



Malcolm Moore

# Increasing process understanding through data mining and statistical modelling

## ABSTRACT

*A lean and simple approach to increasing process understanding that aligns scientific and engineering thinking with statistical principles is presented. This allows a larger community of scientific and engineering users to apply statistical methods to increase process understanding based on data and use this knowledge to increase product quality.*

## INTRODUCTION

The essence of the PAT initiative driven by the FDA (<http://www.fda.gov/cder/guidance/6419fnl.pdf>) is to enable Pharmaceutical manufacturers to improve product quality through increased process understanding based on data. The initiative encourages a data driven approach in process development and manufacturing, specifically one of measuring the relevant process features, integration of the resulting data, and statistical modelling to drive process understanding based on data. It is expected that companies who are able to demonstrate process understanding based on supporting data will be treated differently and may benefit from relaxed regulation and a transition towards continual verification that allows improvements throughout the life-time of manufacturing to be made provided the proposed changes are supported by data and statistical analysis of that data. This change allows pharmaceutical manufacturers to catch up with other manufacturing sectors with regard to strategies for dramatically improving quality and reducing costs that have occurred over the last twenty years or so through Total Quality Management (TQM), Six Sigma and related initiatives.

The FDA Guidance lists four tools of PAT:

- Multivariate tools for design, data acquisition and analysis
- Process analysers (at-line, on-line, in-line measurement tools)
- Process control tools
- Continuous improvement and knowledge management tools

Much of the focus to date has been in the area of installing process analysers, calibrating these new measurement systems against existing off-line QA systems using multivariate analysis methods and installing a new control system based around the

newly calibrated in-line or at-line measurement system with the aim of increasing in-process or end-process quality. For brevity this paper pools the first three tools and calls them PAT calibration methods. It then contrasts PAT calibration methods and continuous improvement tools with regard to the likelihood of achieving dramatic improvements in product quality and process understanding.

## HOW DOES PAT CALIBRATION LEAD TO BETTER QUALITY?

Process analysers include mid-infra red, near-infra red and laser probes that deliver spectral data that need to be calibrated against a direct measurement of the entity of interest using predictive modelling techniques such as Partial Least Squares (PLS). Once calibrated a process analyser might be used for end-point detection (i.e. stopping a process once the desired state has been achieved), e.g. particle size in milling, blend uniformity in blending, moisture content in drying, and so on. Additionally process analysers once calibrated can obtain in-line measurements of key process indicators such as tablet coating thickness, API content in blended material, API form and stability, etc. Figure 1 depicts a typical analysis procedure for calibrating a process analyser against an existing off-line measurement system for measuring the water content of tablets. NIR transmittance measurements for a sample of tablets result in a spectra for each tablet, the spectra are pre-processed to remove noise, the off-line water content is then modelled as a function of the de-noised spectra using predictive modelling techniques such as Partial Least Square (PLS) and providing the resulting model adequately predicts the off-line measurement the new process analyser can be used for in-line or at-line control purposes. This calibration process uses a sophisticated multivariate toolset that requires a highly skilled analyst such as a Chemometrician or Statistician. The process illustrated in Figure 1 is representative of many PAT projects to date, the goal is to increase product quality through in-line control and end-point detection based around in-line or at-line process analysers, however this approach does not necessarily drive increased process understanding (at least in the context of the definition given in section 3).

## SIX SIGMA: DRIVING INCREASED PROCESS UNDERSTANDING THROUGH CONTINUOUS IMPROVEMENT

Six Sigma is a process improvement or problem solving methodology that helps us focus on solving improvement opportunities (problems) with the biggest business impact and to base decisions on data. It requires us to think in terms of the statistical principles of variation, specifically variation in product quality (Y's) is due to variation in one or more process inputs (X's). The key is to find the specific process inputs (hot X's) that drive variation in product quality (Y's) and to understand the transfer function (at least at an empirical level) as to how the hot X's transmit variation into product quality (Y's). DMAIC (Define, Measure, Analyse, Improve, and Control) is a common problem solving process deployed within Six Sigma. As illustrated in figure 2, it involves collecting data on a large number of process inputs and modelling their impact on

