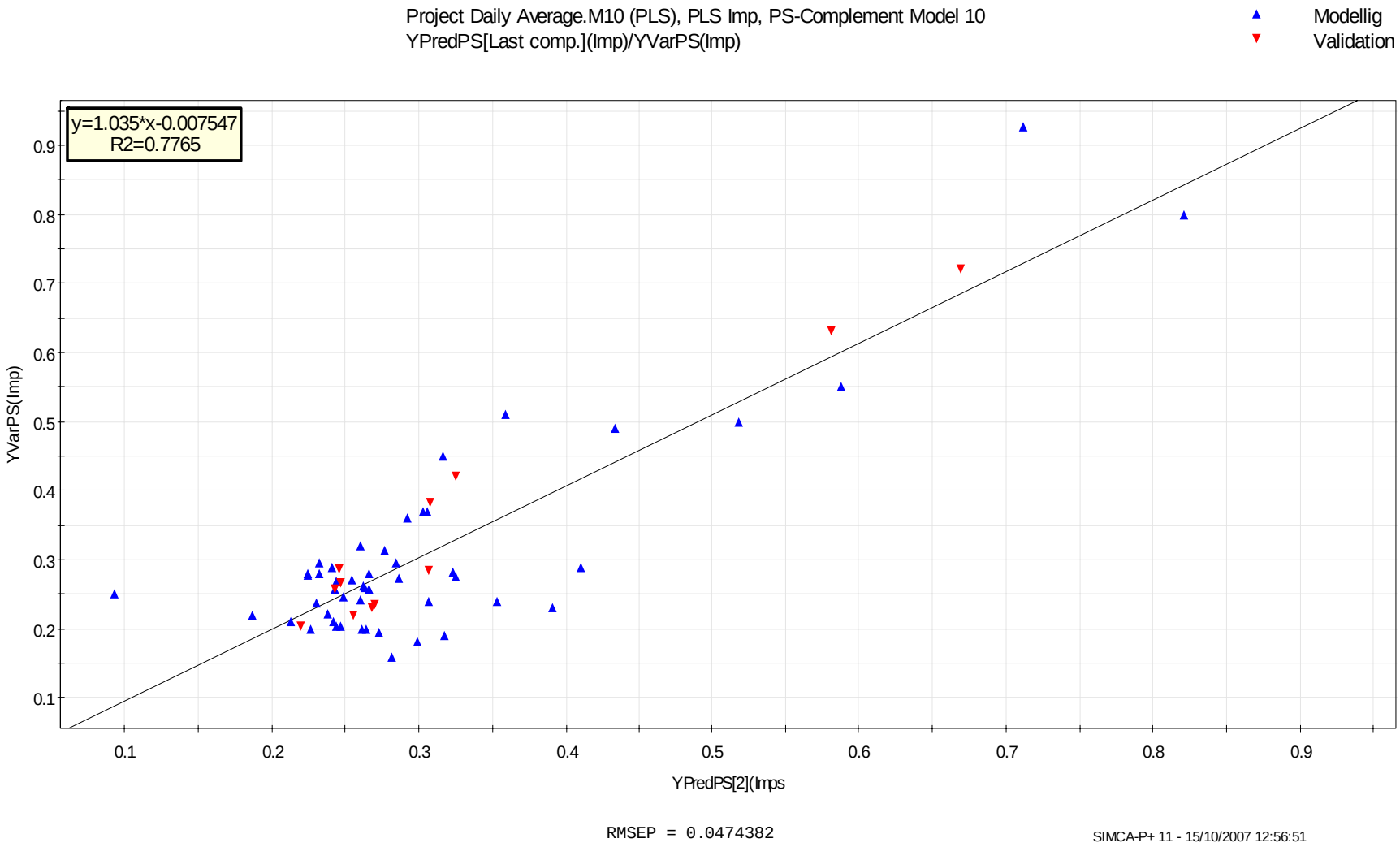
PLS Example: Prediction of level of impurity in a product



