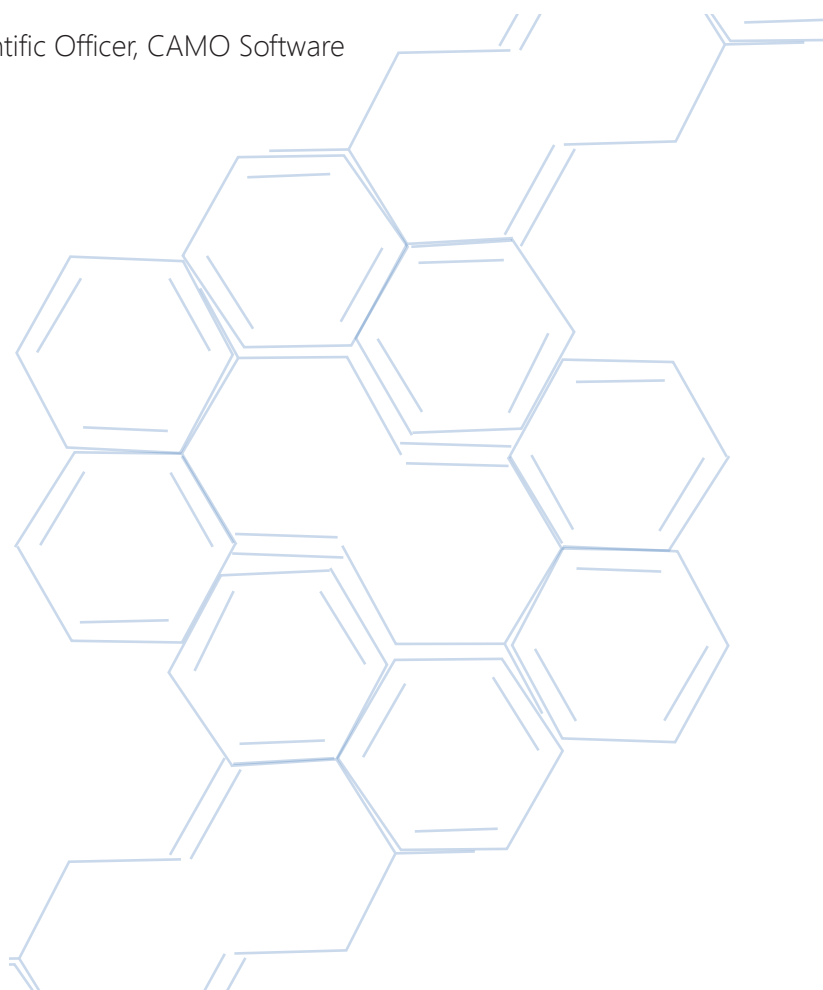


Monitoring chemical processes for early fault detection using multivariate data analysis methods

by Dr Frank Westad, Chief Scientific Officer, CAMO Software



Makers of

The Unscrambler[®] 

CONTENTS

Executive summary	03
About CAMO Software	03
Introduction to Multivariate Process Monitoring and Control	04
Figure 1: Comparing Multivariate and Univariate views of a process	05
Common multivariate methods and statistics	06
Industry example: Background and Data	07
Table 1: Variables and product attributes in a paper manufacturing process	08
Figure 2: The Scores plot to visualize samples based on all variables	09
Figure 3: Line plot of Scores to visualize process trends	10
Figure 4: Weighted coefficients showing impact of variables on the model	10
Figure 5: Predicted values in real-time process monitoring	11
Figure 6: Hotelling T ² statistic	12
Figure 7: Drill down diagnostics for out of limit variables	12
Process optimization	13
Figure 8: Optimizing process settings within given constraints	13
Chart 1: Closed loop process improvement	14
Summary	14
CAMO Software products and services	15

EXECUTIVE SUMMARY

Multivariate statistical methods can be used to monitor process variables and predict final product quality at an early stage, while also providing deeper understanding of the process. This allows engineers and production managers to optimize their processes, thereby realizing significant cost and time savings.