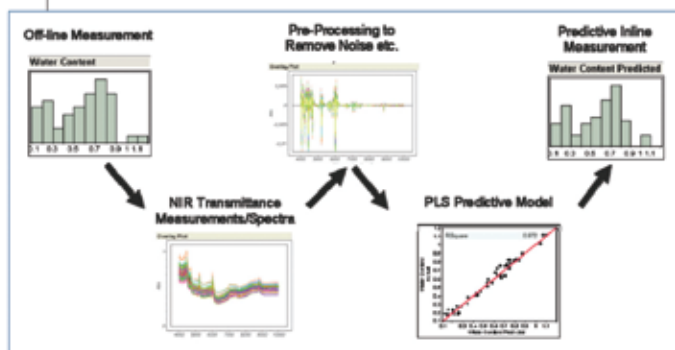


Figure 1

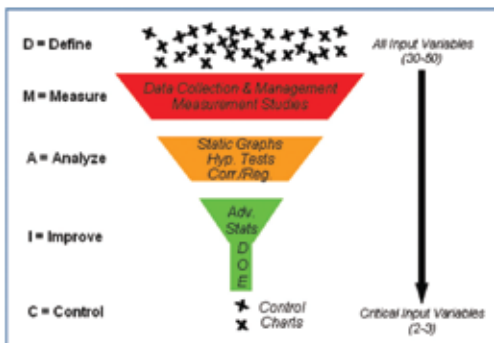


Figure 2. The Essence of DMAIC

product quality to determine the hot X's and the way in which they drive variation in product quality.

Unlike the PAT Calibration approach, Six Sigma aims to drive increased process understanding/quality through fewer controls not more. Six Sigma thinking delivers process understanding with the goal of informing us how to operate the process for robustness, i.e. minimise transmission of variability from process to product (Taguchi).

It takes the stance that we don't want to be routinely measuring lots of things in manufacturing. However, measurement of several X's, is necessary to deliver the data required to build models of the dependence of product quality on process inputs. Effective implementation of Six Sigma is not without its challenges. One challenge is that Six Sigma requires us to collect data on the potential process inputs or X's, which may not be routinely available for the manufacture of mature pharmaceutical products. Another challenge is the effort required to learn and master the statistical concepts listed in table 1. First projects often have an elapsed time of six months to one year, and the two issues that contribute to these long project cycles are the time to:

1. Collect data on the X's and verify the adequacy of measurement systems for the Y's.
2. Time to become proficient in applying the statistical toolset in table 1 to solve problems. A secondary issue is that if statistical analysis is performed on an infrequent basis, users tend to forget how to apply statistical methods to solve problems, meaning that project cycle times rarely get shorter than a few months.

Six Sigma is often overly engineered with respect to statistical complexity. Section 4 will present some ways of reducing this complexity and enabling engineering and quality departments to speed problem solving and incrementally drive increased process understanding through the application of the principles of statistical thinking. A lean, simpler, and faster approach to problem solving is presented.

### INTEGRATING PAT AND SIX SIGMA

An appropriate balance of PAT calibration and Six Sigma approaches are needed to deliver dramatic improvement in product quality and reduction in waste. PAT calibration can deliver timely data on some process inputs, intermediate product and final product quality, which feeds some of the data sources required by Six Sigma. As illustrated by Moore (1) the additional control capabilities offered through PAT calibration may offer an incremental improvement in product quality but PAT calibration and control alone will not deliver defect or near defect free processes. As illustrated in figure 3 a blend of three data analysis approaches are needed:

1. Calibration to deliver timely data to deploy in process understanding and process control.
2. Process understanding based on a lean, simpler, and faster approach to Six Sigma.
3. Process controls where needed, i.e. when the process cannot be made sufficiently robust.

The enabling technology required to support this framework falls into three categories as illustrated in Figure 4:

1. Data integration technology is required to deliver

integrated, clean data that is analysis ready. Note: the time to deliver analysis ready data often represents more than half the time spent in performing data analysis. Data integration technology would integrate and cleanse data from disparate manufacturing and quality data sources including LIMS, SCADA, MES, Process analysers, as well as desktop applications such as Excel.

