The Use of Statistics in Understanding Pharmaceutical Manufacturing Processes

Thesis submitted by

Keeley Burke

For the degree of Engineering Doctorate

School of Chemical Engineering and Advanced Materials

December 2015

Abstract

Industrial manufacturing processes for pharmaceutical products require a high level of understanding and control to demonstrate that the final product will be of the required quality to be taken by the patient. A large amount of data is typically collected throughout manufacture from sensors located around reaction vessels. This data has the potential to provide a significant amount of information about the variation inherent within the process and how it impacts on product quality. However to make use of the data, appropriate statistical methods are required to extract the information that is contained. Industrial process data presents a number of challenges, including large quantities, variable sampling rates, process noise and non-linear relationships.

The aim of this thesis is to investigate, develop and apply statistical methodologies to data collected from the manufacture of active pharmaceutical ingredients (API), to increase the level of process and product understanding and to identify potential areas for improvement.

Individual case studies are presented of investigations into API manufacture. The first considers prediction methods to estimate the drying times of a batch process using data collected early in the process. Good predictions were achieved by selecting a small number of variables as inputs, rather than data collected throughout the process. A further study considers the particle size distribution (PSD) of a product. Multivariate analysis techniques proved efficient at summarising the PSD data, to provide an understanding of the sources of variation and highlight the difference between two processing plants.

Process capability indices (PCIs) are an informative tool to estimate the risk of a process failing a specification limit. PCIs are assessed and developed to be applied to data that does not follow a standard normal distribution. Calculating the capability from the percentiles of the data or the proportion of data outside of the specification limits has the potential to generate information about the capability of the process. Finally, the application of Bayesian statistical methods in pharmaceutical process development are investigated, including experimental design, process validation and process capability. A novel Bayesian method is developed to sequentially calculate the process capability when data is collected in blocks over time, thereby reducing the level of noise caused by small sample sizes.

Acknowledgements

Firstly I would like to thank my academic supervisors, Professor Elaine Martin and Professor Gary Montague for all their advice and guidance throughout my time at Newcastle. Thanks to Elaine for reading through all of my work and providing lots of useful suggestions and improvements. It was great working with my fellow EngD students and getting to know Joana, Grace, Jodie, Charlotte and Jules over the last few years, good luck for your future careers, I'm sure they will all be successful. I am also grateful to EPSRC for providing funding for the course.

I would also like to thank my industrial supervisor Jim Baker for his support while I was based at AstraZeneca, and for reading through all of my reports and chapters. It was great to feel part of the team at Avlon and I really enjoyed working with everyone there, I've certainty gained a lot of valuable experience for my future career. Thanks to Angela and Richard for keeping me entertained in the office! I also enjoyed working with the Global Statistics Forum, who were a great source of advice and information.

I would like to thank my family for all their love and support, especially Mum and Steve for driving me around the country throughout my university years. Finally to my husband Richard for all your love, support and advice that has helped me to be complete this thesis.

Contents

List of Figures

List of Tables

Notation

Φ	Normal distribution function
θ	parameter of a statistical distribution
Σ	Variance matrix
μ, $\bar{x}$	Population / sample mean
σ, s	Population / sample standard deviation
A	Number of retained principal components
$CaCl_2$	Calcium Chloride
C_{Np}, C_{Npk}, $C^{\#}_{pk}$	Distribution free process capability indices
C_p, C_{pk}	Process capability indices
D	Distance between two datasets
D10, D50, D90	10th, 50th, 90th percentiles of the data
E	Error matrix
E[ ]	Expectation function
$f(\theta)$	Probability distribution function of theta
F(USL), F(LSL)	Proportion of data below USL, LSL
$f_{hist,i}$	Feature from historical data
$f_i( )$	Individual neural network
$f_{new,i}$	Feature from new data
K	Number of variables
L( )	Loss function
M	Median
N	Number of samples
P	Loadings matrix
P(A|B)	Probability of event A given event B has occurred
P(USL), P(LSL)	Percentile of USL, LSL
PIT05, PIT08	Pressure sensors
P_p, P_{pk}	Process capability Indices
Px	xth percentile
Q^2	Proportion of variation that can be predicted in new data
R^2, R^2Y	Proportion of variation explained in response data

R^2X	Proportion of variation explained in input data
Spk(%), Spk(percentile)	Distribution free process capability indices
T	Scores matrix
t_in	Inlet temperature
t_out	Outlet temperature
$\mathbf{t_1}$, $\mathbf{u_1}$	Latent variable vectors
U( )	Utility function
$v_{hist,i}$	Variable from historical data
$v_{new\ i}$	Variable from new data
$\mathbf{w_1}$, $\mathbf{u_1}$	Weights vectors
X	Input data matrix
Y	Response data matrix
$\hat{\mathbf{Y}}$	Predicted response

Acronyms

8D	Eight Disciplines
API	Active Pharmaceutical Ingredient
CBR	Case Based Reasoning
CPP	Critical Process Parameter
CQA	Critical Quality Attribute
CUSUM	Cumulative Sum
DModX	Distrance to Model
DoE	Design of Experiments
EWMA	Exponentially Weighted Moving Average
FDA	Food and Drug Administration
FMEA	Failure Mode Effects Analysis
GMP	Good Manufacturing Practice
HEPA	High Efficiency Particulate Absorption
IBC	Intermediate bulk container
LD	Laser Diffraction

LOD	Loss on Drying
LSL	Lower Specification Limit
LV	Latent Variable
MLR	Multiple Linear Regression
MPCA / MPLS	Multiway PCA / PLS
MSE	Mean Squared Error
N_2	Nitrogen Gas
NIR	Near Infra-Red
NN	Neural Network
NOR	Normal Operating Range
PAT	Process Analytical Technology
PC	Principal Component
PCA	Principal Components Analysis
PCI	Process Capability Index
PCR	Principal Component Regression
PLS	Partial Least Squares
PSD	Particle size distribution
QbD	Quality by Design
QC	Quality Control
RJF	Reverse Jet Filter
SD	Standard Deviation
SE	Standard Error
SEM	Scanning Electron Microscopy
SPC	Statistical Process Control
SPE	Squared Prediction Error
USL	Upper Specification Limit
UV	Ultra-violet
VIF	Variance Inflation Factor

1 Introduction

The global population has a need for affordable and quality medicines to support improvements to public health and to help fight disease. As a result, pharmaceutical companies need to develop efficient and robust large scale manufacturing processes that can be relied upon to produce the medicines that people need, at the lowest possible cost. Throughout the lifecycle of a pharmaceutical product, improvements will be continually made to the manufacturing process, as more knowledge is gained about the processes being run. To meet the requirements of the patient, a manufacturing process is required that consistently and cost effectively produces a high quality product.

The manufacture of a pharmaceutical product is a complex process, involving the manufacture of the active pharmaceutical ingredient (API) and then the formulation of the API into a product that can be taken by a patient. The focus of this thesis is the manufacturing processes of API products, which comprise of a number of unit operations, including dissolution, distillation and crystallisation. The manufacture of API is generally a batch process. A number of raw materials will be added at different stages of the process and the resulting product is typically a solid powder that is the active ingredient within a medicine.

During the manufacture of an API, key process variables are controlled to specific levels, including quantities of raw materials, temperature and pressure settings. However not all the inputs are controllable, for example the characteristics of raw materials or the ambient conditions of the plant can vary between batches. As a result, although the process is run with the same set up, variation in the inputs will result in variability in the final product. Process outputs describe the quality of the product and include the purity, crystallisation structure and particle size. Business objectives such as reaction yield, cycle time and cost are also important outputs.

Due to the complexity of API processes, it is difficult to represent an entire process through physical models and hence empirical methods have become important for process and product characterisation. A number of measurements can be collected throughout a process providing a large amount of data, for example with temperature and pressure probes inside of reaction vessels. This data has the potential to provide important information about the source of variability in the process.

By understanding the variation in batches that have previously been manufactured, inferences can be made about the behaviour of futures batches. More specifically the data from previous batches can be used to understand how variation in the inputs can

affect the outputs, for example to show the impact of the reaction temperature on the process yield. Typically there are a number of variables that will impact the process, and hence to understand and quantify these relationships, multivariate statistical techniques can be applied. These techniques form the basis of the research undertaken in thesis.

The knowledge that is gained from the data analysis can be used to drive process and product improvements. For example, by identifying which inputs have the strongest relationship with the outputs, the key input variables can be controlled to minimise their impact on product quality. In addition, statistical models of the process will allow for optimal settings for the controllable inputs to be determined that will result in the desired outputs. From these analyses, the most important inputs can then be monitored so that the onset of a process change can be detected early and problems mitigated before they have an impact on product quality. The outcome of the analysis should be a process that is robust to variation in uncontrollable inputs and consistently and efficiently produces a product to the desired level of quality. In this thesis, a number of statistical methods for handling industrial process data are researched and then applied to specific investigations on an industrial pharmaceutical manufacturing process.

1.1 Industrial Partner

This research project was undertaken in collaboration with AstraZeneca, the industrial partner. AstraZeneca are a global pharmaceutical company with research interests in both small molecule and biopharmaceutical products. Research for new medicines is targeted at disease areas including cancer, cardiovascular, respiratory and inflammation. AstraZeneca are involved in the whole lifecycle of a product, starting with the identification of a potential medicine for an unmet medical need, through to safety and efficacy studies, clinical trials, regulatory submissions and manufacture. While the new medicine is being developed, production processes are scaled up from the laboratory through to large scale manufacturing. Manufacturing stages include the production of the active pharmaceutical ingredient, its formulation into the product taken by the patient, and the packaging of the product to be distributed to vendors.

This thesis focuses on the manufacture of AstraZeneca's API products. Research topics focus on the investigation of statistical methods that can be applied to manufacturing processes that are run undertaken by AstraZeneca. The specific processes under investigation have been established for a number of years and hence the priorities for process improvement are to increase robustness, reduce manufacturing costs and maintain a reliable supply chain to the customer. These

improvements will be achieved by working to reduce the potential for batch losses, reducing cycle times, cutting waste and improving process efficiency. Other priority areas for consideration are process safety and compliance with the principles of good manufacturing practice.

A further consideration for AstraZeneca when planning process improvements is that of regulation. Since the final products are pharmaceuticals, the processes are registered with a number of regulatory authorities, including the United States Food and Drug Administration (FDA). The registration contains details of the manufacturing process, including ranges for various process parameters, such as reaction temperatures and timings of reactions. The processes must be run within the registered ranges for the final product to the approved for release to the customer. Therefore any changes to a process must be within the registered range of the process or else the process is required to be re-registered with the regulators. The purpose of regulation is thereby to ensure that the manufacturing processes have been well controlled so that product quality can be assured prior to the quality control (QC) testing being undertaken**.** However batch failures may still occur if there is poor understanding of the relationship between the critical process parameters and critical quality attributes. The ability of a process to consistently be run within the specification limits can be quantified through the use of process capability indices (Chapter 6).

A large amount of data is collected on the production plants, from measurement probes such as temperature, pressure and flow rate meters located in and around the reaction vessels. AstraZeneca wish to make use of the information contained in the data to increase process understanding and consequently to make improvements that align with the company priorities. The data collected is also used to support problem solving investigations, for example to indentify when and why a change has occurred within a process.

1.2 Data challenges

The data that is collected by AstraZeneca presents a number of challenges that need to be addressed to allow the data to be analysed effectively.

Quantity of data

Measurements from sensors are collected every ten seconds, with some process stages taking hours or days to complete, as a result many thousands of data points are collected for a single batch. Therefore the data needs to be summarised to capture the underlying trends without removing useful information contained within the data.

Additionally, data from industrial processes may be limited in terms of the number of samples, particularly for a batch process when batches are not produced at a high rate. Therefore analysis techniques are required that are suitable for analysing small sample sizes. One approach is to utilise Bayesian statistics, since existing information can be combined with new data to produce inferences about a process (Chapter 7).

Quality of Data

The data may contain a lot of noise, which will hide the trends between variables. For example, poor control of a process variable may result in a large amount of variation in the data that does not necessarily impact on the process outputs, although critical process parameters are controlled to be within the acceptable ranges. Additionally some sensors are known to drift over time, resulting in a change to the measured data that does not correspond to a change in the process; however unacceptable drift is avoided by calibration. Appropriate statistical methods are required that can separate the signal from the noise, to extract the information that is contained within the data.

Multivariate Data

When a large number of measurements are collected from the same process, correlations will exist within the data. Consequently there are fewer underlying trends than the number of variables and traditional linear modelling techniques may not be suitable. Multivariate analysis techniques are thus required to handle the correlations in the data, both to highlight the trends in the data and for developing predictions models.

Process Complexity

A number of complex chemical reactions are involved in the manufacture of API products, which can result in non-linear trends between variables that must be captured by process models. In addition interactions between variables will add to the complexity of the data. Processes may be run across multiple stages, which will each impact the on characteristics of the final product. As a result, the relationship between input and output variables may not be adequately represented with linear models and hence suitable modelling techniques, for example neural networks (Section 3.5), are required that are able to capture the complex trends within the data.

Batch Processes

Since API processes are run as batch rather than continuous processes, batch to batch variation must be considered along with variation within a batch. With batches

processes the duration of certain process stages will vary between batches, resulting in data that is of different lengths for each batch. This type of data needs to be handled appropriately to allow batches of different durations to be compared. In addition the processing conditions can vary between batches, for example as a result of ambient conditions or a process shutdown. Consequently the data will exhibit a large amount of batch to batch variation and data analysis is required to determine how this variation can have an effect on the final product.

1.3 Research Themes

The following research themes were identified that are of relevance to the challenges identified at AstraZeneca and the data available to analyse the processes.

- Multivariate data analysis is a widely applied statistical technique in the process industries (Chapter 3). Hence multivariate methods are investigated to determine the specific applications, benefits and challenges. These techniques then are applied to processes at AstraZeneca in Chapter 4 and Chapter 5.
- Complex processes often exhibit non-linear behaviour. Therefore non-linear modelling methods are investigated (Chapter 3) and applied to the case studies in Chapter 4 and Chapter 5, to compare to the results from linear techniques.
- AstraZeneca wish to monitor the capability of its processes with respect to in-process specification limits. Various process capability indices, including the standard P_{pk} metric and distribution free metrics, are investigated that may be suitable for application to the data that is available (Chapter 6)
- Bayesian statistics is growing in importance in the pharmaceutical industry (Chapter 7). Bayesian methods are investigated to determine how and where they have been applied to pharmaceutical processes, through a detailed literature review and then by identifying novel opportunities at AstraZeneca

1.4 Aims and Objectives

The aim of this thesis is to investigate and develop statistical methodologies to apply to data from API manufacturing processes to increase the level of process understanding and to identify potential areas for process improvement. To meet this aim, the following objectives were identified:

1. Compile a literature review of the role of statistics in the manufacturing industries.
2. Identify key statistical methodologies that are applicable to data from industrial batch processes.
3. Compare advanced modelling techniques to predict batch drying times.

4. Apply multivariate and non-linear methods to analyse the particle size distribution of a product.
5. Investigate and develop potential process capability indices to be applied to data from industrial processes.
6. Investigate the uses of Bayesian statistics in process development.
7. Develop a novel Bayesian methodology for calculating the process capability from data that is collected sequentially.

For each investigation in to support objectives 3 to 5 and 7, potential statistical methods are identified, where necessary developed and then applied to the data available to gain information about the process being studied and to propose potential process improvements. Consideration was given to the specific complexities of data generated from API manufacturing processes, such as large data sets and batch level data. The importance of regulatory control is also considered. For each methodology, the practicality of implementation is assessed, to determine how much information is gained and how a method can be effectively applied to industrial process data.

1.5 Thesis Outline

Chapters Two and Three are a general introduction to the research topic of the thesis, focusing on statistical techniques that are relevant to the analysis of industrial process data. Chapters Four to Seven are individual case studies relating to the API manufacturing processes at AstraZeneca. The overall conclusions and contributions of the thesis are given in Chapter Eight, along with recommendations for future research.

An initial literature survey is undertaken in Chapter Two that describes how statistics has become an important tool for contributing to improvements to manufacturing processes. Specific topics discussed include statistical process control, Quality by Design and continuous improvement methodologies such as Lean and Six Sigma.

In Chapter Three, modelling methodologies are investigated that are applicable to data generated from API manufacturing processes, to gain process understanding and develop prediction models. In particular multivariate statistical analysis methods, techniques for handling batch data and artificial neural networks are considered. The background theory is presented along with examples of applications and a consideration of how the methods can be implemented.

In Chapter Four, statistical methods are compared in terms of their applicability to make predictions of end process characteristics from data collected throughout the duration of a batch process. An investigation was undertaken into the length of time required for a drying process to be completed, using online temperature and flow rate

measurements. The aim was to estimate the duration early in the process so that plant resources could be planned for the batch completion. Linear models and artificial neural networks were investigated and applied to a small number of uncorrelated input variables. Additionally, multi-way partial least squares and case based reasoning were applied to the data that is collected throughout the duration of the batch.

A study is presented in Chapter Five to understand sources of variation in the particle size distribution of a product. The resulting data is multivariate in nature, comprising of a frequency distribution of the proportion of particles of various sizes. Multivariate techniques are applicable to data from particle size distribution measurements, since the data comprises a series of particles sizes and the corresponding frequencies. Summarising the data in terms of a limited number of latent variables enables an efficient comparison to be undertaken between samples. Multivariate prediction methods and artificial neural networks are also applied to assess which process variables have the strongest influence on the particle size of the final product.

Process capability indices are used to calculate the risk of data from a process falling outside of a specification limit, resulting in a batch failure. AstraZeneca uses process capability indices to identify the risk of a specification limit being breached for in-process measurements. Standard process capability indices may not be appropriate for data that does not follow a normal distribution. In Chapter Six, alternative distribution free indices are investigated and a novel method proposed, which is applied to both simulated and process data to compare the accuracy and precision of the various capability indices.