Modelling data – 49 days

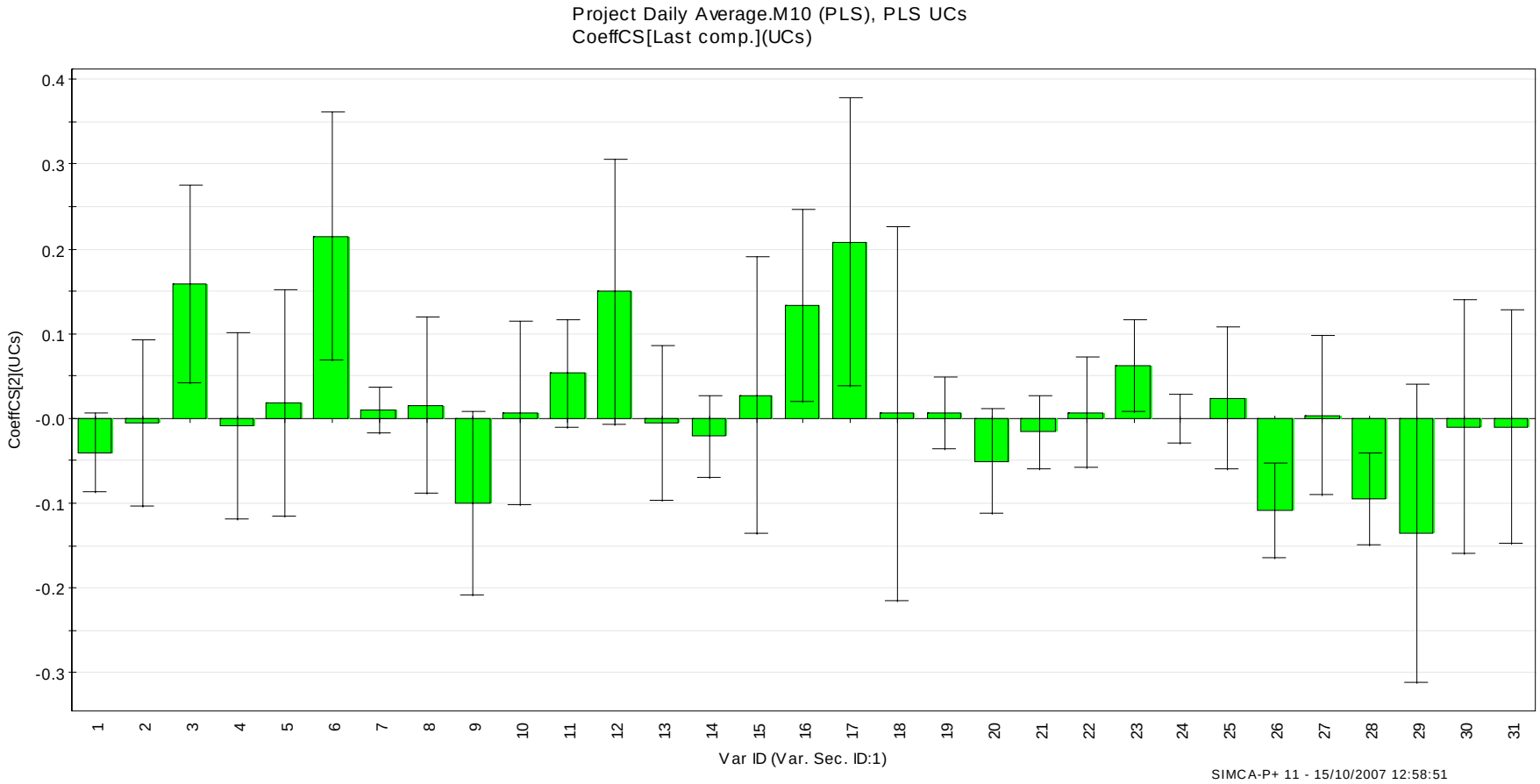
syngenta

PLS Example: Prediction of level of impurity in a product



syngenta

PLS Example: Prediction of level of impurity in a product



Information: Important variables related to - moisture analysers, temperatures and differential pressures