This white paper includes a background and explanation of some of the key multivariate methods, as well as examples of how to interpret typical multivariate plots. It uses a real-world example from a paper manufacturing company that was able to improve a key quality parameter, *Print Through*, by better understanding the variables impacting it.

Additionally, the company was able to optimize the process going forward, by adjusting the process inputs to find the "sweet spot" where the significant variables affecting quality were all within the acceptable limits.

ABOUT CAMO SOFTWARE

Founded in 1984, CAMO Software is a recognized leader in Multivariate Data Analysis and Design of Experiments software. Today, we have offices in Norway, USA, Japan, India and Australia. Multivariate analysis is a powerful set of data mining techniques that help identify patterns and understand the relationships between variables in large and complex data sets.

Our software is used by many of the world's leading companies, universities and research institutes in the life sciences, food & beverage, agriculture, energy, oil & gas, mining & metals, industrial manufacturing, pulp & paper, automotive, aerospace and technology sectors.

The Unscrambler® X software range is a preferred choice of engineers, scientists and data analysts because of its ease of use, world-leading analytical tools and data visualization. Our solutions are used by more than 25,000 people in 3,000 organizations to analyze data, monitor process or equipment performance and build better predictive models. This gives them valuable insights to make more informed decisions, improve market segmentation, research & development, manufacturing processes and product quality.

INTRODUCTION TO MULTIVARIATE PROCESS MONITORING AND CONTROL

Multivariate Statistical Process Monitoring (MSPM) - also referred to as Multivariate Statistical Process Control or MSPC - is a valuable tool for ensuring reliable product quality in the process industry.

MSPM Multivariate Statistical Process Monitoring

However, many organizations today are still not fully utilizing their potential to make significant improvements in their production environment. The MSPM approach to process monitoring involves the use of multivariate models to simultaneously capture the information from as few as two process variables, up to thousands.

The methodology provides major benefits for process engineers and production managers, including:

- Increased process understanding
- Early fault detection
- On-line prediction of quality
- Process optimization

With MSPM approaches, it is possible to monitor the data at the final product quality stage, but also all of the available variables at different stages of the process, to identify underlying systematic variations in the process.

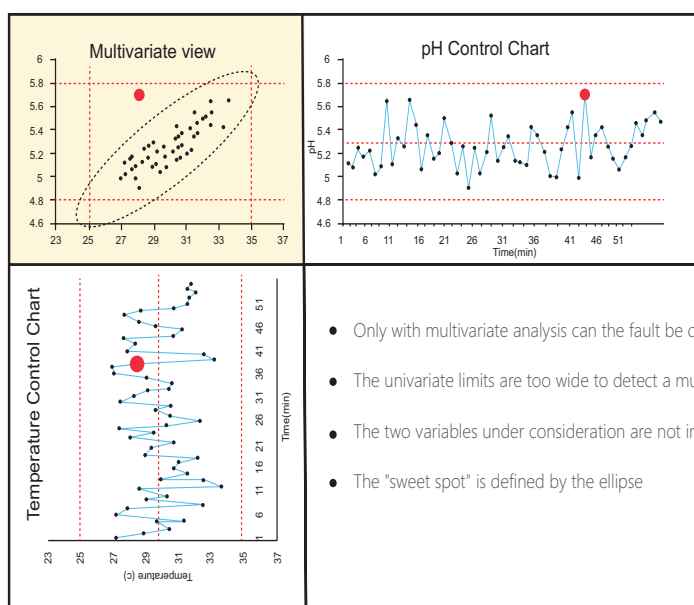
The variables measured in a process are often correlated to a certain extent, for example when several temperatures are measured in a distillation column. This means that the events or changes in a process can be visualized in a smaller subspace that may give a direct chemical or physical interpretation. If we want to keep such a process "in control", traditional univariate control charts - due to the covariance or interaction between variables - may not assure this efficiently. This is because univariate analysis visualizes the relationship to the response variable one at a time and thus does not reveal the multivariate patterns between the variables simultaneously, which both for interpretation and prediction are vital for industrial processes.