Bayesian statistical methods are an important area of research in statistics, allowing for existing information to be combined with information from data, to produce a posterior distribution that represents parameter being estimated. In Chapter Seven, Bayesian methods for experimental design, process validation and process capability are investigated to assess how each technique could enhance process development. A novel Bayesian approach to process capability is developed for application to data that is collected in sequential blocks, for which the process capability is reported for each individual block. When the sample size is small, Bayesian methods allow information from older data to be combined with the most recent data, in an attempt to produce a reliable estimate for the process capability.

2 Literature Review: Applications of Statistics to Manage Robustness and Quality in Pharmaceutical API Manufacture

Statistics has been widely applied in the manufacturing and process industries to increase process efficiency, improve process control and optimise the quality of the final product. Statistical methods are typically implemented to make use of the data collected from production processes, with applications including process monitoring, modelling and data based continuous improvement initiatives. In addition to specific applications, areas of discussion in the literature have included the history of how and why statistics has become an important tool within the process industries and the advantages and limitations of the methods available.

Examples of data based techniques are presented in the subsequent sections, with a particular focus on the challenges of pharmaceutical API manufacture, including process monitoring, operational excellence and process analytical technology (PAT). The influence of regulatory authorities and the importance of quality are also considered.

2.1 The Role of Statistics in Industry

The application of statistics has played an important role in the development of industrial processes during the past century, which has seen a ‘quality revolution’ of improvements to manufacturing processes. This change has been in part driven by the development of methodologies for the analysis of data (Box and Kramer, 1992, Does and Trip, 2001, Korakianiti and Rekkas, 2011). Industrial statistics has its origins with Fisher and Shewhart in the 1930s (Box, 1994). Fisher developed the basis of experimental design and the analysis of variance approach, whilst Shewhart (1931) proposed and implemented the first statistical process control (SPC) charts to visualise data and identify when a process may be moving out of statistical control (Section 2.2).

From the 1950’s, the use of industrial statistics to improve quality progressed rapidly. Deming (1986) worked within the Japanese manufacturing industry to develop methods for improving quality through the control of variation in a process, with the aid of Shewhart’s control charts (Box and Kramer, 1992). Overall there was a shift in philosophy from detecting errors through end product testing, to preventing errors from occurring through good control of a robust process. Consequently the focus has shifted from the product to the process. Deming (1986) emphasised that alongside data based

methods, quality could only be achieved through a framework of total quality management, which requires commitment from all levels of management.

In the 1980's, Taguchi developed a design of experiments methodology that potentially identifies the most efficient experimental design that maximises the information gained from the experiment (Bendell *et al*, 1999). The approach is based on selecting an experimental design that allows interactions between input variables to be quantified. The results from the experimental work are then used to build a statistical model that relates the process conditions to the outputs of the process and consequently the process inputs can be selected to optimise the outputs (Section 2.3.1).

More recently there has been an emphasis on lean and operational excellence methodologies to further optimise manufacturing processes (Karokianiti and Rekkas, 2011). Operational excellence methodologies, including Six Sigma and continuous improvement, use data-based techniques to identify opportunities to reduce variation within a process and therefore enhance the quality and consistency of the final product (Section 2.4).

In the pharmaceutical industry, there has been a push from the Food and Drug Administration (FDA) to consider product quality from the design stage of a new process by adopting Quality by Design (Section 2.3) and implementing process analytical technology (PAT) to increase the level of in-process monitoring to ensure the final quality of the product (Section 2.3.3). The implementation of QbD and PAT involves the use of data to increase the understanding of a pharmaceutical process and hence the use of statistical methods has gained importance. For example PAT methodologies emphasise the importance of process control and hence SPC and control charts are important tools in improving manufacturing processes. Pharmaceutical process can be highly complex and a large amount of data can be collected, therefore the implementation of QbD and PAT encourages the use of multivariate statistical methodologies (Chapter 3).

QbD and PAT are applicable to both new and existing processes. At AstraZeneca, QbD methodologies are driving the use of statistical methods to improve process understanding, monitoring and control. For example, prediction methods are assessed in Chapter 4 to determine how measurements taken early in a batch drying process can be used to estimate the required drying time, allowing for drying times to be optimised and an increased understanding of the process variables that affect the rate of drying. Similarly in Chapter 5, multivariate methods are used to assess the factors that influence the particle size distribution (PSD) of a product, so that the PSD can be

controlled to a level that is not expected to affect the quality of the final product. Process monitoring can be achieved through the use of process capability indices, allowing adverse changes in the process to be detected. By monitoring the capability of in-process variables, product quality can be assured by control of the process rather than through end product testing. Methods for measuring the capability of in-process variables are explored in Chapter 6 and Chapter 7.

Today a range of statistical methods are available for the analysis industrial data, including multivariate techniques (Chapter 3) and Bayesian methods (Chapter 7). However more complex methods may be more challenging to implement on an industrial process. For example, limited data processing systems may be available on the plant or a particular method may require that data satisfies an underlying assumption, such as normality.

Although many complex methods exist, Ishikawa proposed that seven basic tools for quality were the most effective for making use of the data that is collected (Karokianiti and Rekkas, 2011). The tools are Pareto charts, cause and effect charts, check sheets, histograms, scatter plots, stratification and control charts. In particular, visualisation of the data can highlight the greatest potential sources of variation in the process and help identify areas for improvement.

Banks (1993) suggested that the more straightforward statistical methods, such as data visualisation techniques, are the most effective because they can be easily implemented and interpreted. Bendell *et al* (1999) believed that the complexity of some techniques can result in statistical work being undervalued due to the challenges of interpretation. It is therefore important to find a balance between using a more complex method that is fit for purpose for the objectives of the analysis, and a method that is practical to implement and can be understood by others.

2.2 Process Monitoring

Process performance monitoring, in the form of statistical process control (SPC) charts, is used to visualise the performance of a process over time. SPC charts are used to detect changes in the process that could have resulted from failures or operational changes, leading to a process moving out of statistical control (Box and Kramer, 1992). SPC is believed to be one of the most widely applied statistical tools in the manufacturing industries (Stoumbos *et al*, 2000) and has been extended to handle multivariate data (Section 3.4.2).

Shewhart (1931) defines a process as being in a state of statistical control when, with some probability, an interval can be predicted within which future data will be expected to lie, i.e. the level of variability in the data is expected to fall within a given range. This type of variation is termed common cause variation and is inherent within the process; these sources of variation may include ambient conditions and measurement error. When a process is out of statistical control, special cause variation will result in data being observed outside of this range. For example, an excursion from normal operation may result from the failure of a component in the process or damage to a measurement probe. The aim of SPC is to determine when special cause variation has occurred, so an alarm can be raised and action taken to bring the process back into statistical control (Woodall, 2000).

SPC charts can be used to generate alarms when it is expected that the process is out of control. When the process is expected to follow a normal distribution, control limits are set based on the mean and standard deviation of the process. For example a general rule is to generate an alarm if a data point occurs more than three standard deviations from the process mean, i.e. three sigma limits. 99.7% of the data is expected to fall within this range, so a data point outside the limits may indicate a change to the process. The Western Electric Rules set further control limits (Levinson, 2010): an alarm is generated if two or more out of three consecutive points occur outside of two sigma limits, four or more of five consecutive point occur outside of one sigma limits, or eight consecutive points occur on one the same side of the process mean.

The standard Shewhart x-chart is effective at detecting unusual observations and large changes in the mean of the process; however it can be insensitive to smaller changes (Box and Kramer, 1992). Alternative charts include the exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) charts, which are more appropriate for detecting small changes to the process mean (Stoumbos *et al*, 2000). The EWMA chart tracks the mean by calculating a moving average, weighted towards the most recent data points (Roberts, 1959). The CUSUM chart plots the cumulative sum of the distances of each data point to the mean or target of the process (Johnson, 1961).

From an SPC chart it is possible to determine how capable a process is of meeting the specification limits, which can be quantified by process capability metrics. Process capability is a concise summary that enables a comparison to be performed between different parts of a process, which can be tracked over time. An underlying assumption when calculating the process capability is that the process is in statistical control, which

can be determined from control charts. Process capability is discussed in more detail in Chapter 6.

2.2.1 Opportunities of SPC

The use of SPC supports the principles of process validation (FDA, 2011). Once the critical process parameters and quality attributes have been identified, control charts can be used to confirm that the process remains in control throughout its lifecycle.

Plotting the data onto a control chart allows a visual representation of the data, so unusual trends can be identified (Box and Kramer, 1992). A control chart will indicate the extent and direction of a deviation, and suggest the time that the onset of a problem occurred, providing information that can be used for problem solving.

Pharmaceutical manufacturing processes can have a long lead time between the early stages of the process and final QC testing. By monitoring the critical process parameters (Section 2.3) continually with SPC charts, an issue that could affect quality can be detected and addressed as soon as it occurs. Therefore an adverse trend may be halted before there is an impact on quality, or the amount of affected product can be minimised.

2.2.2 Challenges of SPC

For the development of an SPC strategy, alarm rules must be set so that they are effective at detecting a genuine change in the process, but minimise the rate of false alarms. Box and Kramer (1992) suggested that standard rules for SPC can be inefficient at detecting problems, i.e. a genuine change in the process may not cause the data to fall outside of the control limits, so an alarm will not be generated.

A further challenge of SPC is that of multivariate data. When a number of correlated measurements are taken, the number of variables to be monitored can be reduced through the implementation of multivariate SPC (MSPC, Section 3.2.1.3). In addition MSPC can detect the presence of multivariate outliers, where a data point does not follow the correlation structure of the rest of the data.

There are a number of assumptions about the data that must be met for an SPC strategy to be implemented successfully. For example it is assumed that the samples are independent, i.e. there is no autocorrelation between consecutive observations and that the underlying distribution of the data is normal (Box and Kramer, 1992). Autocorrelation in the data can cause a high false alarm rate, since several data points are likely to occur on one side of the mean (Stoumbos *et al*, 2000).

Although there has been a lot of research methodologies into undertaken for control charts and SPC, such as Bayesian and multivariate methods, there may be a gap between the techniques being developed and practical applications within the process industries (Stoumbos et al, 2000). Since a wide variety of methods are available, it is necessary to identify the most appropriate techniques that can effectively differentiate between common and special cause variability (Woodall, 2000), so that SPC can be effectively implemented.

2.3 Quality by Design

Traditionally, the validation of a new pharmaceutical manufacturing process has been based on the variation observed when the initial batches were manufactured. The level of variation observed in these initial batches is judged to be a suitable range for the long term process (Yu, 2008). Furthermore, the quality of each batch is confirmed by QC testing of the final product. However, the FDA (2009) promoted a more robust approach to process development, Quality by Design (QbD), in which quality is designed into the process. The International Conference on Harmonisation Q8 (ICH Q8) guideline for pharmaceutical development suggests that:

"Quality cannot be tested into products; i.e. quality should be built in by design."

Quality by Design is a structured approach to the development of a pharmaceutical process based on scientific knowledge and information gained from data generated throughout the development of a manufacturing process. This approach enables an understanding how the variation in a process affects the quality of the product. The information gained is then used to develop a process which is optimised for product quality. The result should be a robust process that is tolerant to variation in the inputs and quality can be demonstrated by good control of the process rather than by QC testing (Yu, 2008).

The QbD process begins by defining product quality in terms of the needs of the patient. From this definition, the critical quality attributes (CQAs) of the product are identified, which are measurable characteristics of the product that describe the quality of the product. The CQAs can be identified through a risk assessment which uses scientific knowledge to determine which features of the product will have the greatest impact on quality. Specification limits are set for the CQAs which must be met for the product to be released. For an API product, the CQAs may include particle size, crystallisation structure and level of impurities (am Ende *et al*, 2007). These attributes will affect how the drug product dissolves when taken by the patient and whether the product is free from harmful impurities.

Following the identification of the CQAs, a failure modes and effects analysis (FMEA) can be implemented to determine which parts of the manufacturing process are most likely to have a significant effect on the CQAs (Brueggemeier *et al*, 2012). From the FMEA, the critical process parameters (CPPs) will be highlighted, which potentially have the greatest impact on the CQAs. The CPPs must be controlled so that the CQAs will meet their specification limits. The CPPs may include characteristics of raw materials and process conditions such as temperatures and timings.

The relationship between the CPPs and the CQAs must be well understood so that the process can be developed for the CQAs to meet their specification limits. Experimental testing and statistical modelling can be combined with scientific knowledge to build an understanding of this relationship. Two important tools in QbD are design of experiments (DoE) and process analytical technology (PAT).

To complete a QbD process, a control plan is required to ensure that the process remains well controlled throughout the duration of the product life cycle (ICH Q8). The control plan will include monitoring the critical process parameters and quality attributes, and updating prediction models with new data that is collected.

2.3.1 Design of Experiments

The concept of design of experiments (DoE) originated with Fisher in the 1930s, and was developed by Taguchi (1987) as a method for the design and analysis of a set of experiments with the aim of determining how a set of inputs to a process affect the outputs or responses. More recently this methodology is used in QbD to quantify the relationship between the CPPs and the CQAs, to determine which potential CPPs have the greatest impact on the quality of the product, and to identify the optimal processing conditions to operate a robust process.

DoE is an improvement over the one factor at a time experimental approach, in which each input is assessed individually. Studying factors individually does not allow interactions between variables to be considered, consequently important relationships between the input variables may be missed and the optimal processing conditions may not be found. When a DoE approach is implemented, the input factors are assessed simultaneously and the experiment is designed to maximise the information that is gained from as few experiments as possible (Eriksson *et al* 2008).

The first stage of a DoE process is an initial screening study. A large number of input variables can be considered, to determine which could have an important influence on the response. Typically a high and low setting is defined for each variable, to cover the range of the potential operating space. Rather than testing every possible combination

of high and low settings, a fractional factorial design is typically implemented so that enough experiments are run to assess the main effects of each variable on the response. From the results of the screening design it may be possible to discount some variables that are shown not to have an effect on product quality (Fahmy *et al*, 2012). Additionally the results may indicate whether the optimal processing conditions are within the range of the screening study, or whether the ranges of some factors should be changed for further experimental work to ensure that the optimal conditions are found.

Following the results from the screening study, a further experimental design can be carried out to provide a detailed representation of the process. Each factor may be tested at more than two levels, thereby allowing interactions and curvature to be quantified (Maltesen *et al*, 2012). Response surface modelling can be used to fit a statistical model to the results, to measure how each input to the process can affect the quality attributes. From the resulting model, predictions can be generated from specific combinations of input factors. Using an optimisation approach, various combinations of inputs can be identified that will produce optional outputs. These combinations of input settings define the optimal operating region of the process.

When the optimal operating region has been identified, the robustness of the region should be tested. Robustness testing will determine how sensitive the quality measurements are to small changes in the inputs, such as ambient conditions and raw materials. By running a final set of experiments around the proposed operating space, it is possible to determine which process inputs have to potential to adversely affect product quality and therefore which inputs require the tightest control. Designing a process that is robust to variation will ensure that high quality is maintained throughout the lifecycle of the product.

2.3.2 Design Space

From the optimal operating region, the design space of the process can be identified, which is defined as:

"The multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality." (ICH Q8).

The design space is registered with the regulatory authorities, so the process can be operated anywhere within this space. Within the design space the normal operating range (NOR) for the process will be determined, in which the process will be designed to be operated. The NOR may be influenced by business objectives, such as process

yield, cost or duration (Burt *et al*, 2011). For a robust process, the NOR will fit well within the design space, so that some variation in the inputs will not move the process out of the design space.

2.3.3 Process Analytical Technology

For the implementation of a QbD programme, new methodologies are required to be applied to enhance the understanding and control of pharmaceutical manufacturing processes. One such methodology is Process Analytical Technology (PAT). The FDA (2004) defines PAT as:

"… a system for designing, analyzing, and controlling manufacturing through timely measurements of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality."

The aim of PAT is to use scientific knowledge and statistical methods to understand the relationship between the inputs, such as raw materials and processing conditions, and the final product, which fits well into the QbD framework.

The application of PAT can involve the collection of large amounts of data, including spectroscopic (Gabrielsson *et al*, 2006, Brülls *et al*, 2003) and process measurements, such as pressure and temperature (Kosanovich *et al*, 1996, Neogi and Schlags, 1998). Multivariate data analysis techniques (Chapter 3), such as principal component analysis and partial least squares, are required to handle the data and extract the information contained within (Kourti, 2006).

From a model of the process, specification limits can be set for the critical process parameters, outside of which there is a risk to quality. SPC charts can be used to monitor the process to ensure that it stays in control and in specification. When there are a large number of correlated variables to monitor, multivariate analysis is used to reduce the dimensionality of the dataset, so that fewer variables are monitored but no information is lost, known as multivariate SPC (Section 3.4.2).

Monitoring the CPPs in real time allows a potential problem to be detected as soon it is occurs, reducing the impact on the rest of the process. In addition, when a regression model is used to predict the CQAs from the CPPs, monitoring of the CPPs allows the quality of the final product to be assured before undertaking end product testing (Chew and Sharratt, 2010).

2.3.4 Opportunities

The use of the Quality by Design framework allows for a structured, risk based approach to be applied to the development of a pharmaceutical manufacturing process. The tools allow a detailed understanding to be gained about the process, so that it can be operated to optimise the quality and reduce the risks of a quality failure.

The adoption of QbD can lead to more flexible regulatory controls (ICH Q8). The whole design space is registered with the regulatory authorities even if the normal operating range is smaller. The company then has the flexibility to make changes within this design space, for example to optimise the yield of the process, without the need for regulatory approval.

Quality by Design may be considered to be more suited to be applied to new process as they are developed, rather than to established processes. However QbD principals can be applied to make improvements to existing processes, which may benefit from a large database of information that has been collected during processing (Yacoub *et al*, 2011). For example, Lourenco *et al* (2012) used QbD and PAT techniques to define the design space of a fluid bed granulation process, with the aim of reducing the variability in granule quality. Firstly PAT and multivariate analysis was applied to process data to gain process understanding, resulting in the identification of a seasonality effect. Then a DoE approach was applied to pilot scale batches to quantify the relationship between the CPPs and the CQAs. Finally a design space was identified to optimise the CQAs. This design space was within the registered range of the process, so re-registration was not required.