2. Model Building technology is required to help process development, engineering, and quality groups to effectively and efficiently apply the principles of statistical thinking and drive increased process understanding through the application of statistical principles.

3. Model deployment refers to the process controls needed to ensure product quality when the process cannot be set to deliver adequate robustness/quality.

Category	Technique
Data Management	Query Quality Checks Data Cleansing Merge/Split Transformation
Basic Statistics	Descriptive Statistics Distribution Fitting
Static Graphs	Histogram Stem & Leaf Run Chart Pareto Scatterplot 3d Scatter Plot Box plot Individual Value Multi-vari Matrix plot Bar Chart Pie Chart Time-series plot
Measurement Studies	Linearity and Bias Gauge R&R Process Capability
Hypothesis Testing	Compare Means/Medians Compare Counts
Correlation & Regression	Correlation Matrix Simple Regression Simple Logistic Regression
Advanced Statistics	Multi-Way ANOVA Multiple Regression Generalized Linear Modelling
DOE	Screening RSM Taguchi
Control Charts	Run Chart Xbar/R/S IMR MA, EWMA, Cusum C, U, P, NP

Table 1. Statistical Tools Used in Six Sigma

### Lean Model Building Process

Moore and Cox (2) proposed a lean data analysis process for the model building step supported by visual analysis software such as JMP from SAS. The first two steps of this process



Figure 3. Data analysis through manufacturing life cycle

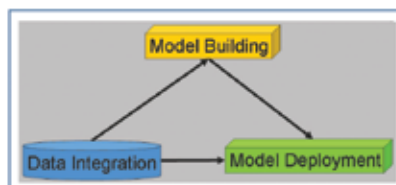


Figure 4. Enabling technology

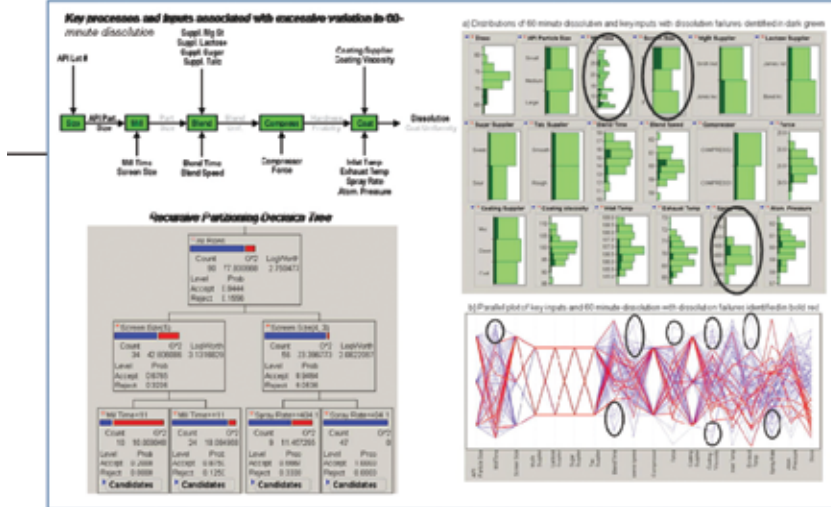


Figure 5. Visual Modelling.

the process for robustness, or by determining additional controls required to obtain consistency of product or a combination of both. Moore (1) and Moore and Cox (2) have presented case studies based around this lean model building process, one of these cases is summarised.

### CASE STUDY

A fictional case study was presented in (1) based around a fairly typical situation in pharmaceutical manufacturing. It related to a process for tablet production at a single concentration where the key performance metric was 60-minute mean dissolution. Historically, 16% of production batches failed to meet the 60-minute mean dissolution requirement of not less than 70%. Figure 5 summarises the key graphical tools used in framing the problem and identifying the hot X's:

1. A process map was used to identify the data that can be easily collected for past production lots.
2. Side by side histograms with the lots with a 60-minute mean dissolution below 70% identified in darker shading help identify three hot X's – mill time, screen size and spray rate.
3. A data mining decision tree shows the acceptance rate and reject rate according to four sub-groups defined by splits based on the values of the top three hot X's. One quick solution to reduce the reject rate is offered in the right-hand branch of the decision tree which involves tighter controls on screen size and spray rate.
4. A parallel coordinates plot identifies passing lots in blue and failing lots in red and illustrates the opportunity to tighten the control range of several X's in order to reduce defect rate, these X's include mill time, blend time, blend speed, compression force, coating viscosity, exhaust temperature and spray rate.