Figure 1 exemplifies a typical situation where two process variables are both inside their univariate control limits (given as two standard deviation) but fails to detect that the general trend of correlation between these two variables is broken for the sample shown in red.

Figure 1.

In many processes, the variables have important interactions affecting the outcome (e.g. final product quality) which cannot be detected by traditional univariate statistical process control charts.

Comparing Univariate and Multivariate views of a simple process involving only two variables, Temperature and pH. In this case, the sample appears to be in specification when seen with two separate univariate control charts (Temperature Control Chart and pH Control Chart) but is actually out of specification when seen with the Multivariate view.



Multivariate analysis

Multivariate data analysis (MVA) is the analysis of more than one variable at a time. Essentially, it is a tool to find patterns and relationships between several variables simultaneously. It lets us predict the effect a change in one or more variables will have on other variables. Multivariate analysis methods include exploratory data analysis (data mining), classification (e.g. cluster analysis), regression analysis and predictive modelling.

Univariate analysis

Univariate analysis is the simplest form of quantitative (statistical) analysis. The analysis is carried out with the description of a single variable and its attributes of the applicable unit of analysis. Univariate analysis is also used primarily for descriptive purposes, while multivariate analysis is geared more towards explanatory purposes. Source: Wikipedia

COMMON MULTIVARIATE METHODS AND STATISTICS

The most frequently applied multivariate methods are Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR).

PCA answers the question "Is the process under control?" but does not provide a quantitative model for the final product quality. Typical applications of PCA for this purpose are raw material identification and on-line testing of product quality.

In addition to the monitoring aspect, PLS Regression also provides quantitative prediction of the final product quality based on all or a subset of the process variables. One vital aspect in this context is to reduce the off-line laboratory work, both to have the prediction at an early stage as the product properties are not available on-line, and to reduce the labour-intensive work.

Critical statistical limits can be derived from the empirical data chosen to establish a model for when the process is under control. One limit is based on the space defined by the model, the so-called Hotelling T^2 statistic. This statistic indicates if there is too high or too low concentration of the quality variable of interest. The other limit is based on the distance to the model, meaning there is something new e.g. there is a change in the raw material.

Multivariate statistical methods are also excellent tools to develop processes further. With these methods we can look inside the process to gain the necessary information for optimizing them.

Principal Component Analysis (PCA)

A method for analyzing variability in data. PCA does this by separating the data into Principal Components (PCs). Each PC contributes to explaining the total variability, with the first PC describing the greatest source of variability. The goal is to describe as much of the information in the system as possible in the fewest number of PCs and whatever is left can be attributed to noise i.e. no information. Maps of samples (scores) and variables (loadings) give valuable information of the underlying data structures.

Partial Least Squares Regression (PLSR)

A method for relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

INDUSTRY EXAMPLE: BACKGROUND AND DATA

A paper producer monitors the quality of newsprint by applying ink to one side of the paper. By measuring the reflectance of light on the reverse side of the paper, a reliable, practical measure of how visible the ink is on the opposite side is obtained. This property, *Print Through*, is an important quality parameter. The paper is also analyzed with regard to several other product variables and raw material variables.

The data used in this example is taken from a real-world paper manufacturing process. Samples were collected from the production line over a considerable period of time to ensure the measurements would capture the important variations in production.

Model

A model is a mathematical equation summarizing variations in a data set. Models are built so that the structure of a data table can be understood better than by just looking at all raw values. Statistical models consist of a structure part and an error part. The structure part (information) is intended to be used for interpretation or prediction, and the error part (noise) should be as small as possible for the model to be reliable.

Calibration

Stage of data analysis where a model is established with the available data, so that it describes the data as well as possible. It is imperative that this is based on model validation and not the best numerical fit.