2.3.5 Challenges

One of the biggest challenges when implementing QbD is the additional cost that will be incurred early on in the development of a product. Compared to traditional pharmaceutical development, more time and experimental work will be required prior to knowing the outcome of the final clinical trials and hence the product is not guaranteed to be launched. A number of published examples of QbD are for products that were terminated by the end of phase III clinical trials (am Ende *et al*, 2007, Brueggemeier *et al*, 2012).

For the development of an effective statistical model of a process, sufficient variation within the process is required so that the trends within the data can be fully captured (Doherty and Lange, 2006). Collecting this data may involve running the process outside of the normal operating range and could be done under experimental conditions with material that will not be passed on to the customer. In addition,

collecting the required data may require the installation of new sensors onto the processing plant (Doherty and Lange, 2006). To justify the additional costs, a benefit analysis may be required to estimate the expected long term savings that could be made from improving product quality and robustness.

2.4 Operational Excellence and Continuous Improvement

Quality by Design and Process Analytical Technology are suited to the application of new processes that are being developed. For processes that have been established for some time, such as those being studied at AstraZeneca, an alternative to QbD is continuous improvement techniques. The FDA promotes the use of continuous improvement technologies (ICH Q10) to reduce variation and optimise manufacturing processes.

The priorities at AstraZeneca for process improvement include reducing waste, reducing the potential for failed batches, improving the yield of reactions and maintaining a reliable supply chain to the customer. The overall aim is to reach a state of operational excellence, where consistently high standards of quality, reliability, robustness and cost effectiveness are achieved.

Achievement of these goals can be reached through a continuous improvement programme that aims to identify the areas within a process where improvements can be made. There are a number of tools available for continuous improvement, including Lean and Six Sigma. A combination of several improvement tools may be needed to fully achieve operational excellence (Kovach *et al*, 2005).

2.4.1 Lean Sigma

The concept of Lean Sigma for lean manufacturing was introduced by Toyota, who gained a reputation for producing cars with very high reliability (Liker, 2004). Lean manufacturing is an approach to manufacturing whereby the aim is to reduce waste and create a flexible and efficient process. Waste within a process can be defined by seven categories: over production, waiting, transport, inappropriate processing, unnecessary inventory, unnecessary motion of components and defects (Hines and Rich, 1997).

Toyota created the Toyota Production System, which was based on five principles of Lean (Womack and Jones, 2003, Bicheno, 2004). Firstly, value should be defined by the requirements of the customer rather than by what the company can offer. Adding features to a product that do not bring a benefit to the customer will result in unnecessary cost and processing. Secondly value stream mapping is used to assess

the whole production process to determine which steps add value to the customer. A large number of non-value adding steps will result in inefficiencies in the process. Thirdly the flow of a process should be considered, so that time is not wasted by waiting for the next stage of a process to become available. By reducing the lead time of a process, customer needs can be met quickly. Next the production line should be run based on the demand from the customer, rather than building up a large stock of intermediate or final products. Making products 'Just in Time' allows the company to respond and adapt quickly to changing demands, allowing greater flexibility of the final product. Finally a Lean process is designed to result in perfection. Systematically standardising and error-proofing a process should reduce and remove the potential for defects to occur.

2.4.2 Six Sigma

Six Sigma is a structured data based approach to process improvement with the aim of increasing customer satisfaction through focusing on improving processes rather than fixing the resulting product, so errors are prevented rather than detected (Deshpande *et al*, 1999). There is an emphasis in quantifying information, including customer requirements, so that improvements can be measured and the benefits observed (Hahn *et al*, 1999).

The Six Sigma methodology was first developed by Motorola in the 1980s, driven by a need to improve customer satisfaction to remain competitive. A Six Sigma project follows a series of five steps to achieve an improvement: define, measure, analyse, improve, control. Throughout a project, data and statistical techniques are used to identify the main sources of variability in the process and suggest where improvements are required to reduce the variation and optimise the outputs. The methods used include control charts, Pareto charts, cause and effect analysis, measurement system analysis and design of experiments. At the end of a project, there is an emphasis on standardising the process so that it will always happen in the same way and controlling improvements so that they remain in place after the project has finished.

2.4.3 Opportunities

The use of Six Sigma and continuous improvement techniques has produced substantial cost savings and quality improvements in a variety of industries, including electrical, pharmaceuticals and other high value products (Kovach *et al*, 2005). The application of Lean and Six Sigma tools to a manufacturing process allows areas for improvement to be identified and improvements to be made in a structured way so that their benefits are realised and quantified. Köksal *et al* (2011) suggested that the success of Six Sigma has been in part due to good training and affordable, user

friendly software that is available for data analysis, allowing practitioners without a strong statistics background to make use of statistical tools.

A number of pharmaceutical companies have applied Lean and Six Sigma to their manufacturing processes, resulting in cost savings from the removal of waste, an increase in quality and efficiency, and reduced cycle times (Tolve, 2009). In particular AstraZeneca have collaborated with Lean experts from a Jaguar car assembly plant to understand how Lean methodologies from the car manufacturing industry can be implemented to remove waste from pharmaceutical manufacture (Tolve, 2009). Dassau *et al* (2006) presented an example of using Six Sigma to make iterative improvements to a penicillin production process. Process capability analysis identified the least capable unit operation and then process modelling and control techniques were used to drive improvements. This process was repeated to make improvements to the poorest performing stages of the process, resulting in a 40% reduction in batch time and 17% increase in yield.

The benefits from continuous improvement projects may be expected to become smaller over time, after the most beneficial improvements have been implemented, however Box (1994) argued that rather than expecting diminishing returns, a process is constantly evolving as new inputs, technology, people and targets are available, so substantial gains can continue to be made.

2.4.4 Challenges

The setting up and implementation of a Lean or Six Sigma programme within a company requires a number of practitioners to be trained to use the tools, alongside their usual job. Therefore a successful programme requires commitment from all levels of management and also a willingness in the organisation to accept change (Nave, 2002).

When implementing the continuous improvement tools it is assumed the current process design is the most appropriate and hence the process can be optimised by making small changes rather than being re-designed (Nave, 2002). This is particularly important in the pharmaceutical sector, since any changes outside of the registered process will have to be approved by the regulatory authorities. Additionally, it is assumed that improvements will bring an overall benefit to the company, so for example an improvement will not involve high costs or substantially longer cycle times for a process.

2.5 Conclusions

During the past century, the use of statistics has played an important role in the development of modern manufacturing processes, leading to large improvements to processes that subsequently lead to an increase the quality of the final product. By understanding and controlling the variation in a process, the focus of quality control has shifted from the product to the process, so that defects are prevented rather than detected. One of the most useful tools in the quality revolution has been the control chart, which allows the variation in a process to be visualised, so that abnormal events can be detected as soon as they occur. The use of SPC charts relies on assumptions about the distribution and independence of the data, to ensure that errors are detected efficiently without a large number of false alarms.

In the pharmaceutical industry, there is a need to further develop manufacturing methods by increasing the level of process understanding, so that quality can be designed into a manufacturing process rather than tested into the final product. Through the use of risk assessment tools, including FMEA, combined with scientific knowledge, the most important parameters in the process can be identified and investigated to ensure that they are controlled to a level that is not expected to affect the critical quality attributes of the product. Using a design of experiments approach, the relationship between the inputs and outputs of the process can be identified, so that the optimal settings can be found to run an efficient and robust process.

Although many complex statistical techniques are available, in some cases the more simple methods may be most effective at extracting information from data and using the knowledge to drive improvements to a process. Continuous improvement methodologies such as Six Sigma use graphical tools and data summaries to identify and implement improvements to a process, using data to justify decisions that are made.

Overall a variety of techniques are available to make use of the data from a process, including methods for increasing understanding, making improvements and controlling a process, so that it can be run to optimise quality and robustness. Where possible, the more simple methods should be implemented initially. However when the process is complex or the data does not meet the required assumptions, then more complex methods may be necessary. In Chapter Three, methodologies are presented for the analysis of multivariate, non-linear or batch level data.

3 Modelling Methodologies

Modern process industries have the capability to record a large amount of data, both during the production process and from the final product (Nomikos and MacGregor, 1994). The data will contain a large amount of information about the process, such as how the measurements vary over time, the unit operations for which the greatest variation is observed and the differences between batches or samples. The trends in the data will show both the normal process behaviour and also indicate when a change or disturbance has occurred that could adversely affect the product. Information gained from exploring the data can be used to gain an enhanced understanding of how the processing conditions relate to the characteristics of the final product.

A wide variety of methods exist for interrogating data and extracting information about the process or product. When there are a small number of variables, methods such as summary statistics (e.g. mean, median, standard deviation) and graphical representations can be effective tools for visualising the trends in the data. Statistical process control charts, such as x-bar and range charts, are widely used to monitor a process and detect the onset of abnormal occurrences. Multiple linear regression (MLR) can be applied to establish a linear relationship between the inputs and outputs of a process, allowing the most important inputs to be identified. A prediction model of the can be applied to optimise the processes, by determining the input settings that are expected to result in optimal outputs.

MLR is a well established method for model development and is relatively simple to implement (Doherty and Lange, 2006). However when constructing a linear model, a number of assumptions are required to be satisfied for the information from the model to be accurate and useful. It is assumed that a linear relationship exists between the predictor and response variables. A linear model cannot be used to represent non-linear relationships unless appropriate transformations of the input variables can be identified. Additionally MLR is not able to handle strong correlations between the input variables (Wold *et al*, 2001), resulting in erroneous regression coefficients. It is also assumed that the residuals from the model are independent, identically distributed and follow a normal distribution with constant variance.

Advances in automation and computer technology have made it possible to collect data from a large number of sensors, throughout the duration of a process (Köksal *et al*, 2011). When sensors are located close together, the resulting variables are likely to be correlated (MacGregor *et al*, 2005). Additionally chemical processes, such as API manufacture, will involve complex interactions between the variables. As a result, the

data collected will not satisfy the assumptions required to implement methods such as MLR, consequently alternative techniques have been developed to analyse more complex data, including multivariate analysis and non-linear modelling. For each specific problem, the most appropriate method needs to be identified, to maximise the information that can be extracted from the data.

When there are a large number of correlated variables, multivariate analysis (MVA) methods are able to reduce the dimensionality of the dataset through the derivation of a smaller set of latent variables, which are a linear combination of the original variables (Section 3.2). MVA can be used to both explore the patterns within a dataset, through the use of principal component analysis (PCA), and to create a regression model between the inputs and outputs of a process, utilising multivariate regression methods, including partial least squares (PLS) and principal component regression (PCR).

A further consideration is that many pharmaceutical processes are run as batch processes as opposed to continuous processes (Doherty and Lange, 2006). For a continuous process, data is collected for each variable over time, whilst for a batch process data is collected in a similar manner, but for each batch. The resulting dataset is three dimensional, requiring multivariate methods to be adapted by unfolding the dataset into a standard two dimensional data matrix, enabling the trends both within and between batches to be identified (Section 3.3.1). Additionally pattern recognition techniques, such as case based reasoning, can be applied to compare batch profiles and make inferences about new batches using information from a set of historical batches to identify batches with similar profiles (Section 3.3.2).

Chemical processes will typically involve interactions between the variables and the presence of non-linear relationships between the measured variables and the output of the process. As a result traditional linear methodologies may not be able to provide an accurate representation of the process (Nascimento *et al*, 2000). Artificial neural network modelling is one approach for developing a non-linear relationship between inputs and outputs (Section 3.5). These models are created by finding the strongest fit to the data provided, without using any information about the underlying relationships in the process, and hence are termed black box models.

The objective of this chapter is to explore modelling techniques that are appropriate for data generated from industrial processes. In particular, the analysis methods discussed in this chapter are applied to the case studies in Chapters Four and Five. In this chapter, multivariate analysis techniques, methods to handle batch data and artificial neural networks are introduced, along with a consideration of how these methods can

implemented to a manufacturing processes, with an emphasis on pharmaceutical API production.

3.1 Multiple Linear Regression

A simple regression method to find a linear relationship between a set of k input variables (x_1, x_2,... x_k) and a single response variable, y, is multiple linear regression (MLR). The regression equation takes the form (Montgomery *et al*, 2012):

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon \qquad \text{Equation 3-1}$$

Where b_0,... b_n are the regression coefficients to be determined and ε is error. It is assumed that the error values are independent of each other and follow the same normal distribution for any value of the input variables.

Given a data set of n observations arranged in an (n x k) matrix of inputs, **X**, and an (n x 1) vector of responses, **Y**, then the vector of coefficients is found as:

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{X^T X}]^{-1} \boldsymbol{X^T Y} \qquad \text{Equation 3-2}$$

When developing an MLR model, hypothesis tests can be applied determine whether each input variable is a significant predictor for the response. Each hypothesis test takes the form:

H_0: $b_i = 0$

H_1: $b_j \neq 0$

for j = 0 to k. A p-value is calculated for each significance test and typically a p-value of greater than 0.05 will suggest that the regression coefficient could be equal to zero and hence the input variable is not a significant predictor of the response.

The level of the model fit is measured by the metric R^2, with an R^2 value close to one indicating a strong fit:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})} \qquad \text{Equation 3-3}$$

Colinearity of the input variables can be identified through the use of the variance inflation factor (VIF). For each input x_j, the strength of the linear relationship with the other variables is measured by R_j^2. Then the VIF is calculated as (Fahrmeir *et al* 2013):

$$VIF_j = \frac{1}{1-R_j^2}$$ Equation 3-4

In general, a VIF greater than ten is used to indicate a problem of high colinearity.

3.2 Multivariate Analysis Techniques

Large datasets from industrial processes will typically comprise of a number of highly correlated variables (MacGregor *et al*, 2005). For example in-process measurements including temperatures, pressures and flow rates may all follow the same trend over time since sensors located close together will exhibit similar behaviour. Additionally spectroscopic data measures the absorbance at a large number of wavelengths. The wavelengths will be correlated and hence multivariate techniques will be applicable for its analysis.

Although there may be many variables in a dataset, there are typically fewer underlying sources of variability in the process. Using MVA, the dimensionality of the dataset is reduced to a smaller number of latent variables that will represent the majority of the variation in the data set (Wold *et al*, 1987a). An appropriate number of latent variables are selected that capture the main trends in the data, while the remaining latent variables will describe the noise in the process. By summarising the data in terms of a small number of latent variables, it is easier to present the data graphically to assess the patterns between samples and identify which variables indicate the main sources of variation. When there is high variation in individual measurements, summing a number of correlated variables can emphasis the trends and reduce the variability, in effect 'averaging' out the noise.

Multivariate methods are applicable to data sets with or without a response variable. When the objective of the analysis is to explore the relationships between the variables and the pattern between samples, an exploratory data analysis technique such as principal component analysis (PCA) can be applied (Section 3.2.1). For example, PCA may be applied to a set of process measurements recorded for a number of samples, to identify which measurements are correlated and which indicate the greatest causes of variation between samples.

Alternatively, when the objective is to determine how a set of input variables may affect one or more response variables, multivariate regression techniques, including partial least squares (PLS) can be applied (Section 3.2.2). Unlike multiple linear regression, PLS models can handle correlations between the input variables and also multiple response variables. For a manufacturing process, PLS models can be used to predict

the final properties of a product using the process measurements, or to determine which variables have the greatest influence on the output variables of the process.

Multivariate analysis methods have applications in a wide range of industries, including pharmaceuticals, petrochemicals, biotechnology, telecommunications and marketing (Eriksson *et al*, 2006). A number of examples in the pharmaceutical industry are present in Section 3.4.

3.2.1 Principal Component Analysis

Principal component analysis is applied to a data set, **X**, consisting of N samples, each with data for K variables. The data is reduced to a smaller number of principal components, where the weightings of the variables in the individual components are known as the loadings, **P**. The values of the components for each sample are the scores, **T**. When analysing a PCA representation, the scores show the trends between samples and the loadings indicate the relationships between variables.

3.2.1.1 An Overview of PCA Methodology

The individual principal components (PCs) are calculated iteratively, with each subsequent component explaining a smaller proportion of the variation in the data. Details of the PCA algorithm are given in Appendix 1. The first principal component is the linear combination of the original variables that describes the direction of greatest variation in the data (Kourti and MacGregor, 1995). The coefficients of the first principal component are denoted by the loadings vector $\mathbf{p_1}$. The second principal component then describes the next largest source of variation and is orthogonal to the first.

For a data matrix **X**, the rows represent the (N) samples and the columns represent the (K) variables. Following the application of PCA, the (NxK) **X** matrix is represented as the product of two smaller matrices: **T** (NxA) and $\mathbf{P^T}$(AxK), plus an error matrix **E**(NxK) (Wold *et al*, 1987a), where A is the number of retained principal components (Equation 3-5Figure 3-1):

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \qquad \text{Equation 3-5}$$

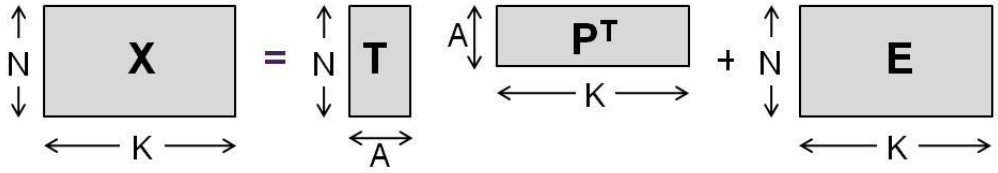


Figure 3-1: Data matrix representation of principal component analysis, with A retained components

The columns of **P** are the loading vectors that provide the weighting for each original variable in each principal component. Variables with a large loading are indicative of the causes of variability captured by a specific principal component and may suggest where tighter control is required to maintain a robust process (MacGregor *et al*, 2005). The columns of **T** are called the scores vectors:

$$\mathbf{T} = \mathbf{XP}. \qquad \text{Equation 3-6}$$

By summarising the data into a set of principal components, the observations can be examined through the scores to identify the main trends in the process. A comparison of the scores for all the observations can indicate clusters of samples exhibiting similar behaviour and also potential outliers. Through inspection of the corresponding loadings, those variables most likely to be associated with the trends in the observations can be identified.