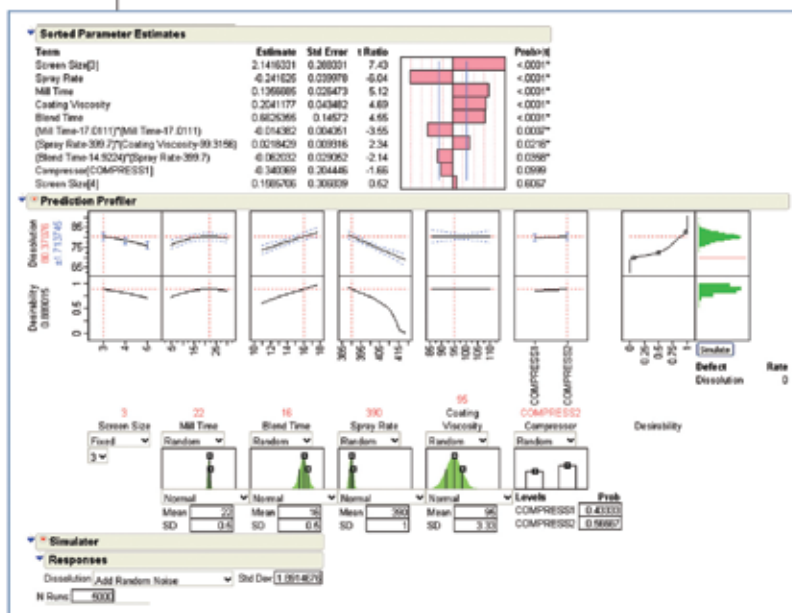


Figure 6. Statistical Modelling

Figure 6 illustrates what is possible when progressing to statistical modelling methods with multiple regression analysis confirming the importance of the hot X's identified by parallel coordinates and running Monte Carlo simulations from the multiple regression model to identify the control tolerances of the hot X's we need to maintain in order to get a near defect free process.

concern the identification of the X's and Y's to measure, collecting the data, ensuring it is free of obvious errors and any measurement errors will not mask the patterns of process variation. Once the data is clean and free of large measurement errors, the model building process consists of two steps:

1. discovering patterns between X's and Y's using dynamic visualisation techniques, and
2. empirical or statistical model building to help understand how to exploit these relationships to ensure increased product quality.

This lean analysis process aligns scientific and engineering thinking with statistical principles, enabling a larger community of users to increase process understanding supported with data and drive quality levels higher by determining how to operate

### SUMMARY

PAT calibration and control methods alone are unlikely to deliver near defect free processes. Continuous improvement methodologies such as Six Sigma are also required to deliver a body of knowledge to ensure processes are made robust wherever possible with process controls deployed to ensure high levels of quality where robustness cannot be designed into the process. A lean and simple approach to Six Sigma concepts that aligns scientific and engineering thinking with statistical principles, and enables a larger community of scientific and engineering users to increase process understanding based on data was presented. This approach speeds project cycles and provides a sound basis for incrementally improving product quality as per the PAT principle of continuous verification.

### REFERENCES

1. Moore, M. Lean Data Analysis: Simplifying the Analysis and Presentation of Data for Manufacturing Process Improvement, Pharmaceutical Engineering (March 2007)
2. Moore, M. and Cox, I. Making Data Analysis Lean, Pharmaceutical Technology Europe (April 2008)

### MALCOLM MOORE

SAS UK  
Wittington House  
Henley Road  
Medmenham  
Marlow, Buckinghamshire  
SL7 2EB United Kingdom

*Reprinted with permission from Chimica Oggi, February 2008*



JMP WORLD HEADQUARTERS SAS INSTITUTE INC. +1 919 677 8000 U.S. & CANADA SALES 800 727 0025 [www.jmp.com](http://www.jmp.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2008, SAS Institute Inc. All rights reserved. 103464\_496325\_0508