After calibration, the variation in the data can be expressed as the sum of a modelled part (structure) and a residual part (noise). 'Calibration samples' are the samples on which the calibration is based. The variation observed in the variables measured on the calibration samples provides the information that is used to build the model. If the purpose of the calibration is to build a model that will later be applied on new samples for prediction, it is important to collect calibration samples that span the variations expected in the future prediction samples.

Prediction

Predictions are performed by collecting new samples, obtain the values for the variables with the appropriate sensors similarly as in the calibration stage and apply the model to give a prediction (estimate) of the product quality. The multivariate methods also have diagnostics for detecting outliers at the prediction stage.

The data consists of 66 samples with 15 process and product attribute variables and the response variable, *Print Through*. In this case, 16 of the samples were test samples used for prediction using the model based on the calibration data of 50 samples. The process variables are given in Table 1.

Table 1.

Variables and product attributes in a paper manufacturing process which determine quality.

X-var	Name	Description
X1	Weight	Weight / sq. m
X2	Ink	Amount of Ink
X3	Brightness	Brightness of the paper
X4	Scatter	Light scattering coefficient
X5	Opacity	Opacity of the paper
X6	Roughness	Surface roughness of the paper
X7	Permeability	Permeability of the paper
X8	Density	Density of the paper
X9	PPS	Parker Print Surf number
X10	Oil absorb	A measure of the paper's ability to absorb oil
X11	Ground wood	The % of ground wood pulp in the paper
X12	Thermo pulp	The % of thermomechanical pulp
X13	Waste paper	The % of recycled paper
X14	Magenf	The amount of additive
X15	Filler	The % of filler

The purpose was to establish a model that could be used for quality control and production management. The objectives were:

- Predict quality from the process variables and other product variables
- Rationalize the quality control process by reducing the number of variables measured i.e. build a model that includes a subset of variables without losing the underlying variability

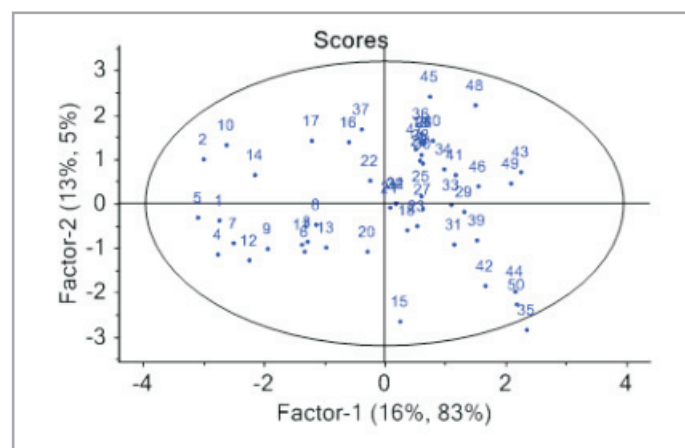
Using The Unscrambler® X multivariate analysis software, a PLS regression model was run with 50 calibration samples and the 15 process and product variables with *Print Through* as the response variable. As mentioned above, an important aspect of multivariate modelling is that the dimensionality of the process is typically lower than the number of process variables measured i.e. there is a redundancy among the observed variables.

This is exemplified in the Scores plot in Figure 2 which summarises the model in the two underlying dimensions ("factors" or "latent variables") for the 15 original process variables. Therefore, rather than plotting the individual variables in one, two or three dimensions, the process can be visualized as a map of the samples in the latent variable space, the Scores plot. The corresponding Loadings plot (not shown) visualizes the relationships between all variables.

Figure 2.

The Scores plot is used to visualize the samples based on all variables.

The Scores plot for Factors 1 & 2 in a paper manufacturing process. In this case, the fact the samples (dots) are evenly and widely scattered indicates there are no clear groupings or outliers, which is a positive result in this instance. However, the timeline in the process can be seen from left to right as shown in Figure 3. The ellipse defines the 95% confidence limit.