Geometrically, PCA can be interpreted as a projection of the data points into K-dimensional space, where K is the number of variables (Eriksson *et al*, 2006). The first PC is the line in the K-dimensional space that represents the direction of greatest variation in the data. Then the second PC is the line in an orthogonal direction to the first PC that represents the next largest direction of greatest variation. The first two PCs define a plane into which all of the samples can be projected; the projections will show the scores for the first two PCs. The subsequent PCs that are found are orthogonal to the preceding components and describe decreasing amounts of variation in the data.

3.2.1.2 Data Set Selection

When applying principal component analysis, it is assumed that the data are normally distributed (Eriksson *et al*, 2006). Deviations from the normal distribution, particularly as a result of the presence of outliers can have a significant influence on the structure of the PCA representation. A strong outlier may influence the representation by causing the direction of greatest variation to be orientated in the direction of the outlier, i.e. pulling the plane towards itself and identifying a correlation structure that only applies to the one observation.

When applying multivariate data analysis it is recommended to standardise each variable to have a mean of zero and standard deviation of unity (Wold *et al*, 1987a). The aim of PCA is to identify the direction of greatest variation in the data, consequently variables that have a large standard deviation are likely to have large absolute loadings in the PCA representation. These variables may not be the most important in the process but can have a significant influence on the structure of the representation. Scaling the data so that each variable has the same level of influence on the PCA structure will allow the true structure of the data to be identified. However Wold *et al* (1987a) noted that variables that are close to being constant should not be scaled because this could introduce additional noise into the data.

From some processes, it is possible to simultaneously collect more than one type of multivariate data. For example, process variables collected at the same time as spectroscopic data can be combined to use as inputs into a multivariate analysis (Gabrielson *et al*, 2006). For this situation, the data can be scaled so that the variance of each data set is equal, and hence each source of data will have similar influence on the resulting model.

3.2.1.3 Selecting the Number of Principal Components

A PCA representation of a dataset can contain as many principal components as the minimum of the number of variables or observations. However several of these components will only contain the noise within the data and so should be excluded. For the effective application of PCA, only those components that explain the majority of the variability in the process should be retained. There are a number of ways of selecting the number of components to be retained (Vallee *et al*, 2009). The most common methods presented in the literature involve assessing the amount of variation that is explained by each component (Mercier *et al*, 2013), and calculating the effect on the error as more components are retained (Mattila *et al*, 2007).

The PCs are calculated in decreasing order of the amount of variation that is explained by each component, which is denoted R^2X (Appendix 1). Plotting the cumulative variation that is explained by including each subsequent PC will typically identify a point at which the gain from adding more components becomes small, for example four components in Figure 3-2. When a component explains a small proportion of the variation, it can be assumed that the component just explains the noise.

Alternatively cross-validation can be applied to compare the prediction errors of models with different numbers of retained PCs (Wold *et al*, 1987a). Samples are removed from the dataset either individually or in blocks and applied to the model after it has been

developed with the remaining samples (Section 3.2.1.5). This process is repeated until all the samples have been excluded once. The final model will be developed from the full set of samples.

Including more PCs to explain the variation in the data will reduce the prediction error and improve the fit of the model. However when the PCs are added that only capture the noise, the model will become over fitted and the prediction error will start to increase. The optimum number of PCs is that which minimises the prediction error or maximises the model fit to new unseen data (Eriksson *et al*, 2006). Measures of the model fit include the squared prediction error (SPE, red line in Figure 3-2) and R^2 of cross-validation, also denoted Q^2 (green line), see Appendix 1 for details. The optimal number of components may minimise the SPE or maximise Q^2. In the example in Figure 3-2 four components would be selected.

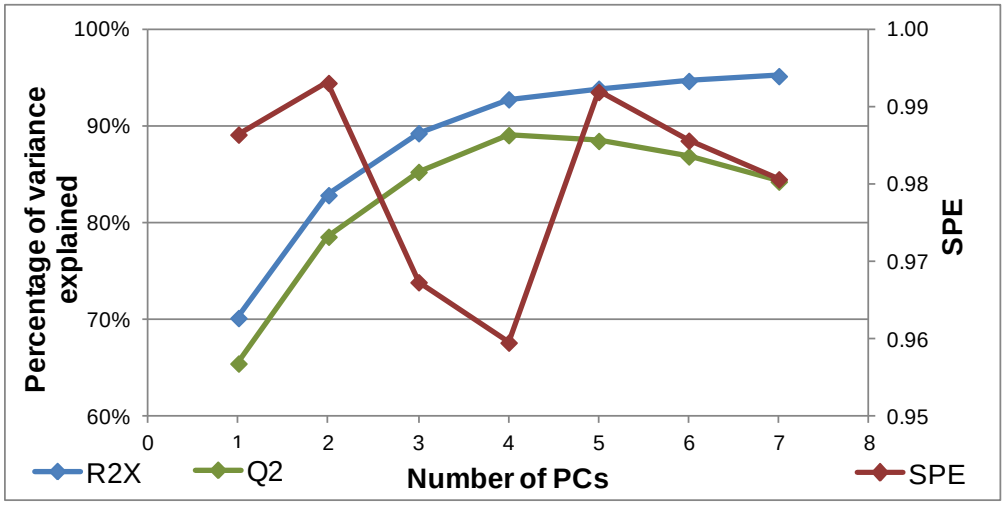


Figure 3-2: Level of fit vs. number of retained PCs, Q2 and SPE are found from cross validation

Vallee *et al* (2009) compared a number of methods for selecting the number of principal components to retain. Along with the methods mentioned above, Vallee *et al* (2009) also found the average variance explained, parallel analysis and variance of reconstruction error to be reliable methods. The first of these methods involves taking the average value of the amount of variance explained by each PC and retaining all those that explain more than the average value. In parallel analysis a second PCA is created with uncorrelated data of the same size as the original data. Then the PCs are selected that explain more than the corresponding PC in the second analysis. For the variance of reconstruction error method, each variable is reconstructed from the remaining variables to be the most consistent with the PCA model. This process is repeated as the number of retained PCs is varied. Then the number of components is selected to minimise the variance of the error from reconstruction.

Nomikos and MacGregor (1995a) suggested that if a PCA model is to be applied to future observations, cross-validation is the most appropriate method to select the number of PCs to maximise the predictive power of the model. If the PCA model is to be used to explore the current set of data, R^2X may be sufficient to indicate the required number of retained PCs.

3.2.1.4 Detection of Outliers

An assessment of how close an observation fits a PCA model can be measured through the use of Hotelling's T^2 and the distance to model in the X space (DModX). An observation with a high value for either of these two measures may not fit with the pattern of the rest of the dataset and should be investigated to identify the cause of the deviance.

3.2.1.4.1 Hotelling's T^2

The Hotelling's T^2 statistic for an observation is a weighted average of the square of the scores for the PCs that are retained in the model (Eriksson *et al*, 2006). This metric denotes the distance of the observation to the origin in the model plane and can also be used to indicate whether an observation follows a multivariate normal distribution. For an observation i, with A retained principal components, Hotelling's T^2 is calculated as:

$$T_i^2 = \sum_{a=1}^{A} \frac{(\mathbf{t_{iA}})^2}{s_{tA}^2} \qquad \text{Equation 3-7}$$

where s_{tA}^2 is the variance of $\mathbf{t_A}$, the A^{th} row of the scores matrix **T**. A large value of T^2 indicates that a sample may not be from the same multivariate normal distribution as the rest of the data. Confidence limits for T^2 are dependent on the sample size and number of retained PCs, and can be obtained from an F-distribution.

3.2.1.4.2 Distance to Model (DModX)

DModX is calculated from the residual matrix **E** and shows the distance from the observation to the plane in the X-space and is calculated for each individual principal component (Nomikos and MacGregor, 1995b), which is equivalent to the square root of the squared prediction error when the model is applied to data that was used in its construction. So for the b^{th} component and i^{th} observation, the distance to model is calculated as:

$$\mathrm{DModX_i} = \sqrt{\sum_{k=1}^{K} (\mathbf{E_{ak}})^2}$$ Equation 3-8

where $\mathbf{E_{ik}}$ is the error for the i^{th} oberservation and k^{th} variable, from a representation with b retained principal components. Confidence limits for DModX can be obtained from an F-distribution.

Hotelling's T^2 can be used to detect strong outliers that do fit within the range of the rest of the data. These observations will therefore be far from the origin and may potentially influence the structure of the PCA. DModX can be used to detect moderate outliers that do not follow the same underlying structure of the rest of the data and so may not fit closely to the model plane. A strong outlier may not have a high DModX because it can have high influence and pull the model plane towards itself. Hotelling's T^2 measures how well a data point fits with the rest of the data, whereas DModX measures how well a data point fits to the PCA model.

3.2.1.5 Application to a Prediction Dataset

Once built, a PCA representation can be applied to a new data set, $\mathbf{X_{new}}$, to assess whether the same patterns are present in the new dataset as the original (Wold *et al*, 1987a). The loadings matrix from the original PCA model is used to predict the scores for the new dataset:

$$\mathbf{T_{new}} = \mathbf{X_{new}}\,\mathbf{P}$$ Equation 3-9

The loadings and predicted scores are then used to infer the values in the new dataset:

$$\hat{\mathbf{X}}_{\mathbf{new}} = \mathbf{T_{new}}\mathbf{P}^{\mathbf{T}}$$ Equation 3-10

This method generates a set of predicted scores for the new dataset, which can be compared to the scores from the original data to see if the two datasets are similar. For each new observation, **x**, the squared prediction error (SPE) is calculated as:

$$\mathrm{SPE} = \sum_{i=1}^{K} \left(\mathbf{x_{new,i}} - \hat{\mathbf{x}}_{\mathbf{new,i}}\right)^2$$ Equation 3-11

The SPE shows how close the prediction is to the original data, so a large SPE suggests the new data is not from the same range or does not have the same correlation structure as the original data. The SPE is equivalent to the square of DModX when the PCA is applied to new data.

3.2.2 Partial Least Squares

The relationship between a set of input and output variables can be assessed by extending the PCA methodology to build a regression model, using partial least squares (PLS) or projection to latent structures (Wold *et al*, 2001). Similar to PCA, the variables are reduced to a smaller set of latent variables that are again a linear combination of the original variables. Then linear regression is applied to find the relationship between the latent variables of the input and output data sets (MacGregor *et al*, 2005). The first latent variable is the direction of maximum correlation between the scores of the inputs and responses.

Unlike multiple linear regression, PLS is able to handle data with multiple response variables, particularly when the response variables are correlated. Furthermore, a PLS model can be constructed from data containing more variables than samples and is able to handle missing data.

In a manufacturing process, PLS models can be implemented to predict the properties of the final product using data captured earlier in the process (Lopes *et al*, 2004). PLS is also widely applicable to spectroscopic methods because the data can contain a large number of highly correlated variables (Haaland and Thomas, 1988). Using a PLS model, small changes in absorbance can be detected over a narrow range of wavelengths, allowing for the concentration of a particular compound to be monitored. Examples of applications are given in Section 3.4.

3.2.2.1 PLS Method Overview

The partial least squares algorithm is used to calculate a linear relationship between a set of K predictor variables, **X,** and a single or set of M response variables, **Y**, with N observations. The algorithm has been described in many sources, including Wold *et al* (2001) and Kumar (2004). The details of the algorithm are shown in Appendix 2 and are summarised below.

First each data set is reduced to latent variables, $\mathbf{t_1}$ and $\mathbf{u_1}$, with associated weight vectors $\mathbf{w_1}$ and $\mathbf{v_1}$, such that there is the maximum possible correlation between $\mathbf{t_1}$ and $\mathbf{u_1}$:

$$\mathbf{t_1} = \mathbf{X}\mathbf{w_1} \qquad \text{Equation 3-12}$$

$$\mathbf{u_1} = \mathbf{Y}\mathbf{v_1} \qquad \text{Equation 3-13}$$

The vectors $\mathbf{t_1}$ and $\mathbf{u_1}$ are the scores of the first latent variable for the input and output datasets respectively. To relate the inputs to the responses, a linear regression is found between $\mathbf{t_1}$ and $\mathbf{u_1}$, termed the inner regression:

$$\mathbf{u_1} = b_1\mathbf{t_1} + \mathbf{e_1} \qquad \text{Equation 3-14}$$

To relate the scores back to the original variables, the loadings, $\mathbf{p_1}$ and $\mathbf{q_1}$ are found by linear regression, to satisfy:

$$\mathbf{X} = \mathbf{t_1}\mathbf{p_1^T} - \mathbf{E_1} \qquad \text{Equation 3-15}$$

$$\mathbf{Y} = \hat{\mathbf{u}}_\mathbf{1}\mathbf{q_1^T} = b_1\mathbf{t_1}\mathbf{q_1^T} + \mathbf{F_1} \qquad \text{Equation 3-16}$$

with error matrices $\mathbf{E_1}$ and $\mathbf{F_1}$.

Subsequent latent variables are found following the removal of the contribution of the first latent variable from the data sets, and the above process is then repeated.

3.2.2.2 Measures of model fit.

The effectiveness of a PLS model is quantified by the amount of variation in the data that is explained by the latent variables (Eriksson *et al*, 2006). The measures R^2X and R^2Y define the proportion of variation that is explained by the input and response data respectively (Appendix 2). Q^2 quantifies the level of fit for the predictions of the response variables, when cross-validation is used to develop the model (Appendix 2). The squared prediction error (SPE) denotes the error of predictions for new data:

$$\mathrm{SPE} = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathbf{y_{ij}} - \hat{\mathbf{y}}_\mathbf{ij}\right)^2 \qquad \text{Equation 3-17}$$

Where $\mathbf{y_{ij}}$ and $\hat{\mathbf{y}}_\mathbf{ij}$ are observations and corresponding predictions from the i^{th} observation of the j^{th} response variable.

Similar to constructing a PCA representation, the Q^2 and SPE metrics from cross validation can be examined as more latent variables are added to the model, to determine the optimal number of latent variables to be included.

3.2.2.3 Non-Linear Partial Least Squares

When constructing a PLS model, it is assumed that a linear relationship exists between the predictor and response variables. However the complexity of industrial processes may result in data that contains non-linear relationships, including curved relationships

between the input and response variables, or interactions between input variables. A number of authors have proposed adaptations of the PLS algorithm to allow the method to capture non-linear behaviour (Wold *et al*, 1989, MacGregor *et al*, 2005).

The inner regression (Equation 3-14) assumes that there is a linear fit between the pairs of scores. Hence studying plots of the relationship between the input and output scores, **t** and **u**, will indicate whether a non-linear PLS model may be required. Wold *et al* (1989) recommend that in general linear models should be implemented initially, since these are more straight forward and interpretable. If linear models do not produce satisfactory predictions and the input-output scores plots indicate non-linearity, then a non-linear model may be more appropriate. However where it is known that non-linear mechanism is present in the process being studied, a non-linear model should be fitted from the start.

In multiple linear regression models, additional terms such as quadratic or interactions can be added to the model as predictor variables. In the same way, new columns can be added to the input data for a PLS model to explain the non-linear relationships in the data (MacGregor *et al*, 2005). However Frank (1990) suggested that if there are a large number of predictor variables, then the **X**-matrix will become too large and the level of noise in the data will be increased. To limit the size of the input data matrix, it may be possible to identify those variables that are expected to show a non-linear relationship with the response, and use these variables to create additional terms. This method requires a detailed understanding of the process being studied and consequently may always be feasible.

Alternatively, to represent a non-linear relationship, Wold *et al* (1989) proposed modifying the inner regression (Equation 3-14):

$$\mathbf{u}_i = f_i(\mathbf{t}_i) + \mathbf{e}_i = f_i(\mathbf{X}\mathbf{w}_i) + \mathbf{e}_i, \qquad \text{Equation 3-18}$$

where $f_i(\mathbf{t}_i)$ can be any function that is continuous and differentiable with respect to the terms $\mathbf{w}_i$. Wold *et al* (1989) used the example of a quadratic inner relationship:

$$\mathbf{u}_i = \mathbf{c}_{0,i} + \mathbf{c}_{1,i}\mathbf{t}_i + \mathbf{c}_{2,i}\mathbf{t}_i^2 + \mathbf{e}_i \qquad \text{Equation 3-19}$$

The coefficients $\mathbf{c}_{0,i}$, $\mathbf{c}_{1,i}$, and $\mathbf{c}_{2,i}$ are estimated by least squares and $\mathbf{t}_i^2$ denotes that each element of the vector **t** is squared.

The above two methods will produce similar results when the non-linear terms are simple quadratic transformations. The first method (MacGregor *et al*, 2005) is potentially more straightforward to implement, since the non-linear terms can be added

to the data matrix and the standard PLS algorithm applied. The second method (Wold *et al*, 1989) requires the PLS algorithm to be edited to modify the inner relationship between the input and output scores. However this method is more flexible since any appropriate function can be used to model the inner relationship. To select a suitable model, cross-validation can be used to compare models of different size and complexity, to determine which can provide good predictions for new data. For either method, it is assumed that an appropriate regression can be found between the input and output scores, so these methods can only be suitable to represent weak non-linear relationships.

3.3 Analysis of Batch Data

Many industrial processes, particularly in the pharmaceutical industry, are run as batch rather than continuous processes (Doherty and Lange, 2006). Process data recorded during the running of a batch process can be collected to show the evolution of a batch from start to finish. The data generated for each batch will show a profile of how each variable changes over time and hence the resulting data matrix is three-dimensional: batch by variable by time. This information needs to be analysed appropriately to be able to understand the differences between batches.

Multivariate methods can be adapted to handle batch level data by unfolding the data matrix into a 2-D data set (Section 3.3.1). Alternatively, pattern recognition tools, such as case based reasoning (CBR) can be used to quantify the differences between the profiles of batches and use information from similar batches to make predictions about new batches (Section 3.3.2).