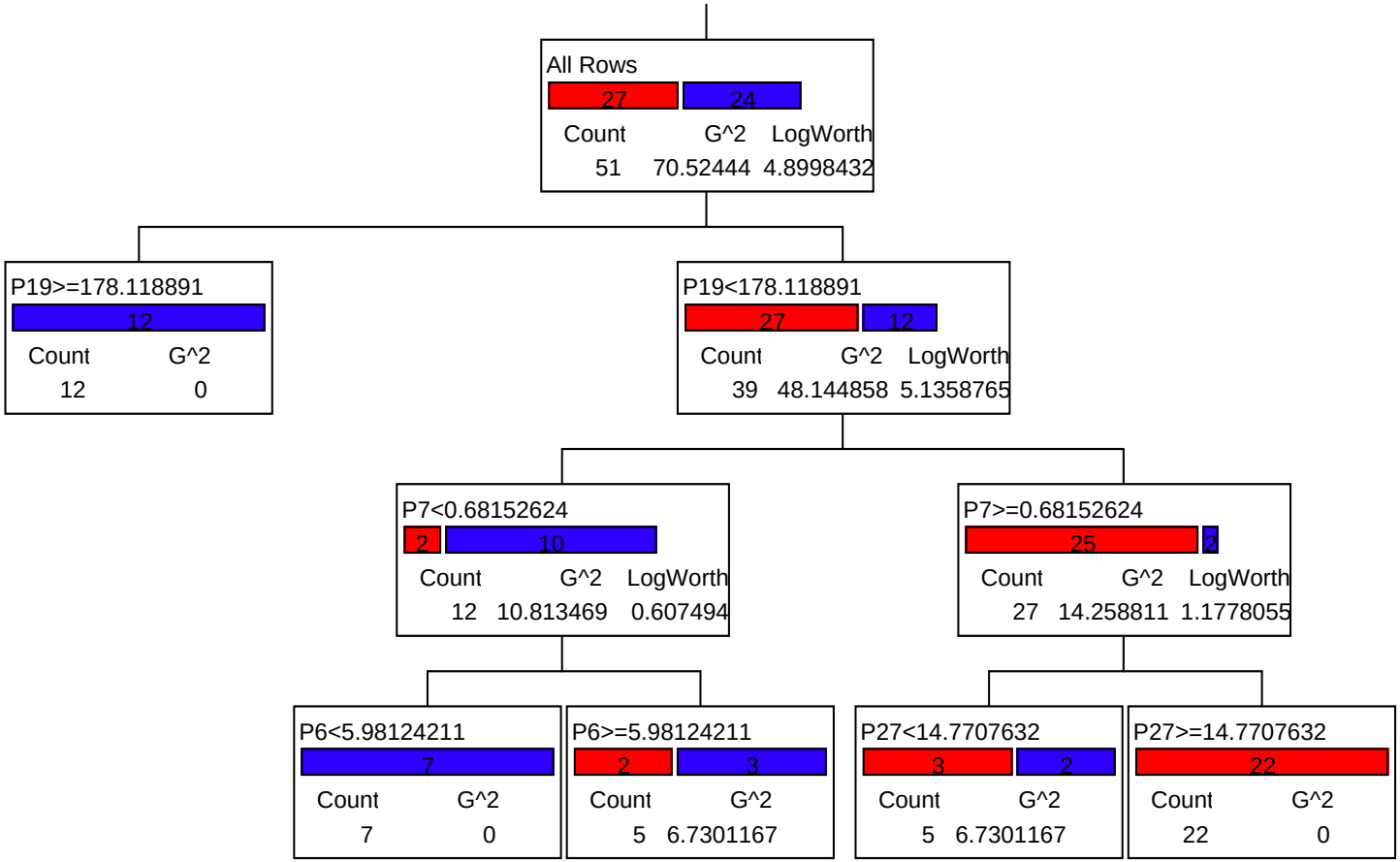
Knowledge: Interpreted by chemical engineers

syngenta

Recursive Partitioning - Classification Tree

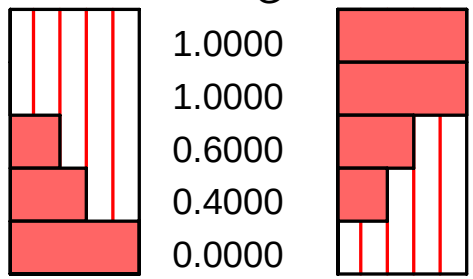
- This technique recursively partitions data according to a relationship between the *X* (predictor) and *Y* (response) values, creating a tree of partitions. It finds a set of cuts or groupings of X values that best predict a Y value. It does this by exhaustively searching all possible cuts or groupings. These splits (or *partitions*) of the data are done recursively forming a tree of decision rules until the desired fit is reached
- In the following example the product is classed as fit to formulate (Good, Blue) or not fit to formulate (Bad, Red)
- The daily averages of 31 process variables are used as predictors
- 61 days of data are available. 12 days are excluded for validation of the model