The Scores plot

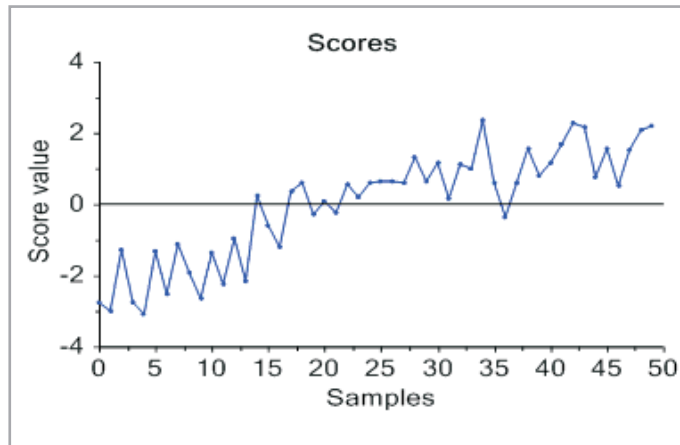
A Scores plot represents each sample in the space defined by a particular Principal Component. They can be plotted as line plots for describing sample trends, or 2D or 3D scatter plots for defining trends and visualizing clusters.

Alternatively one may visualise the change in the process over time as a one-dimensional Scores plot if such a clear trend exists, as shown in Figure 3.

Figure 3.

A line plot of Scores can be used to visualize trends and developments in a process over time.

The line plot of Scores for Factor 1 clearly shows the change over time towards higher Scores, as seen by the upwards trend. In this case this corresponds to higher quality.

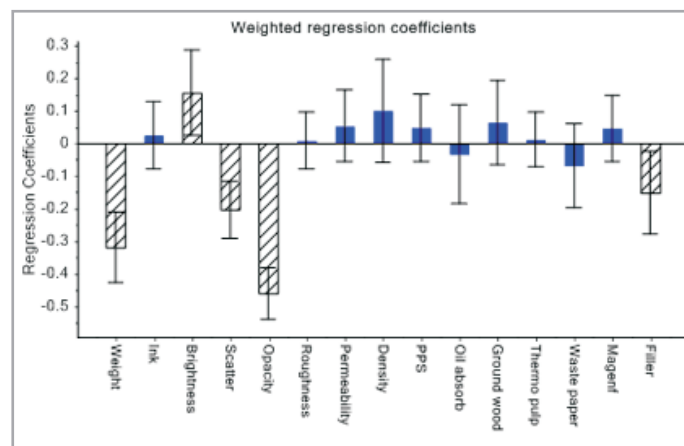


The overall importance of the process variables are most easily depicted in terms of the model coefficients as shown in Figure 4.

Figure 4.

Weighted regression coefficients show which of the variables have a significant impact on the model in terms of the final product quality.

The bars with the diagonal lines are those that have a significant relationship to Print Through. The model shows that Weight, Scatter, Opacity and Filler have an inverse relationship to Print Through (e.g. the lower the weight the higher the Print Through), while Print Through increases with increased Brightness. Where the bars are blue, there is no significant relationship, or there is high uncertainty, as indicated by the 'I' shaped confidence interval lines.



While validating the model robustness, a 95% confidence interval is estimated for each variable, thus indicating which variables are important. The practical benefit of this is that if many variables describe the model in the same way, it is not necessary to measure all of them. Of course, one may decide to continue monitoring all the variables but not to use them for prediction if the parsimonious (most simplistic version possible) model is better for that purpose.

From the results of the first model, a reduced model with only five variables was chosen for on-line prediction using the variables which were shown to be significant:

- Weight
- Brightness
- Scatter
- Opacity
- Filler

Using the Unscrambler® X Process Pulse real-time process monitoring software, process operators or engineers can view interactive plots during the prediction stage which give insight into any changes in the process. Upper and lower control limits for the print through are shown in real-time (Figure 5) and using the Hotelling T^2 statistic (Figure 6 overleaf).