3.3.1 Multivariate Analysis of Batch Data

A dataset consisting of N batches, with J variables recorded over K time points is a 3-D (N x J x K) matrix. To apply PCA and PLS, the 3-D matrix must be unfolded to produce a 2-D matrix. Eriksson *et al* (2006) described two options for unfolding the data matrix, observation level or batch level, each resulting in different multivariate models being created (Figure 3-3).

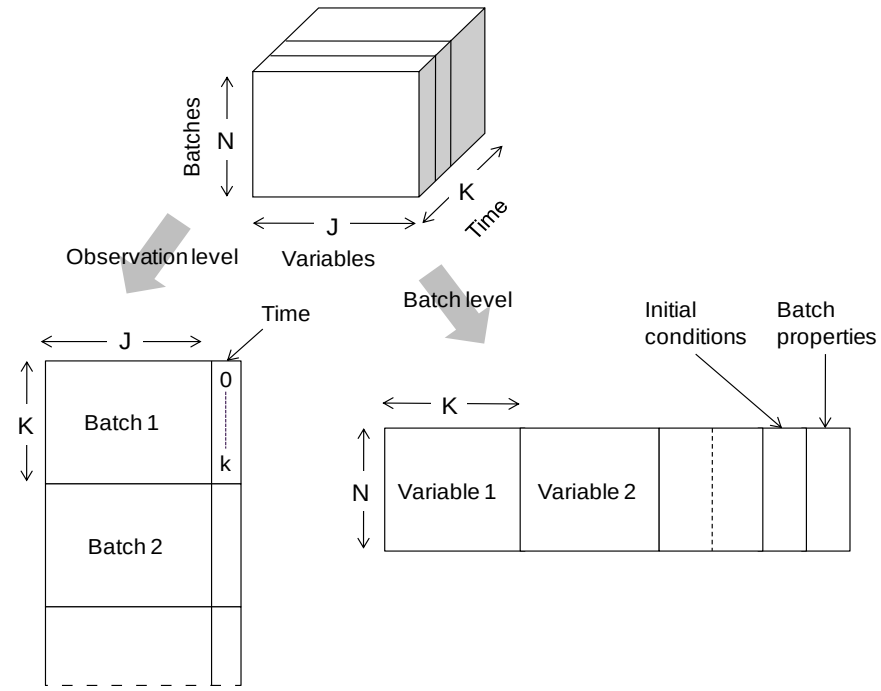


Figure 3-3: Data handling for observation and batch level modelling of 3-dimensional data

3.3.1.1 Observation Level Analysis

The trend of each batch over time can be followed by unfolding the **X** matrix, so that batches are stacked vertically, producing a (NK x J) matrix (Figure 3-3). By applying PCA to the unfolded dataset, the resulting scores will show the evolution of each batch over time (Figure 3-4), allowing unusual batches that do not follow the trend of the other batches to be identified (Kourti and MacGregor, 1995). The control limits are three standard deviations from the mean, calculated from the individual scores at each time point. The scores plot will also identify time points at which differences between batches are observed, while the loadings plot will show which variables are indicative of these differences (Figure 3-5).

Observation level unfolding can be extended to develop a PLS model in which the response variable represents the maturity of the batch (Kirdar *et al*, 2007). A maturity variable is created that indicates how far the batch is through the process. For example the maturity variable may run from 0 at the start to 1 at the end, or alternatively be represented by the concentration of a product that is forming. At any time during the running of a batch, the process variables can be used to predict the maturity of the batch, to infer how close a batch is to completion.

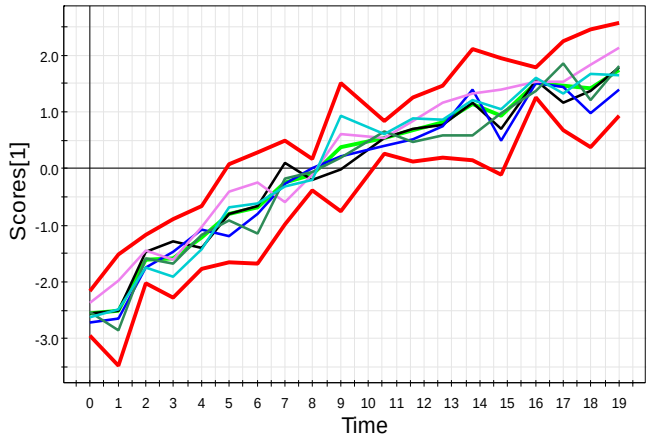


Figure 3-4: Scores plot from observation level modeling, red lines are three standard deviation limits

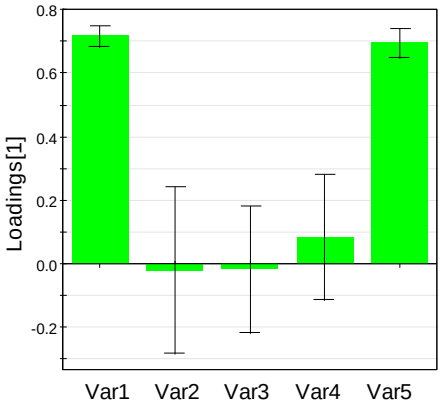


Figure 3-5: Loadings plot from observational level modeling

3.3.1.2 Batch Level Analysis

An alternative way to unfold the data matrix is to have a single row for each batch and a column for each time point of each measurement (Wold *et al*, 1987b), resulting in a (N x JK) matrix (Figure 3-3). This method is known as multi-way PCA or multi-way PLS (MPCA or MPLS).

By applying multi-way PCA or PLS, the data points for each batch are reduced so that each batch is represented by one score for each retained latent variable, allowing the differences between batches to be identified from a scores plot (Figure 3-6). The loadings will contain values for each variable and time point, so the loadings plots will highlight which variables are important at specific times during the process (Figure 3-7).

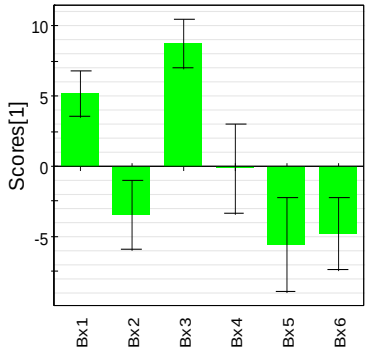


Figure 3-6: Scores plot from batch level unfolding

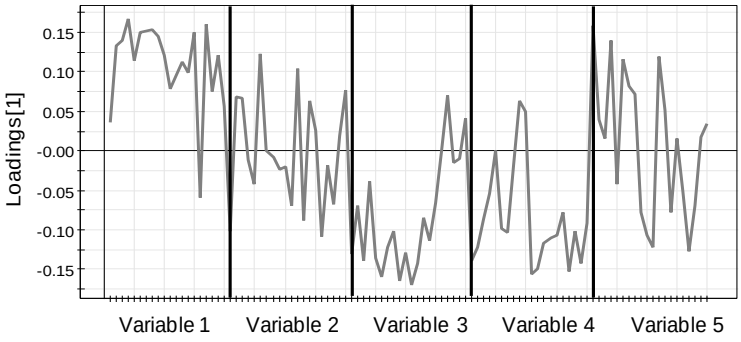


Figure 3-7: Loadings plot from batch level unfolding

When the data is unfolded for multiway analysis, additional variables can be added, for example that represent the initial conditions of the batch or inputs such as raw materials, thereby allowing non time specific information to be included in the data. Additionally, response variables can be included relating to information that is collected

at the end of the process, such as the final properties of the batch or the time required for the process to reach completion. Using MPLS, a prediction model can be constructed to infer the properties of a batch from the information that is collected during processing (Section 3.4.3).

3.3.1.3 Online Process Monitoring

Observation level modelling is particularly suited to online process monitoring. As data is collected throughout the duration of a batch, the scores until that time can be calculated without the need for data from the remaining time points. This method allows the progress of the batch to be monitored in real time and unusual behavoiur to be identified as soon as it occurs so that the source of the issue can be investigated quickly.

The multiway approach is more challenging for use online, since all of the data for one batch is used to calculate the scores. Nomikos and MacGregor (1994) suggested a method for predicting the future observations for a batch that is in progress by assuming that future deviations from the mean trajectory of the batch will remain constant over time. However the observation level approach is preferred because new batches can be monitored over time, without the need for inferring missing data.

3.3.1.4 Time Alignment

Utilising either of the above methods to compare batch profiles may require the data to be time aligned so that time points are matched up that relate to specific stages of the process. For example, when a process operates across several stages, the duration of each stage may differ between batches and hence it will be necessary to align the data from each stage.

One approach to addressing the issue of time alignment is to use a suitable indicator variable that will show how far the batch is from completion (Garcia-Munoz *et al*, 2003). The indicator variable must exhibit monotonically increasing or decreasing behaviour, an example could be the product concentration. For observational level PLS, the indicator variable will be used as the response variable. This method relies on a suitable indicator variable being available.

Alternatively, when a process has fixed stages, the data within each stage can be aligned (Ramprasad *et al*, 2008). A polynomial function can be fitted to the data of each variable over time, and the length of the stage expanded or contracted so that the stage has the same duration for each batch. The data points for each batch may be

interpolated from the polynomial function, so that each batch has the same number of data points within a particular stage. An example is shown in Figure 3-8. Three batches are shown with different durations. The length of batch one is used as the standard batch time, batch two is contracted so that the original curve is represented by fewer data points, while batch three is expanded by interpolating between data points.

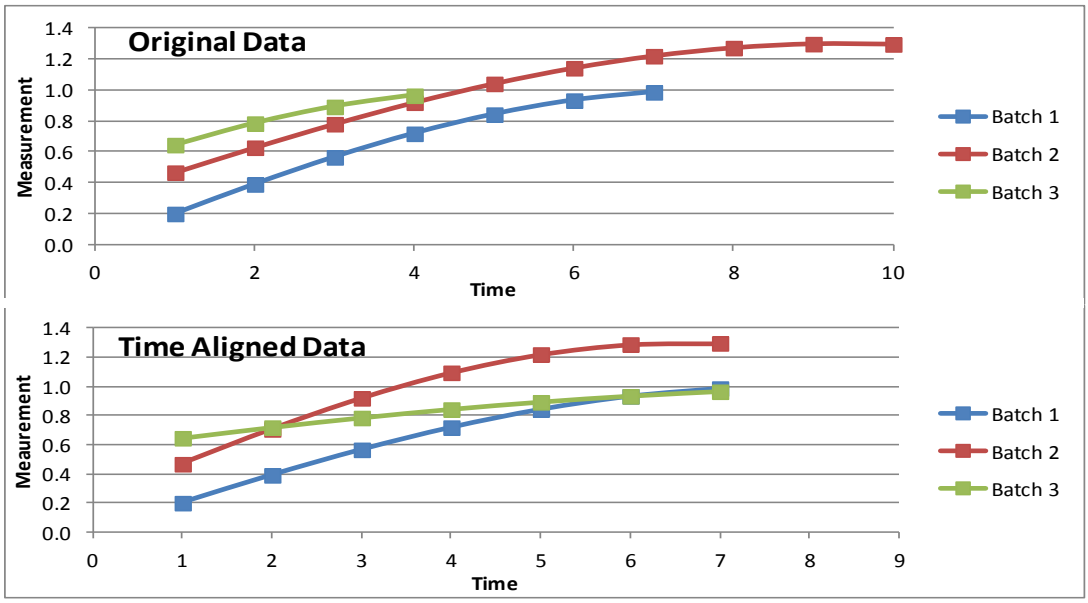


Figure 3-8: Example of original and time aligned data

3.3.2 Case Based Reasoning

All of the multivariate prediction techniques described previously assume that a linear relationship exists between the predictor variables, or transformations of, and the response variables. An alternative technique to compare batch profiles is to use the concept of pattern recognition, or example case based reasoning (CBR).

Originating from the field of artificial intelligence, case based reasoning uses pattern recognition to infer information about a new batch or case, by comparing its profile to a set of cases with known features or properties (Watson and Marir, 1994). Initially, a data set is created comprising a number of historical cases, or samples, with known features. A set of features is collected for each case, so that cases with similar features will have similar properties. Then when a new case becomes available, the features of the new case are compared to those of the historical cases, to determine the most similar historical case. The properties of the new case are then predicted to be the properties of the most similar historical case. Unlike multivariate methods, no assumptions are made about the shape of the data or the relationship between the variables.

CBR can be used to compare the profiles of one or more measurements recorded throughout the duration of a batch. For example, in Figure 3-9 the profile of the new batch is compared to each of the other batches and the selected batch is the batch with the most similar profile to the new batch. The batch properties to be predicted could include measurements taken of the final product or the duration of a unit operation. It would be expected that batches with similar profiles would have similar properties (Montague *et al*, 2008). This method was applied in Chapter 4 to a study of batch drying times. The objective is to compare the profiles of process measurements of batches with known drying times to predict the drying time of a new batch. The predicted drying time will be the drying time of the batch with the most similar profile.

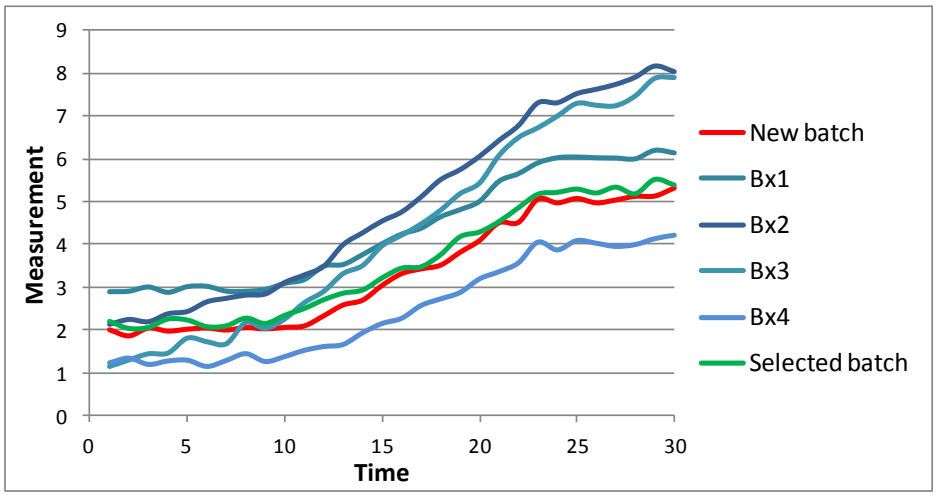


Figure 3-9: Comparison of batch profiles to select the most similar profile to a new batch

3.3.2.1 CBR Methodology

For the comparison of a set of batches, a number of features, such as process variables, are selected that will represent important information about each batch. When features are collected over time, the data will show the evolution of each batch and the time that is taken for a processing step to be completed. For example the temperature profile will show the time taken to reach a particular temperature. The same set of features must be generated for each historical batch and each new batch that is subsequently produced.

To compare two batches, the similarity of the features can be quantified by calculating the squared difference between each pair of features. For a set of features, $f_1,\ldots,f_n$, collected for a new and a historical batch, the distance is calculated as:

$$D = \sum_{i=1}^{n} w_i \left(f_i^{new} - f_i^{hist}\right)^2 \Big/ \sum_{i=1}^{n} w_i$$

Equation 3-20

The weights $w_1,\ldots,w_n$ are used to indicate the importance of each feature within the comparison, so features that are expected to be the most important are given a larger weight.

The values of the weights may be estimated from process knowledge of the most important variables. Alternatively a cross-validation approach can be applied to the batches in the historical dataset. By removing each batch individually and using the remaining batches to predict the response, the prediction accuracy can be compared as the weight values are varied, so that the weights can be optimised to minimise the prediction error.

A new case is compared to a set of historical cases by calculating the distance, D, between the new case and each individual historical case. The historical case that provides the smallest value of D is selected as the most similar to the new case and the properties of this new case are predicted to be the properties of the selected historical case.

For a batch process, a number of measurements can be collected over a period of time, which can then be used to calculate the difference between batches. For example, for two process variables, v1 and v2, collected over time points 1 to T, the difference is calculated as:

$$D = \sum_{t=0}^{T} w_1\left(v_{1,new}(t) - v_{1,hist}(t)\right)^2 + w_2\left(v_{2,new}(t) - v_{2,hist}(t)\right)^2 \qquad \text{Equation 3-21}$$

To extend to J variables, Equation 3-21 becomes:

$$D = \sum_{j=1}^{J}\sum_{t=0}^{T} w_j\left(v_{j,new}(t) - v_{j,hist}(t)\right)^2 \qquad \text{Equation 3-22}$$

Data from an industrial process may contain missing data points. At a time point when either the new or historical batch has a missing value, the term in the summation of Equation 3-22 cannot be computed. The above method can be modified so that the two batches are only compared at times points for which both batches contain data. For each comparison with a historical batch, an average is taken of the terms that can be computed, so the value of D is not biased by the number of missing data points in any batch. If the set T_j is defined as the set of time points for which both batches contain data, for the variable v_j, and T_j contains N_j points, then Equation 3-22 becomes:

$$D = \sum_{j=1}^{N_j} \sum_{t \epsilon T_j} \frac{w_j}{N_j} \left(v_{j,new}(t) - v_{j,hist}(t) \right)^2$$ Equation 3-23

3.4 Applications of Multivariate Analysis

Multivariate analysis techniques have found applications in a wide variety of industries outside of the chemical processes for which the techniques were developed (Wold *et al* 1987), because these allow relationships to be explored within correlated datasets. Applications in include marketing, sociology and finance. In sociology, characteristics of a population can be analysed to identify similarities and differences between groups of people. For example, Hutcheson and Sofroniou (2010) used PCA to analyse results of tests for various skills such as writing, coordination and memory, to highlight clusters of people with similar skills and to determine which skills are closely associated with others.

In marketing, applications of PLS models have included developing prediction models for sales forecasting (Paliwal and Kumar, 2009) and assessing the drivers of customer satisfaction (Hair *et al*, 2014). In a study into brand preferences to mobile phones, Vinzi (2010) used PLS modelling to relate customer characteristics such as demographics, along with brand identity to determine the strongest factors that influences brand preferences.

PLS has also found many applications in the finance industry, including bankruptcy prediction. For example, Yang *et al* (2011) used PLS models for feature selection of financial indicators, such as assets, liabilities and sales levels, to predict whether companies would go bankrupt in the next five years.