Classification Tree



Classification Tree – Decision Rules

Leaf Label	B	G
P19>=178.118891	0.0000	1.0000
P19<178.118891&P7<0.68152624&P6<5.98124211	0.0000	1.0000
P19<178.118891&P7<0.68152624&P6>=5.98124211	0.4000	0.6000
P19<178.118891&P7>=0.68152624&P27<14.7707632	0.6000	0.4000
P19<178.118891&P7>=0.68152624&P27>=14.7707632	1.0000	0.0000



80% of validation data correctly classified

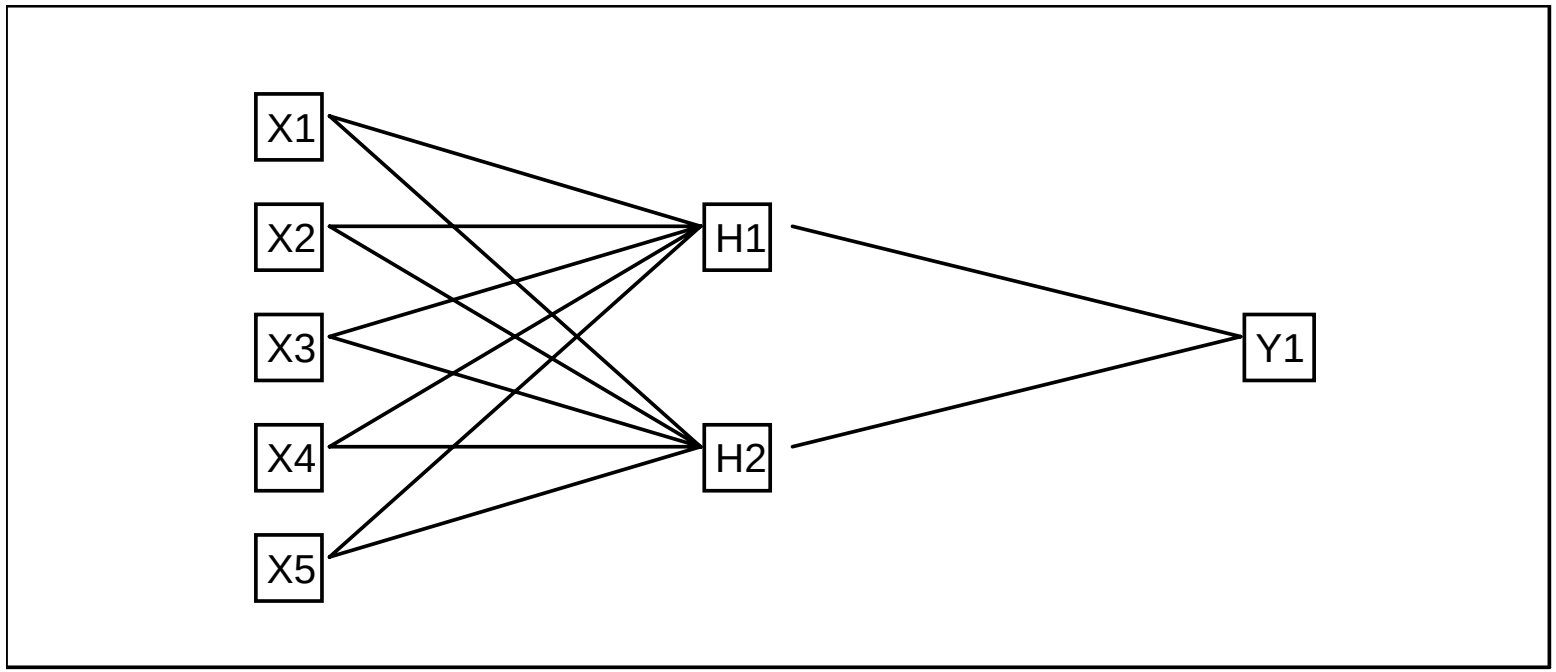
Information

Key variables related to moisture, distillation unit temperatures and plant throughput.