Hotelling T^2 statistic

A linear function of the leverage that can be compared to a critical limit according to an F-test. This statistic is useful for the detection of outliers at the modelling or prediction stage. The 'Hotelling T^2 ellipse' is a 95% confidence ellipse which can be included in scores plots and reveals potential outliers, lying outside the ellipse.

Figure 5.

Predicted values: Multivariate methods allow the quality variable to be represented by simple upper and lower limits using a model based on all five 5 input variables, above.

This chart shows the predicted values for Print Through. In this case, Sample 14 is out of specification (above the red critical limit line), indicating a problem with the process. See Figure 7 for further explanation.

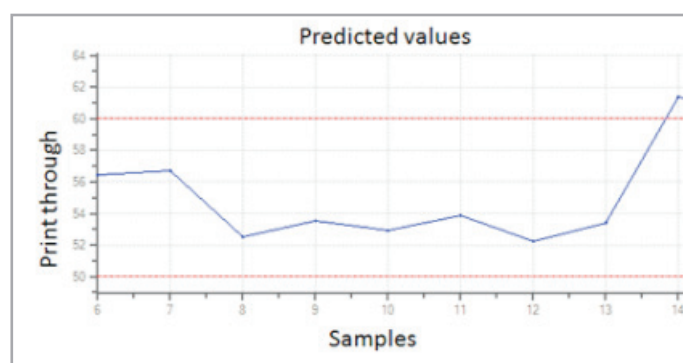
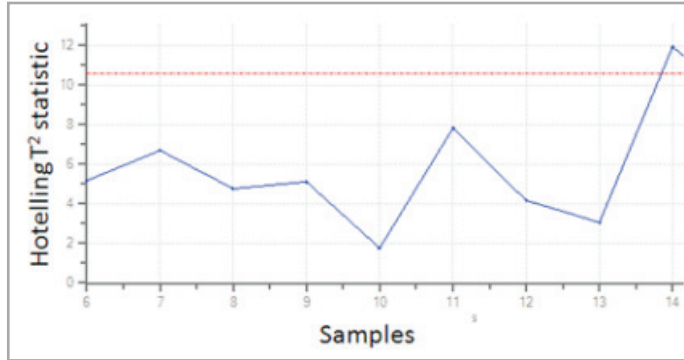


Figure 6.
Hotelling T² statistic: Critical limits

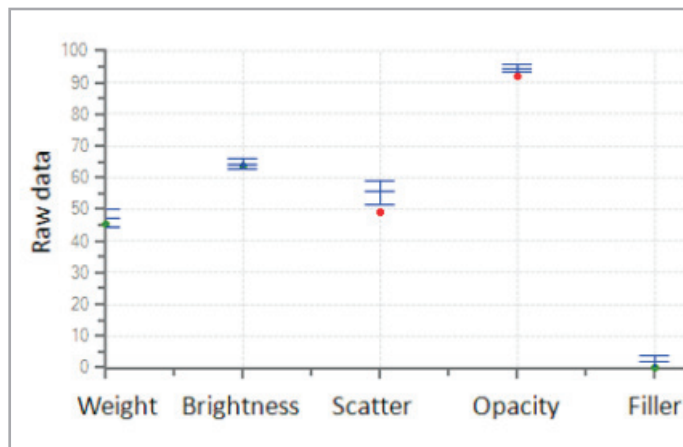
The model shows a univariate representation of the confidence ellipse shown in Figure 2.



Using the Unscrambler® X Process Pulse, when a new sample falls outside the critical limit the process operator or engineer can simply click on the suspect data point in the plot to immediately see which variable is outside the limits as defined by the calibration (Figure 7).

Figure 7.
The diagnostics in Process Pulse allow the process operator to 'drill down' and identify the specific variables which are out of limit in a sample.

After drilling down into Sample 14, the process operator can see that Opacity and Scatter are outside the min and max limits from the calibration, illustrated by the red dot below the blue min and max lines.



PROCESS OPTIMIZATION

Once a model has been established the next step may be optimization of the process with constraints on the process variables as well as the product quality.