In a process analysis context, multivariate data techniques fit well into the framework of process analytical technology (Section 2.3.3). The can be applied to explore the data to gain an understanding of the most important sources of variation within a process (Section 3.4.1). Process measurements are identified that have the greatest influence on the final quality of the product. These measurements are then monitored in real time to detect a shift from the normal operating range as soon as it occurs (Section 3.4.2). Additionally, PLS prediction models can be used to estimate the results from off-line testing, so information about a batch can be inferred from in-process measurements ahead of test results being generated (Section 3.4.3).

Process data from sensors placed around a reactor will typically generate highly correlated measurements and hence there are many examples of the application of multivariate analysis techniques to process data. Similarly data from particle size analysis can show the frequency density of various particles sizes, which also results in data that is suited to multivariate analysis (Section 3.4.4).

3.4.1 Using MVA to enhance process understanding

Multivariate techniques have been applied a number of times to tablet manufacturing processes, to gain understanding of the relationship between the inputs and outputs. For example, Huang *et al* (2009) used multivariate analysis to assess the impact of input material properties and process variables on tablet dissolution times. PCA analysis identified that the particle size distribution of the API material exhibited the greatest variation of the input variables, and PLS identified a number of material and process parameters that had the greatest impact on the dissolution time. Tomba *et al* (2013) used a similar methodology to identify three key process parameters that determined the tablet properties on a paracetamol manufacturing process. In order to reduce the number of input variables of a PLS model of a drug product manufacturing process, Cui e*t al* (2012) systematically removed those variables that were ranked as least important and observed the effect on the Q^2 and MSE of the model. A dataset of 25 input variables was reduced to three with little effect on the model fit, and the remaining three variables were identified to be focused on for a control strategy.

Bioprocesses present additional challenges of dynamic variables and processes that vary in length. Mercier *et al* (2013) applied PCA and PLS to process data from a biopharmaceutical cell cultivation process. Online process measurements such as substrate concentrations were recorded as an average over every 30 minutes to reduce the dataset to a manageable size. Two models were created: firstly data of the first seven days, which was the length of the shortest batch, and secondly data of the first 11 days, for batches that ran for at least 11 days. Both PCA models showed clusters of batches that corresponded to differences in batches sizes. PLS models to predict the products critical quality attributes showed a low fit, suggesting that the data that was captured did not exhibit enough variation to be related to the CQAs, or other sources of variation were present that were not captured.

Observation level data unfolding can be applied to investigate how process variables evolve during the operation of a batch. Brülls *et al* (2003) used near-infrared spectroscopy to monitor the freeze drying process of a pharmaceutical product. By applying observation level PCA to the spectroscopic data, the trend over time of the scores from the first principal component gave a good indication of the rate of drying

and showed step changes that indicated when the phase of the drying process had changed.

For the fermentation of an API product, Ferreira *et al* (2007) compared an MPCA model of the process data to an MPLS model that related the process data to the final product concentration. The retained components in the MPCA captured 75% of the variability in the process data, whereas the PLS only captured 53%, suggesting that not all of the variation of the measurements that were included had an impact on the final product.

When MPLS is used to unfold batch data, each variable has a loading for each time point, so by investigating the loadings, the time periods in the process that have the greatest effect on the final product can be identified. For a fermentation process, Lopes and Menezes (2003) used MPLS to relate the process variables to the final API concentration. Loadings plots showed that measurements taken early in production had the largest weighting and hence greater control was required early in the process.

To investigate the causes of a number of out of specification quality testing results for a batch drying process, Garcia-Munoz *et al* (2003) used MPLS to combine data relating to the chemical composition of the initial product with the drying process variables, to predict the chemical composition of the dried product. The duration of each batch varied, so each batch was separated into stages and an indicator variable, such as temperature and receiver level, used to align the data within each stage (Section 3.3.1.4). From the loadings of the resulting model, the variables with the largest weights could be identified and further investigated to understand their effect on the product quality.

3.4.2 Multivariate Process Monitoring

When multivariate data is collected from a process, standard SPC charts can be extended to multivariate SPC (MPSC) charts, where the scores, SPE and Hotelling's T^2 from a PCA model are trended (Kourti, 2006). By trending the scores rather than individual variables, the number of variables to monitor can be greatly reduced without losing information contained in the data. Additionally the presence of multivariate outliers can be detected.

For continuous processes or batch processes where results are collected at one time point for every batch, data from a PCA model can be trended in the same way as for standard SPC charts. Rocha *et al* (2010) used MSPC to monitor NIR results from samples of a pharmaceutical formulation. The product concentration and impurity

levels could be estimated from the PCA scores of the NIR data, providing a rapid representation of the quality of the product. Control chart limits were calculated based on a set of 'in-control' samples, and then applied to new samples to determine when the process was moving out of statistical control.

Real-time monitoring of data collected during a batch process can be implemented through observation level PCA and PLS. Gabrielsson *et al* (2006) demonstrated how process and spectroscopic data can be combined to enhance the monitoring of a process. Observation level PLS was applied to combine data from process variables with ultra-violet (UV) spectroscopic data from a chemical reaction, with the reaction time as the response variable. The two datasets were scaled so that the sums of squares were equal for each dataset. By trending the scores throughout a batch, potential deviations in the process were highlighted, allowing the onset of problems to be detected online and resolved quickly before the completion of the batch.

For a powder blending process, Puchert *et al* (2011) used a control chart of Hotelling's T^2 to determine when the blending process was complete. From a PCA model built using NIR data of well blended samples, control limits were set for Hotelling's T^2 that identified when a batch was suitably blended. Using observation level PCA, NIR data collected online from a batch could then be applied to the PCA model and the Hotelling's T^2 values monitored until they fell inside of the control limits, suggesting that the blended process is complete.

3.4.3 Prediction models

The implementation of a prediction model of a process allows inferences to be made about the properties of the final product while the batch is still being manufactured. For batch processes, the time required for a process stage to complete may vary between batches. Prediction of the batch end point from online measurements will allow for efficient determination of the end of a process stage. End points can be estimated either by following the progress of the batch, for example by estimating the concentration of the product as it forms, or by using data collected early in the process to infer the time required to reach completion.

The use of PLS modelling for end point prediction has been found to be applicable to drying processes to estimate the required drying time. For example, Lopes *et al* (2004) used a PLS model of on-line near-infrared spectroscopy data to predict the moisture content of an API product during a drying process. The resulting model could predict the moisture content with a high level of accuracy so that the on-line analysis could replace the standard off-line laboratory test, increasing the efficiency of the process.

Bioprocesses, such as fermentation reactions to produce therapeutic proteins, have the potential to gain large benefits from the use of process analytical technology (PAT) techniques (Lopes *et al*, 2004). Variation in raw materials and seed cultures and high sensitivity to process conditions can result in high batch to batch variability and processes that are difficult to monitor and control (Read *et al*, 2009). Additionally, products and impurities from bioprocesses can be difficult to characterise, so good understanding and control of the process conditions is necessary to ensure the quality of the final product. Testing samples off-line during processing will provide limited data of the state of the process, so PAT tools can be implemented to gain information from data that is collected online, such as pH, substrate concentrations and waste gas compositions.

Furthermore, the time required for the fermentation process will be variable. Taking samples from a batch to be measured off-line creates a delay in information being available about the process. Kaiser *et al* (2008) showed how on-line measurements of fluorescence spectroscopy could be used to identify the optimal time to harvest the product from a bioprocess to produce a recombinant protein. Using PCA to reduce the spectroscopic data to two principal components, a change in the trend of the scores of the two PCs showed when the product had stopped forming and should be harvested.

For a fed batch antibiotic fermentation process, Ramprasad *et al* (2008) used an MPLS model to predict the expected batch time and process yield, using process measurements such as temperature, pH and dissolved oxygen concentration as predictor variables. The model was found to produce good predictions for the batch length, from data collect during approximately the first 10% of the required processing time, allowing downstream operations to be scheduled well in advance.

A further use of PLS models is to assess how the inputs to a process can be changed to optimise the outputs. Shi *et al* (2013) applied PLS modelling to identify the design space of a continuous hydrogenation process to manufacture an API product. Using data from historical experiments, a PLS model was developed between four process parameters and two CQAs: reaction extent and enantiomeric excess. Then an optimisation process was run by simulating the outputs across the scores space and identifying a potential design space.

Muteki *et al* (2011) demonstrated how process parameters can be optimised based on the characteristics of raw materials. For a dry granulation process for tablet manufacture, a PLS model was created using the material attributes and process parameters in input variables, and the tablet hardness and dissolution process as the

responses. Then for a batch of raw materials, an optimisation process could be run to identify the setting for the process, including the roller speed and compaction force, to ensure that the tablet properties would be within the specification limits.

3.4.4 Application to particle size data

Multivariate analysis methods are also applicable to data generated from particle size distribution (PSD) measurements. PSD data consists of a series of sizes and the corresponding frequency densities from the sample that has been analysed. Although information can be summarised in terms of the mean or percentiles of the distribution, these methods do not use all of the information that is available. The PSDs of several samples can be compared graphically by overlaying the curves of each distribution and identifying whether differences exist. However to compare a number of samples effectively, the information from each distribution can be summarised using multivariate techniques. For example, principal component analysis could be used to identify the differences between samples by comparing the scores and loadings plots. If a small number of PCs are able to explain a large amount of the variation in the data, then little information will be lost by summarising the data.

Ma *et al* (1999 and 2000) proposed using PCA to detect small quantities of large particles from PSD data obtained by laser diffraction measurements (Section 5.2.5). When collecting data, several sweeps are taken of a sample and the particle size distribution is obtained by taking the average light intensities recorded by each of a number of detectors. The signal will fluctuate across several sweeps due to movement of the particles in the measuring zone. If there are a small number of large particles present they will not be detected in every sweep, so their signal may be lost when the results are averaged. Ma *et al* (1999 and 2000) applied PCA to raw data from laser diffraction measurements to identify the sweeps that detected the large particles. Datasets were used for the laser diffraction analysis of aluminium oxide powder and a simulated dataset. The use of PCA allowed the individual sweeps to be analysed, rather than the average. It was seen that the second or third principal components were able to detect the presence of a small number of large particles. This method could be useful to detect large particles that may affect the content uniformity of a small number of tablets. When analysing the PSD of a pharmaceutical powder, a small amount of large particles will have a large impact on the content uniformity.

Mattila *et al* (2007) developed a multivariate statistical process control approach by applying PCA to particle size data obtained from on-line laser diffraction on a mineral processing plant. Several datasets were compared and in each case two or three components were sufficient to explain 99% of the variation in the data and to minimise

the prediction residual sum of squares. Control charts of the squared prediction error and Hotelling's T^2 were used to detect disturbances in the process. Reducing the data from the whole particle size distribution into a small number of PCs enabled the process to be monitored efficiently and disturbances to be detected quickly.

Sandler and Wilson (2009) applied PCA and PLS to understand the relationship between particle size and shape measurements, and the downstream packing behaviour of a pharmaceutical product. PCA of the size and shape measurements highlighted clusters of samples that used the same input material. Then PLS models to predict flow and density characteristics revealed that a small amount of highly circular particles can have a large effect on the behaviour of the particles.

3.4.5 Application of Case Based Reasoning

Montague *et al* (2008) showed how CBR and multi-way PLS could be applied to compare batch profiles from the first part of a batch to infer information about the end point of the batch. A lager fermentation process was considered and CBR and MPLS were used to predict the time that the process would finish so that the next batch could be prepared in advance. Temperature and alcohol measurements were recorded during the first part of the brewing process for a number of batches. For CBR these readings were compared to the same measurements from a new batch and the most similar historical batch used to predict the end time. Both CBR and MPLS produced comparable results for this case study. However when these methodologies were applied to a more complicated fed-batch pharmaceutical fermentation process, the CBR results showed a higher level of accuracy, exhibiting a 30% reduction in mean squared error compared to MPLS. This difference may be a result of the MPLS model assuming that a linear relationship exists between the input and response variables, which may not be the case. The case-based reasoning method does not rely on the data satisfying any assumptions, so is more suited to handling non-linear data.

3.5 Artificial Neural Networks

Creating a representation of a process using linear modelling techniques relies on the assumption that there is a linear relationship between the input and output variables, over the range that is being studied. For many complex processes, such as manufacturing processes, a linear model is not appropriate for producing a meaningful representation of the process (Sukthomya and Tannock, 2005). An alternative approach is to use mechanistic models based on first principles of physical or chemical relationships in the process. However these models can be difficult and time consuming to produce for complex processes (Hussain, 1999).

Artificial neural networks are a modelling approach for calculating non-linear relationships between input and output variables. Rather than creating a model based on the expected shape of the relationship, a set of non-linear functions are defined and optimised to find the highest level of fit to the data. Other than selecting the input variables to include in the model, no process knowledge is used to determine the structure of the model, so neural networks can be considered to be a black box modelling method (Wilcox and Wright, 1998). Although other modelling techniques, such as PLS, are based on the relationships inherent within the data rather than physical relationships, there is an assumption of linear relationships, whereas no such assumption is made for neural networks. In general, as less process knowledge and scientific principles used to create a model, a larger amount of data is required (Figure 3-10).

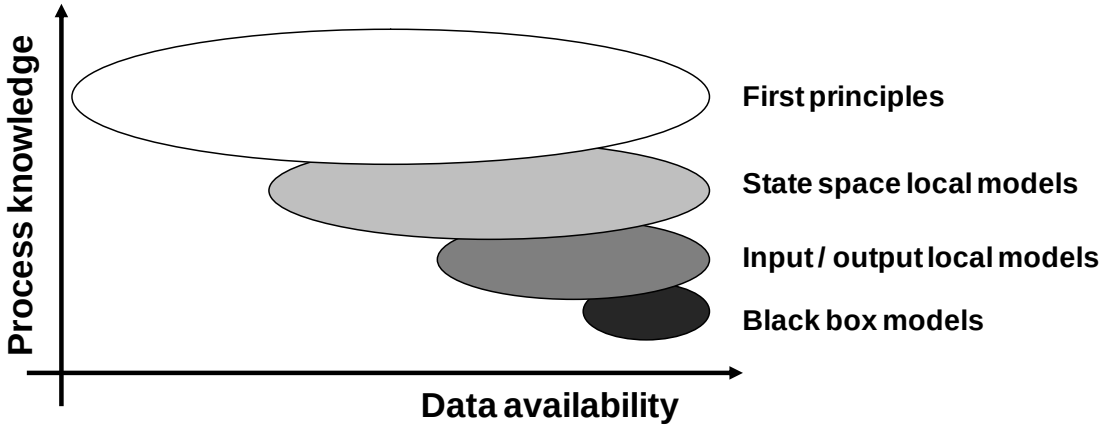


Figure 3-10: Model types and data requirements

The development of neural networks has progressed over the past 40 years, beginning in the field of neuroscience, to represent the complex neurology of the brain (Grossberg, 1988), with applications including speech and image recognition (Lippmann, 1987). Since then, neural networks have found uses in a wide range of industries, including engineering, finance and management (Vellido *et al*, 1999).

3.5.1 Methodology

An artificial neural network is a collection of interconnected nodes that take the data from input variables, apply non-linear functions and produce an output that is a prediction of the response variable. The parameters of the network can then be optimised using an iterative process, to determine the best fit to the data that is available.

3.5.1.1 Network Structure

Within a neural network, a node is a single processing unit (Himmelblau, 2000), Figure 3-11. Multiple inputs into the node are summed to produce a weighted sum. A transfer function is then applied to the weighted sum to provide the output from the node. The transfer function can be any mathematical function, but is typically a continuous non-linear function, such as the sigmoid function (Figure 3-12). During training of the network, the weights of the input variables are adjusted to provide the optimal predictions for the response variables.

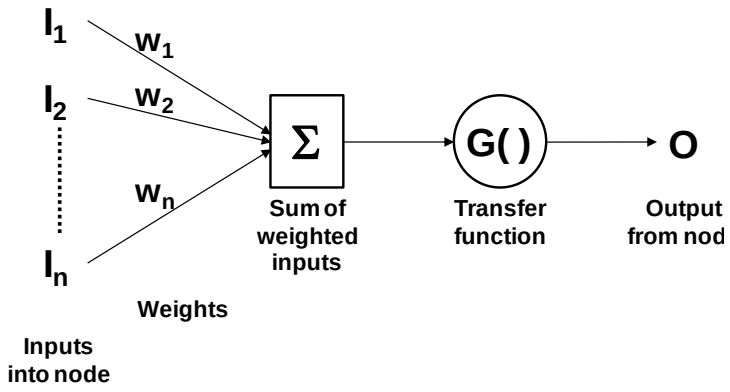


Figure 3-11: An individual node in a neural network

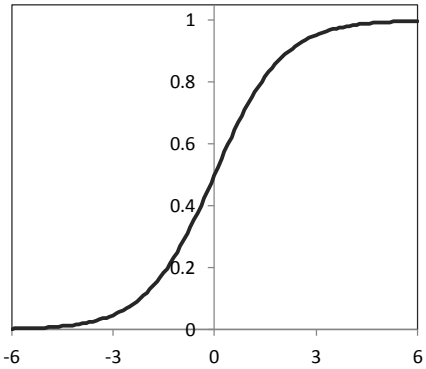


Figure 3-12: The sigmoid function

The nodes of a neural network are arranged in layers, with nodes connecting to other nodes in adjacent layers. Overall, information is passed from the input layer, through one or more hidden layers to the output layer (Figure 3-13). The input layer takes the signal into the network and, if necessary, scales the data appropriately for the transfer function that is used. There will be as many nodes in the input layer as there are input variables. The hidden layers are used to find the relationship between the data in the input and output layers. Finally the output layer is used to provide the output of the network, or the prediction of the response data. The output layer will contain as many nodes as there are response variables. The hidden and output layers each contain an additional bias node, which are constant values used to allow non-zero outputs to be produced from a zero values input.

The simplest form of a neural network is a feed forward network with three layers: input, hidden and output (Himmelblau, 2000). In a feed forward network information flows in one direction from the input to the output layer. Alternatively in a recurrent network, information can be fed backwards to previous hidden layers.