Knowledege

Interpretation by Process Chemists/Engineers

Artificial Neural Network (ANN)



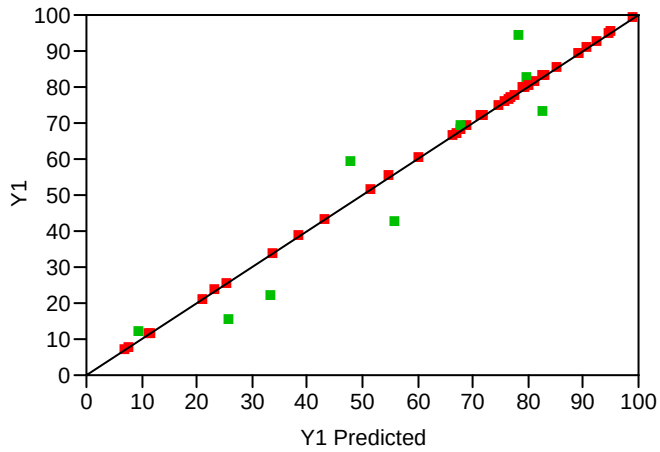
ANN relates a series of input variables, X, to one or more response variables ,Y

The hidden layer consists of a number of constructed variables (nodes) each of which is a sum of weighted contributions of the original variables. These constructed variables are then related to the response variable, usually via a non linear function

Artificial Neural Network (ANN)

Example: 31 process variables, daily average from a continuous process. 51 days production, 9 held back to validate the model. Used to model the amount (%) of a particular polymorph in the material

An ANN model with 3 nodes in the hidden layer was used



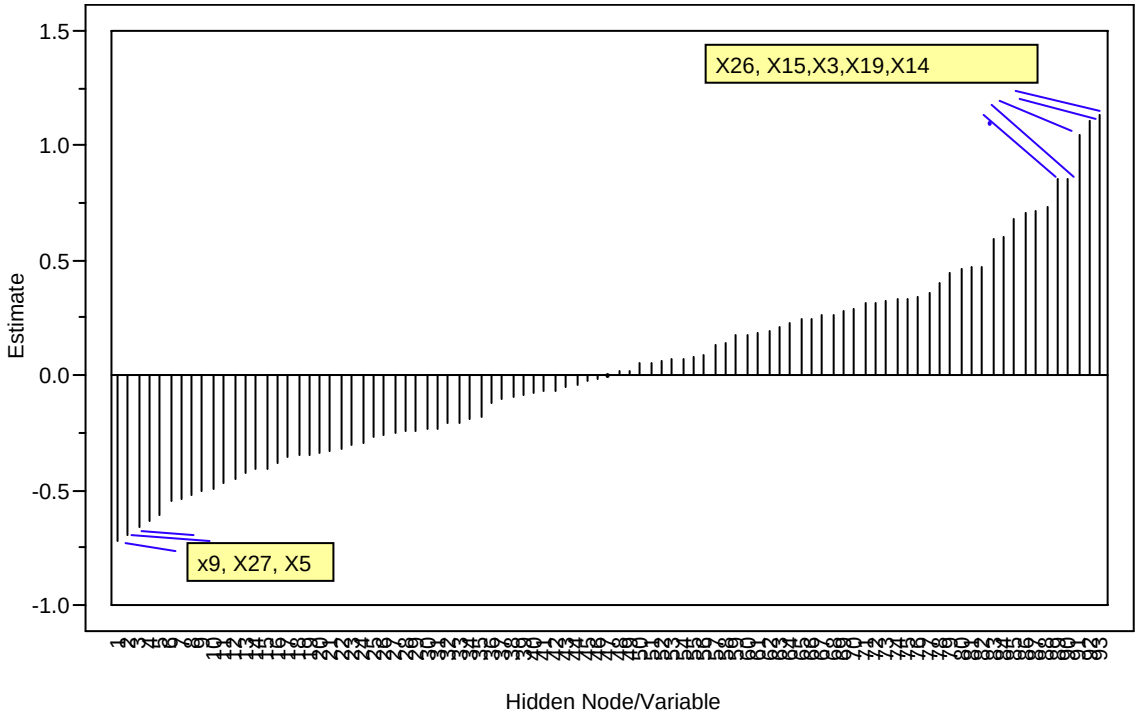
Red – Model Data

Green – Validation Data

syngenta

Artificial Neural Network (ANN)

Weightings can be used to help identify the important variables



Caution: Very easy to over fit ANN models - essential to use a validation data set

syngenta

Summary

- Most organisations are data rich but probably information poor
- Application of simple graphical tools can help identify when a process is no longer operating with only random variation
- Process Data is usually highly correlated
  - Enables the use of techniques such as PCA to aid in fault detection
- Other statistical modelling techniques (eg PLS, ANN and Classification trees) can be used to extract information from data sets
  - Interpreted in terms of the process chemistry