In this case, some constraints were given for the five important variables in the final model, while at the same time the target range for the *Print Through* was set to be between 35 and 40.

Figure 8 shows the result of the optimisation using the Unscrambler® Optimizer software. It is also possible to add interaction and square terms but this was not pursued in detail as there were no indications that such additional model terms improved the model.

Figure 8.

Optimization software enables manufacturers to choose the best process setting within specified constraints.

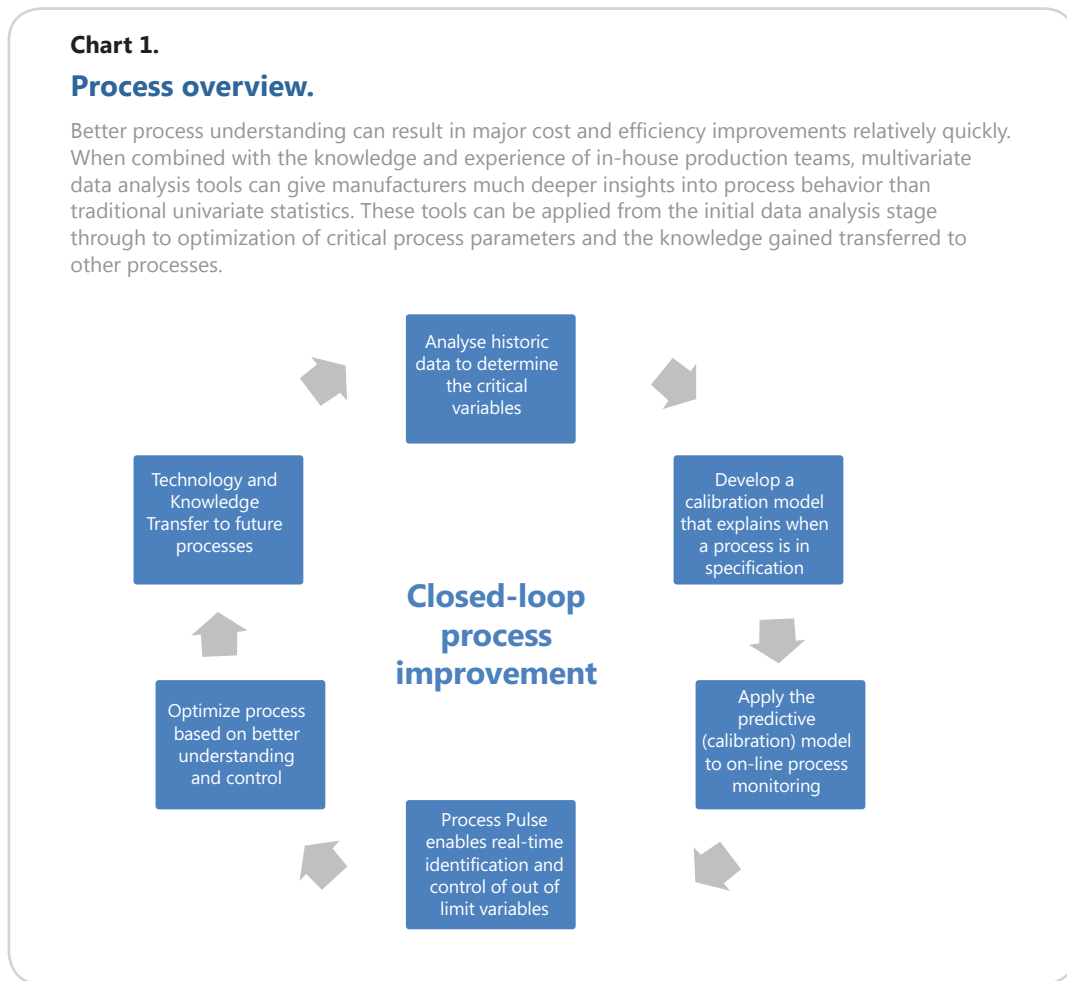
By setting the upper and lower limits for the five key process variables, as defined by the earlier analysis, they are then able to identify the optimal process settings given the specified constraints using Unscrambler® Optimizer software.