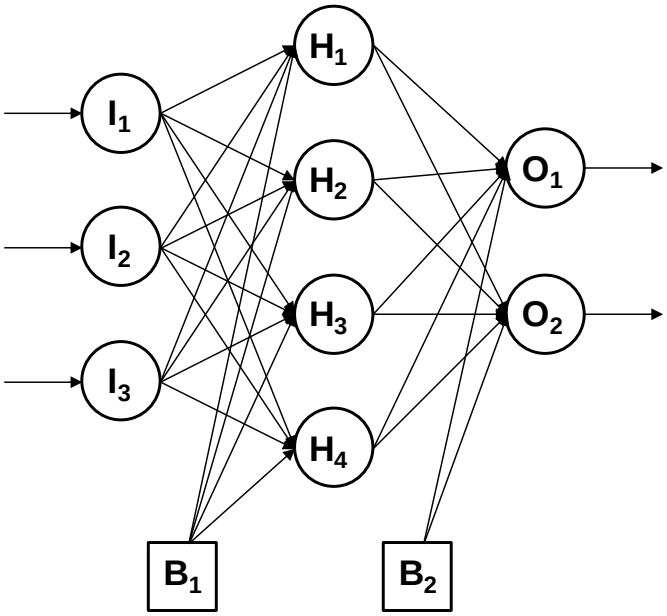


Figure 3-13: Example of a network with 3 input variables, 4 hidden nodes and 2 output variables

3.5.1.2 Network size

The number of hidden layers and number of nodes in each hidden layer will depend on the complexity of the process being modelled. Since neural networks are a black box modelling approach, the topology of the network cannot be found from the expected relationships of the variables being modelled (Wilcox and Wright, 1998). For a successful network, enough nodes are required so that a reasonable fit can be found between the inputs and responses. However a network that is too complex may become over fitted, with some of the nodes modelling the noise in the data, resulting in poor predictions when the network is applied to new data. Although there is no rigorous method to find the ideal topology for a network, systematically adding or removing nodes or layers from a network and comparing the error level when applied to new data may suggest the optimal network structure for making predictions (Sukthomya and Tannock, 2005).

3.5.1.3 Training, Validation and Test data sets

As well as limiting the number of nodes in a hidden layer, the number of inputs into the network should be limited to variables that are expected to have a relationship with the response variables. Including potentially uninformative inputs may add noise to the network and increase the level of error in the predictions (Behzadi *et al*, 2009). It is useful if the number of samples in the data set is larger than the number of weights for which values need to be assigned, to reduce the level of over-fitting (Nascimento *et al*, 2000).

When the structure of the network has been defined, the weights of the inputs to a node must be optimised using an iterative process. This stage is known as network training. For the training of a network two data sets are required, training and validation, each containing samples of the input and response data (Himmelblau, 2000). Firstly the training data is used to optimise the values of the weights. After each iteration, the network is applied to the validation data to measure the error when applied to unseen data. The training algorithm is stopped when the error of the validation data starts to increase, suggesting that the network is becoming over fitted to the training data. An example is shown in Figure 3-14; the error of the validation data set is minimised after the fifth iteration, so the model produced at this point would be the selected to apply to new data.

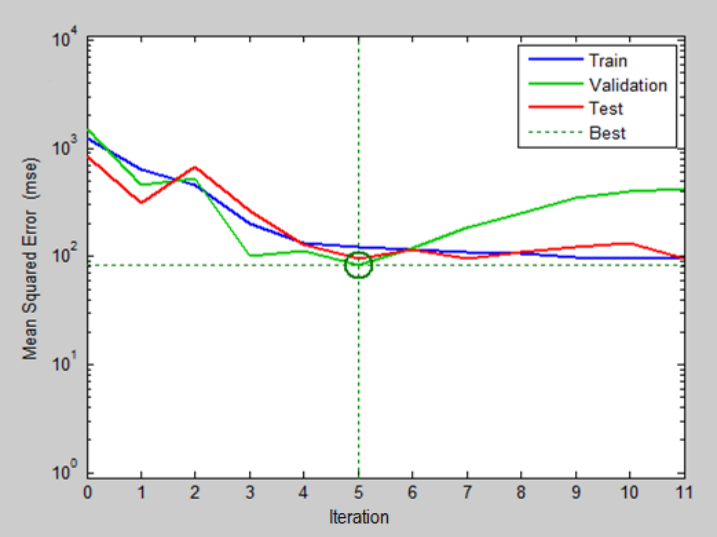


Figure 3-14: Mean squared error during network training

The training data must cover the range of the input and output data for which the model is to be used, since neural network models cannot extrapolate outside of the training data range (Himmelblau, 2000). In the literature, suggestions for the ratio of training to validation data range from one to one, to four to one (Sukthomya and Tannock, 2005). Additionally a set of testing data can be used to compare different models, by testing their performance when applied to new data that has not been used in the development of the model.

3.5.1.4 Training Algorithms

A training algorithm is used to update the weights in the network. The most common training algorithm for a feed forward network is the back propagation algorithm (Sukthomya and Tannock, 2005). The network is trained using a set of training data and then applied to validation data. The algorithm proceeds as follows (Lippmann, 1987, Himmelblau, 2000):

1. Initialise weights with small random numbers
2. Apply the input data to the network to generate predictions for the output data
3. Calculate the error between the actual and predicted output
4. Calculate the gradient of the error with respect to each weight individually, to determine whether increasing or decreasing each weight will reduce the error
5. Starting with the output nodes, adjust the weights to optimise the error
6. Repeat from step two until the error of the validation data starts to increase

The algorithm is named back propagation because updating of the weights starts with the output layer and works backwards through the network. Since each training session begins with randomly generated numbers, the final weights in the model differ each time the algorithm is executed.

3.5.1.5 Dataset selection

The data set used to train a neural network model must cover the whole range of inputs that may be fed into the model (Himmelblau, 2000). If suitable data is not used to train the network then the model may produce poor predictions (Karim *et al*, 2003) and will not represent the process enough to provide useful information.

When possible, the data set can be taken directly from data that is automatically collected during the normal running of the process. Using readily available data will avoid any extra costs and may include a large number of data points. However to model the process thoroughly, a wider range of input data may be required, potentially outside of the normal operating range of the process. For example when optimising a process, the optimal settings may be not be included in the normal operating range and so may not be represented by the model (Sukthomya and Tannock, 2005).

For the whole of the potential operating space to be covered, a designed experiment can be used to explore all possible combinations of input variables, resulting in a balanced data set with which to train the model (Coit *et al*, 1998). However the time and cost of running an experimental programme may limit the amount of data that can be collected, particularly if the product produced during the experimental work cannot be sold. A possibility is to obtain data collected from normal processing and then to use experimental work to cover areas of the operating space that are not included in the process data.

When a mechanistic model of the process is available, an alternative approach is to simulate the data that would be collected from a designed experiment (Nascimento *et al*, 2000). The data from a computational model has the advantage of being free from

noise that would be observed in data collected on the process (Sukthomya and Tannock, 2004). Once developed, a neural network model is faster to run than a first principles model and so can be used in an optimisation algorithm to find the ideal settings for the process (Hussain, 1999, Mohammed and Zhang, 2013).

3.5.2 Stacked Neural Networks

When a neural network is trained, the training algorithm starts with a random set of numbers for the weights. Consequently each network that is trained will result in a unique set of values for the weights and will produce different outputs for the same input data (Zhang *et al*, 1997). The temptation may be to train a number of networks and then select the network that produces the minimum prediction error when applied to the testing data. However, the structure of a neural network is flexible so it is possible that a model can become over fitted to the data this is used to create the network and hence the model may not give good predictions when applied to unseen data.

For good predictions, a neural network model is required to generalise well so that the trends in new data will be represented by the model. One solution is to create a number of networks and combine the outputs of each network, known as stacking. The concept of generalising by stacking was introduced by Wolpert (1992), who proposed running an algorithm a number of times and taking an average of each output. The aim of stacking was to achieve generalisation accuracy rather than learning accuracy, so that the resulting model is applicable to all data, rather than just the data used to create the model (Sridhar *et al*, 1996). When a neural network is trained a number of times, each resulting model may capture a different aspect of the process behaviour, so improved predictions may be achieved by combining the outcomes of all of the models.

A stacked neural network is created by training a number of networks ($f_1, f_2, \ldots, f_n$) and the output of the stacked network is a weighted sum of the individual network outputs (Figure 3-15). So for input data, **X**, and weights $u_1, u_2, \ldots, u_n$, the output of the stacked network is (Sridhar *et al*, 1996):

$$f_{stack}(\boldsymbol{X}) = \sum_{i=1}^{n} u_i f_i(\boldsymbol{X}) \qquad \text{Equation 3-24}$$

The stacking weights, u_i, could be found through multiple linear regression. However the outputs from each network would be expected to be correlated, so it may be more appropriate to use principal component regression (PCR). Zhang *et al* (1997) proposed the following methodology:

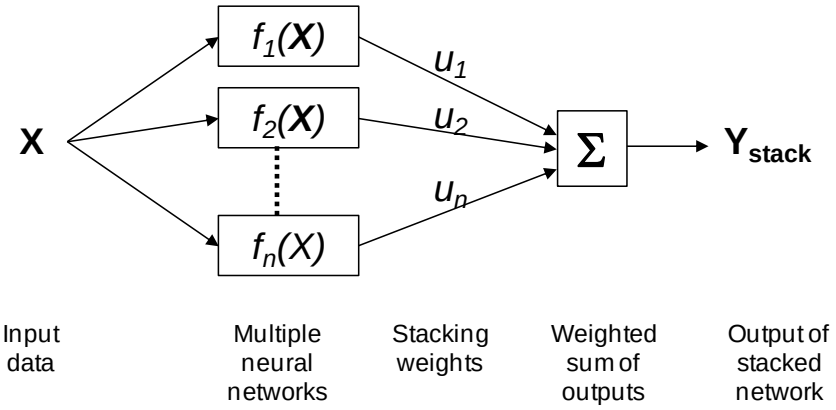


Figure 3-15: Stacked neural network

Create *n* neural networks, each time randomly resample the data to form different training and validation datasets. Create a matrix, $\hat{\boldsymbol{Y}}$ of the outputs from each individual network, $\hat{\boldsymbol{Y}}_1, \hat{\boldsymbol{Y}}_2, \ldots, \hat{\boldsymbol{Y}}_n$:

$$\hat{\boldsymbol{Y}} = \left[\hat{\boldsymbol{Y}}_1, \hat{\boldsymbol{Y}}_2, \ldots, \hat{\boldsymbol{Y}}_n\right]$$

Equation 3-25

Then the output from the stacked neural network is:

$$\hat{\boldsymbol{Y}}_{stack} = \boldsymbol{Yu}$$

Equation 3-26

where $\boldsymbol{u}$ is a vector of weights, $u_1, u_2, \ldots, u_n$.

Using principal component analysis, decompose $\hat{\boldsymbol{Y}}$ into scores, $\boldsymbol{T}$, and loadings $\boldsymbol{P}$. By inspecting the variance explained by each principal component, retain *k* components that are shown to explain the majority of the variation in the data (Section 3.2.1.3), so that:

$$\boldsymbol{T}_k = \hat{\boldsymbol{Y}}\boldsymbol{P}_k$$

Equation 3-27

Then from Equation 3-26, let $\boldsymbol{u}=\boldsymbol{P}_k\boldsymbol{\theta}$, for some (kx1) vector $\boldsymbol{\theta}$, so:

$$\hat{\boldsymbol{Y}}_{stack} = \hat{\boldsymbol{Y}}\boldsymbol{P}_k\boldsymbol{\theta} = \boldsymbol{T}_k\boldsymbol{\theta}$$

Equation 3-28

Since $\boldsymbol{T}$ is the scores matrix, the columns are orthogonal and therefore not correlated, so $\boldsymbol{\theta}$ can be found by least squares regression:

$$\boldsymbol{\theta} = (\boldsymbol{T}_k^T\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k^T\boldsymbol{y} = \left(\boldsymbol{P}_k^T\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\boldsymbol{P}_k\right)^{-1}\boldsymbol{P}_k^T\hat{\boldsymbol{Y}}^T\boldsymbol{y}$$

Equation 3-29

where $\boldsymbol{y}$ is the vector of actual responses. Finally the weights are found as:

$$\boldsymbol{u} = \boldsymbol{P_k}\left(\boldsymbol{P_k^T \hat{Y}^T \hat{Y} P_k}\right)^{-1} \boldsymbol{P_k^T \hat{Y}^T y}$$ Equation 3-30

3.5.3 Application of Neural Networks to Manufacturing Processes

Artificial neural networks have been widely applied in the manufacturing and process industries. The main advantage of using neural networks is the ability to learn the relationships present in complex non-linear processes, without the need to develop mechanistic models (Nascimento *et al*, 2000). In addition, there is a requirement that the data comes from a normal distribution (Lippmann, 1987).

3.5.3.1 Identification of important process variables

When constructing a neural network model, no indication is given of the importance of each variable for generating a good prediction of the response. An effective method to assess the importance of the input variables is to remove each variable systematically and measure the effect on the error when the model is applied to validation data (Sukthomya and Tannock, 2005). An increase in the error when a variable is removed suggests that the input is needed to predict the response, conversely when the error remains the same or reduces, the removed input may only be contributing noise and should be excluded from the model.

Nascimento *et al* (2000) made use of this approach to reduce the number of input variables for a neural network model of a chemical production process. A total of 31 input variables were reduced to 13 by comparing the model error with and without each variable. Behzadi *et al* (2009) applied neural networks to the validation of a pharmaceutical granulation process to determine which process conditions had the greatest influence on the properties of the final product and therefore required greatest control. The final model was used to assess how variation of each input variable would affect the properties of the product.

Neural networks can be applied as part of a predictive process control system (Section 3.5.3.3), however an initial task is to identify which variables to include and the structure of the model that will be used. For a spray drying process, Neshat *et al* (2011) created a neural network model to predict the weight of the granules that were produced, with inputs including the density and viscosity of input materials, and the temperature and pressure of the process. Then correlation analysis of the variables was used to determine an improved structure of the model, whereby the processing conditions were used to predict the exhaust air temperature, which was them combined with the input material characteristics to predict the granule weight. Gaining

understanding of the relationships between the variables then allowed a more effective predictive control system to be developed.

3.5.3.2 Process optimisation

Once a suitable neural network representation of the process has been developed, the model can be used to determine the optimal values of controllable inputs into the process, to produce the desired level of quality, energy use or cost efficiency of the process. In some cases, a first principles model can be used in a similar way to optimise a process. However the equations for a first principles model may take time to compute the predicted response, so the optimisation process will require a large amount of computer power. Once a neural network model is constructed, it is quick to run the model multiple times through an optimisation algorithm to find the required levels for the process inputs.

Nascimento *et al* (2000) presented an example of a polymerisation process for which a neural network model was used to analyse multiple combinations of seven input variables to produce the optimal values of three quality measurements on the final product. Firstly a mechanistic model was developed and combined with experimental data to produce a noise free simulation of the process. Simulated data was then generated to cover the whole operating space and then used to fit a neural network model. The resulting model produced a large number of predictions over a grid of process conditions, enabling the identification of the settings that would produce a high quality product, while also considering the cost of the process. The optimal settings were successfully implemented on the industrial process.

In a semi-conductor manufacturing process, Chou and Chen (2012) applied a design of experiments approach to collect data to be used to develop a neural network model. The process being studied involved depositing a dielectric material between two metal layers. Inputs variables included the composition of the material and the temperature of the process and the outputs were the defect rate of the devices produced, results of a voltage stress test and the amount of fluorine retained in the devices. Following the collection of data, a neural network model was fitted that allowed the optimal settings for the input variables to be determined, resulting in an increased yield and reduced production costs.

A further advantage of neural networks is the speed that models can be run, enabling the continual optimisation of the controllable settings for a process based on measurements that are collected during a run. Mohammed and Zhang (2013) demonstrated how this approach could be applied to a polymer moulding process, by

controlling the temperature set point. Mechanistic models were found to be time consuming to run and hence not suitable for online control. The profile of a batch was divided into seven stages and at each stage the degree of cure was predicted from the temperature and degree of cure from the previous stage. Then predictions of future stages could be used to predict the final degree of cure from the final stage of the process and an optimisation process run to find the optimal temperatures at which to run the remaining stages. A stacked neural network consisting of 30 individual networks was developed to improve the accuracy of predictions to new data.

3.5.3.3 Process monitoring and control

In the literature, a common field for the application of neural networks in manufacturing is for process monitoring and control (Sukthomya and Tannock, 2005). When a mechanistic model is too complex to run, a neural network model can be used for online predictive control. A neural network model can be used to predict the future response of the system, over a specified time and a control signal calculated based on the expected deviation from the target of the controlled variable (Hosen *et al*, 2011). For example, for a packed distillation column, MacMurray and Himmelblau (1995) showed that a neural network approach to predictive control could provide a higher level of control than a first principles model. For a polystyrene manufacturing process, Hosen *et al* (2011) compared the performance of a combined first principles and neural network model to a conventional PID controller. The neural network approach achieved smoother control with less variation in the controller output.

Statistical process control is used to detect faults in a process by identifying unusual patterns in the data. Neural networks can extend this method by attempting to link unusual patterns to specific faults in a process (Zorriassatine and Tannock, 1998). A neural network model can be established to detect specific abnormal patterns in the data, such as a shift or drift, and then to produce an output of the likelihood that a particular change has occurred (Guh, 2007). The type of change may then be used to determine the cause of the abnormal behaviour of the process (Cheng, 1997).

Bioprocesses may be particularly difficult to represent with first principles models because they involve complex processes that vary between batches (Karim *et al*, 2003). However, good control and fault detection is required to run an efficient process. For a protein production process, Karim *et al* (2003) implemented a neural network model to estimate the process yield online based on process conditions such as temperature and substrate concentrations. This method allowed a faster estimate of

the yield to be achieved compared to waiting for an offline analysis. Consequently a problem in the process could be identified and reacted to more quickly.