Process variables						
Factor	Lower bound	Min	Max	Upper bound	Opt Value	Response Chart
Weight	42	39.9	49.7		48.224247	
Brightness	63	62.300003	66.1	65	63.555107	
Scatter	50	48.649998	58.95		56.604237	
Opacity		91.9	95.700005		95.27748	
Filler	1	0.0	3.6000001	2	2.0	

Quality						
Main goal	Factor	Calibration Range		<= Target	<=	Predicted Value
y	Print through	31.5	- 69.0	35	40	37.5

In this case, the paper manufacturer was able to:

- Identify the critical process variables affecting *Print Through*
- Implement real-time process monitoring enabling them to fix possible failures at an early stage
- Optimize the process using their newfound understanding of its behaviour
- Improve end product quality and reduce scrap, re-work and energy costs



SUMMARY

Multivariate methods are a powerful and efficient tool in monitoring process variables as well as for predicting final product quality at an early stage. For complex processes involving several variables interacting, multivariate statistical process monitoring (MSPM) methods are considerably more effective than univariate control charts. They enable the identification of the process “sweet-spot”, while disturbances in the process can easily be detected and the variables causing the upset can be interactively spotted in the on-line monitoring phase.

Importantly, process operators do not need to understand the methodology behind the system, as the plot of the original process variables is shown on screen. The concept of MSPM can also be extended to hierarchical models for classification and prediction of raw material quality in a complete production process quality system.

CAMO SOFTWARE

PRODUCTS & SERVICES

Our powerful yet easy to use and affordable solutions are applied around the world in a wide range of industries

The Unscrambler® X

Leading multivariate analysis software used by thousands of data analysts around the world every day. Includes powerful regression, classification and exploratory data analysis tools.

 [TRIAL VERSION](#) | [READ MORE](#)

Unscrambler® X Process Pulse

Real-time process monitoring software that lets you predict, identify and correct deviations in a process before they become problems. Affordable, easy to set up and use.

 [TRIAL VERSION](#) | [READ MORE](#)

Unscrambler® X Prediction Engine & Classification Engine

Software integrated directly into analytical or scientific instruments for real-time predictions and classifications directly from the instruments using multivariate models.

 [TRIAL VERSION](#) | [READ MORE](#)

Consultancy and Data Analysis Services

Do you have a lot of data and information but don't have resources in house or time to analyze it? Our consultants offer world-leading data analysis combined with hands-on industry expertise.

 [READ MORE](#) | [CONTACT US](#)

Training

Our experienced, professional trainers can help your team use multivariate analysis to get more value from your data. Classroom, online or tailored in-house training courses from beginner to expert levels available.

 [READ MORE](#) | [CONTACT US](#)

Our partners

CAMO Software works with a wide range of instrument and system vendors. For more information please contact your regional CAMO Software office or visit

 www.camo.com/partners

Find out more >

For more information please contact your regional CAMO office or email sales@camo.com

Did you find this useful? Send it to a friend or share it in your network.

 www.camo.com



NORWAY

Nedre Vollgate 8,
N-0158
Oslo
Tel: (+47) 223 963 00
Fax: (+47) 223 963 22

USA

One Woodbridge Center
Suite 319, Woodbridge
NJ 07095
Tel: (+1) 732 726 9200
Fax: (+1) 973 556 1229

INDIA

14 & 15, Krishna Reddy
Colony, Domlur Layout
Bangalore - 560 071
Tel: (+91) 80 4125 4242
Fax: (+91) 80 4125 4181

JAPAN

Shibuya 3-chome Square Bldg 2F
3-5-16 Shibuya Shibuya-ku
Tokyo, 150-0002
Tel: (+81) 3 6868 7669
Fax: (+81) 3 6730 9539

AUSTRALIA

PO Box 97
St Peters
NSW, 2044
Tel: (+61) 4 0888 2007