Zhang (1999) applied stacked neural networks to a batch polymerisation process, with the aim of inferring product quality from in-process measurements, such as temperatures. Only nine batches were available to train the network, of which five were used as the training set and four as the validation set to determine when to stop the training algorithm. Since the dataset size was limited, bootstrap resampling with replacement (Efron and Gong, 1983) was used to select the training data for each network that was created. In addition, two unseen test batches were applied to the models after they had been constructed.

In total 30 individual networks were created and combined to form a stacked network. Using principal component regression, two components were retained and the resulting loadings matrix used to determine the weightings for each individual network within the stacked network. Comparison of the individual models showed that those with the lowest mean squared errors (MSE) for the training and validation data did not have the lowest MSE for the testing data, suggesting that the individual models were not robust for application to new data. However when a stacked neural network was created, a consistently low MSE could be achieved for both the training and unseen data sets. Comparison of the number of individual networks within a stacked network found that the MSE reduced as more individual models were added into the stacked network, until the error stabilised at around 20 networks. Zhang *et al* (1997) recommended that 30 networks are used to construct a stacked network. Stacked neural networks have been found to achieve greater generalisation to the dataset and produce more robust predictions, hence stacked networks are applied in the case studies in Chapter 4 and Chapter 5.

3.6 Conclusions

Data generated from industrial processes may not meet the assumptions required to use traditional multiple linear regression methods. Many alternative methods are available to make use of the data that is collected, including multivariate analysis for handling correlated data and artificial neural networks for modelling non-linear relationships. Multivariate methods can also be adapted to represent non-linear trends. For a batch process, when data is collected throughout the duration of a batch, multivariate methods can be extended to handle batch level data. Alternatively case based reasoning can be implemented to quantify the similarity of batch profiles without requiring the assumption of linearity.

The methodologies described in this chapter have found many applications to industrial processes. Representing the relationships between variables aids process understanding and provides information of the inputs that have the greatest impact on the product. Online predictions allow off-line measurements to be estimated and monitored in real time. Prediction models between the process settings and the final quality allow the optimal process conditions to be identified, to maximise the productivity of the process. For batch processes, the end of a processing stage can be predicted from data collected early in the batch or by continually predicting a maturity variable that will indicate when a process is complete.

To make effective use of the data that can be collected from a process, the most appropriate methods must be identified for the data analysis. The methods described in this chapter are generally more complicated to use than traditional linear methods, so a good understanding of the methodology is required to make use of the results that are produced. It may be beneficial to start by using the most simple methods and using these results to determine whether a more complex technique is required. For example, assessment of the correlation between variables will determine whether multivariate methods are required. The pattern of the residuals from linear models will indicate whether non-linear methods should be used.

The majority of the case studies presented in Section 3.4 and Section 3.5.3 applied one type of modelling technique, although Montague *et al* (2008) compared MPLS and CBR to two case studies. In the case study in Chapter 4, four different methodologies are applied to the same dataset, to identify which is the most appropriate for the process being studied. Initially multivariate models were considered appropriate, since the data is multivariate in nature. However, reducing the number of variables to a smaller number of measurements that are taken directly from the process data may result in a prediction model that is more straight-forward to run and more intuitive to be implemented on the production plant. Both linear and non-linear methods are also compared to determine the level of complexity that is required.

Where possible, examples have been presented of applications within the pharmaceutical industry. However many relevant applications have been found from other sectors, including mineral processing (Mattila *et al*, 2003), polymerisation (Nascimento *et al*, 2000) and semi-conductor (Chou and Chen, 2012). The majority of the applications to pharmaceutical process were for secondary manufacturing, for example Huang *et al* (2009) and Behazadi *et al* (2009), or bioprocesses, for example Merica *et al* (2013) and Karim *et al* (2003). In the literature there are significantly fewer

applications to small molecule API production, although examples were found with Shi *et al* (2013) and Sandler and Wilson (2009).

Therefore in the subsequent chapters of this thesis, the statistical methodologies presented in this chapter are applied to industrial process data collected from API manufacture at AstraZeneca. In particular methods are applied to predict the duration of a batch drying process (Chapter 4) and to represent the relationship between process variables and the particle size distribution of an API product (Chapter 5) Methods of estimating the process capability of an API process are investigated in Chapter 6 and Chapter 7.

A specific challenge of the case study presented in Chapter 4 is the alignment of batch data when stages within each batch vary in duration (Section 3.3.1.4). Garcia-Munoz *et al* (2003) used an indicator variable to represent the extent of process. Ramprasad *et al* (2008) aligned the start and end times of each stage and used a polynomial fit to interpolate the data. However in the case study in Chapter 4, the duration of each stage is a result of the timing of a manual operation rather than the rate of the process. The rate of the drying process increases at the beginning of each stage and slows towards the end. Therefore in this case study it was more appropriate to remove data from stages than ran for longer than usual, and leave gaps in the data for stages that ran for less time than usual (Section 4.6.1.1).

In Chapter 5 PLS is used to investigate how process variables may impact on the PSD of the final product. The majority of examples of using multivariate data analysis to gain process understanding make use of spectroscopic data and data from process variables, such as temperatures and chemical compositions (Section 3.4.1). Fewer examples show how particle size distribution data is also multivariate in nature and therefore applicable to multivariate modelling methods. Sandler and Wilson (2009) used PLS to analyse how the PSD can impact on the downstream packing process. Conversely in Chapter 5, MVA is applied to determine how the process variables can impact on the PSD.

4 Modelling of Filter Drying Times

4.1 Introduction

Data collected during a manufacturing process can provide valuable information about the outcome of a particular stage of the process. Through the application of regression methods, a prediction of the outcome can be attained while the process is running. The outcome of the process could be a quality characteristic of the final product or information on how a particular processing step will progress to completion. The methodologies presented in Chapter Three, multiple linear regression, artificial neural networks, partial least squares and case based reasoning, were found to be useful for predicting the outcome of a process from the in-process measurements (Section 3.3 and 3.4.3).

Of the manufacturing processes being studied at AstraZeneca, one particular unit operation for which improvements were required was the filter drying process. The filter dryers are used to drive off excess water from the solid product, until the water content of a sample is below a target specification limit. This test is used to ensure that the batch will pass the product strength test during the final quality control (QC) testing.

Variation is seen in the performance of the filter dryers and the rate of drying for different batches. Batch drying times typically range from 50 hours to 90 hours and occasionally batches require drying for as long as 200 hours. The drying times increase over a period of time until a decision is taken to clean the filter; following a clean shorter drying times are observed. The drying process can be the rate limiting step of the overall process, so reducing the drying time would allow the capacity of the plant to be increased.

4.1.1 Aims of the Case Study

A large amount of data is collected online from measurement probes around the filter dryers. In particular the temperature inside the dryer and the flow rate of nitrogen gas passing through the filter are expected to be related to the rate of drying. There is the potential to investigate these data sources to gain more information about the variation in the drying times. The aim of this chapter is to investigate which measurements taken during the drying process are indicative of the rate of drying and then to build regression models to predict the drying time from the in-process measurements.

The data collected from sensors on the plant will be investigated to determine which variables indicate the progression of the drying process, and whether there are early

indications of how long a batch will be required to be dried to pass the water content test. From a plant management perspective, it is useful to know in advance when a slow drying batch is expected, so that plant resources and upstream operations can be planned accordingly. Additionally, arrangements can be made in advance for the dryer to be cleaned at the end of a slow drying batch.

Following the identification of the most important variables, a number of prediction methods were assessed to determine if it is possible to predict the drying time of a batch with better accuracy than the current method of following the profile of the outlet gas temperature. Both linear and non-linear models were investigated, along with methods that either utilise data from the whole profile of the batch, or use specific data points within a batch. Ideally a model should be simple enough to be implemented on the plant so that predictions can be generated automatically, however more complex methods will be considered to determine the greatest level of accuracy that can be achieved in a modelling context.

If overall drying times can be reduced, the capacity of the plant would be increased, which is especially important during busy times in production. Additionally, the amount of energy required to dry each batch would be reduced. By ensuring that the product is produced with a consistent final water content, the milling and formulation stages of the process would potentially be easier to control, since there will be less variation in the input material.

4.1.2 Drying Process

The drying process is used to drive water off from the solid product to increase the purity of the API product. Prior to drying, the solid product is formed as a powder from a precipitation reaction, and then the batch is transferred to the filter dryer. The powder is collected onto the filter and the remaining liquid is collected in the filter receiver. Following the batch transfer from the precipitator, a fixed quantity of purified water is washed through the precipitator and dryer to the receiver. A further two washes of purified water are passed through the filter dryer to the receiver. The level of liquid in the receiver can be monitored to measure how quickly the water passes through the filter cake, indicating the resistance of the product in the cake (Section 4.2.2).

Figure 4-1 shows a schematic of the filter drying process and the location of the inlet and outlet gas temperature and N_2 low rate sensors. During the drying process, warm nitrogen gas is passed through the filter cake to drive off the remaining water. The temperature of the gas is controlled to 40^{o}C by the Apovac system to prevent the product from changing form. The flow rate of the nitrogen gas, and the inlet and outlet

temperatures were identified by the process technical experts as the most likely to be indicative of the drying time; these variables are discussed further in Section 4.2. Measurements are recorded every ten seconds. Three filter dryers are run in parallel; the dryers are identical in design and set up, although some differences have been noted between the dryers.

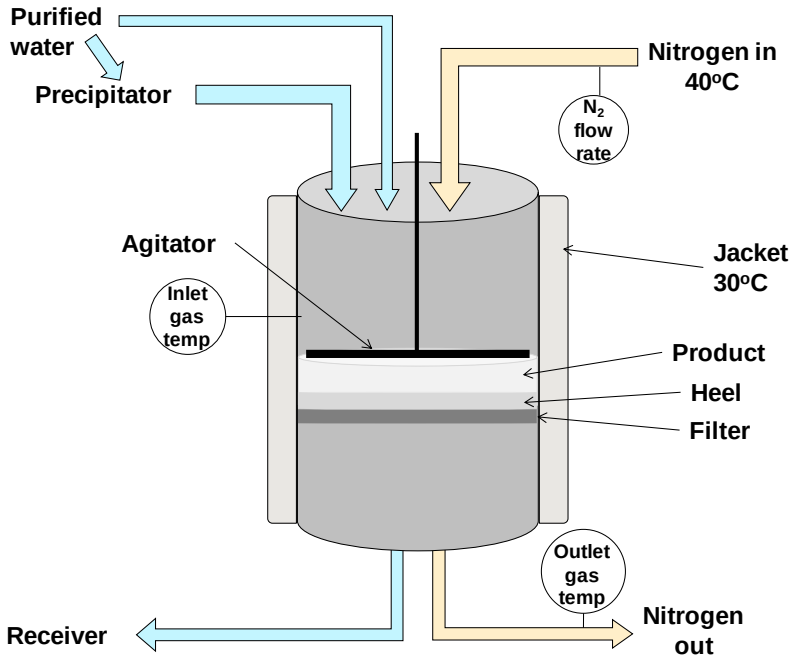


Figure 4-1: Schematic of the filter drying process

At fixed points during the drying process, the agitator is used to mix the product to ensure that the batch is evenly dried and also to allow more water to be removed. As more water is removed, the gas is cooled and the outlet temperature falls (Figure 4-2). The agitator is applied after 18, 30, 36, 42 and 48 hours of drying, and again when the gas temperature reaches 28.5°C.

When the outlet temperature reaches 29.5°C, it is assumed that no more water is being removed so a sample is taken to be analysed for water content, using a loss on drying (LOD) test. The reported LOD measurement quantifies the amount of material that is left when the water is removed. When the LOD result is found to be above the specification limit (95.3%), the drying is stopped and the batch removed from the dryer. If the LOD result is below the 95.3% target, drying is restarted and the time until the next sample is taken is determined by the first LOD result; for a lower result, more time is left until the next sample, which is specified in the batch sheet.

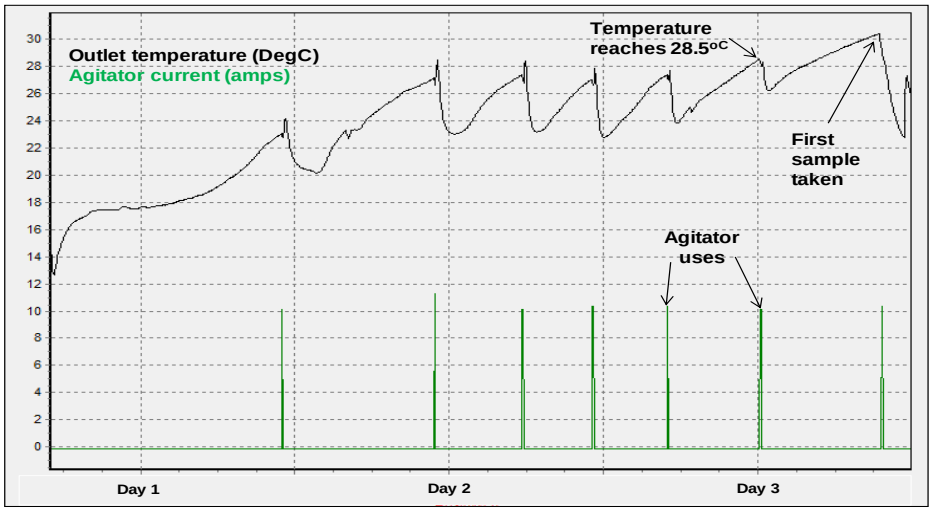


Figure 4-2: Typical outlet temperature profile, with agitator uses.

When the drying process has finished the batch is removed using the agitator. Since the agitator does not touch the base of the filter, a small layer of product remains on the filter, known as a heel. The heel consists of around 10% of the material from one batch. Over several batches the amount of material in the heel builds up and compresses, causing the rate of drying to slow down. Eventually the heel is washed off by dissolving the remaining material in a solution and a new heel will form with the next batch. Typically the heel is washed off every ten to 30 batches.

The initial loss on drying result can be as high as 97.5% for fast drying batches, suggesting that some batches are dried for longer than optimal with this sampling plan. However taking samples earlier to determine the drying end point more precisely can be undesirable. More specifically, taking a sample from the filter cake, the distribution of material can be disrupted, resulting in damage to the soft amorphous particles. If a crack forms in the cake, nitrogen gas can pass through the crack rather than the rest of the cake, resulting in slower drying. Additionally, each time a sample is taken, the flow of nitrogen gas through the dryer is stopped, the sample analysed and then the nitrogen is restarted, adding around an hour to the drying process. The ideal drying scenario is for the LOD test to be passed with the first sample, and with an LOD result just above 95.3%.

4.1.3 Physical Models of a Drying Process

Greater understanding of the mechanics of the drying process can be gained by studying physical models of drying and filtration processes. These models will indicate which inputs and characteristics of the process may indicate the rate of drying.

4.1.3.1 Heat Transfer Model

Nere *et al* (2012) proposed a heat transfer model of an agitated filter dryer for an API process, highlighting that use of an agitator promotes heat transfer across the material on the filter and the timing of the agitator can be adjusted to optimise the drying process. This study focuses on the heat and mass transfer as heat travels to the product from the wall of the vessel, causing the solvent to separate from the product. Therefore the temperature of the vessel jacket can affect the rate of the drying, although it is important to maintain the temperature below the melting point of the product.

To understand the kinetics of the drying process, the process can be separated into three phases (Nere *et al*, 2012). Phase 1 is the removal of the unbound solvent from the cake material, which occurs as the product is transferred onto the filter and the initial gas flow begins. Phase 2 consists of desolvation, the separation of solvent into a liquid form, and evaporation of the liquid solvent. In phase 3 the evaporated solvent is carried away by inert gas. It is assumed that gas flow rate is high enough so that phase 2 is the rate limiting step of the process. The rate of desolvation and evaporation depend on the heat that is transferred from the vessel wall. In the case study at AstraZeneca the inlet gas is also heated and will provide heat to the material.

In the proposed model (Nere *et al*, 2012), the material nearest to the vessel wall will dry first, and then heat is transferred through the dry solids to the wet solids, which are next to dry. There is assumed to be a temperature gradient from the vessel wall (T_{wall}) through the dry solids to the wet solids (T_{bed}), which have a constant temperature profile. The heat transfer rate, Q, is calculated as:

$$Q = UA(T_{wall} - T_{bed})$$

Equation 4-1

where A is the heat transfer surface area. U is the heat transfer coefficient, which depends on the heat transfer coefficients of the first layer of solids at the vessel wall (h_{solid}) and the rest of the solid bed (h_{bed}):

$$\frac{1}{U} = \frac{1}{h_{solid}} + \frac{1}{h_{bed}}$$

Equation 4-2

It is assumed that the material of the vessel wall has high thermal conductivity and hence provides no resistance to heat transfer.

In this model it is assumed that the removal of unbound solvent does not limit the rate of the process. However in the case study as AstraZeneca, compression of the heel on

top of the filter may restrict the flow of gas and limit the rate of drying. Therefore the resistance caused by the filter and material must also be considered.

4.1.3.2 Filtration Model

Richardson *et al* (2002) describe the mechanics of a filtration process. Factors that affect the rate of filtration include the pressure drop across the filter, and the resistance of the filter cake, filter medium and initial layers of the cake. In this case study, the heel on the filter constitutes the initial layer of the cake. The level of resistance will vary from batch to batch, depending on the orientation of the particles in the cake and the extent to which the particles block pores in the filter cloth. For an individual batch, the resistance caused by the cake and the heel will affect the gas flow rate during drying, and therefore the rate of drying that will occur.

The batch that is transferred to the filtration vessel consists of the solid precipitate and the solvent. The resistance to flow increases as the product is builds up on the filter, and can be observed by the rate at which the remaining solvent flows through the filter.

Richardson et al (2002) derive the following model to describe the flow rate of the filtrate through the filter as the cake build up:

$$\frac{1}{A}\frac{dV}{dt} = \frac{-\Delta P}{rl\mu} \qquad \text{Equation 4-3}$$

V is the volume of filtrate that has passed through the filter at time t, A is the area of the filter, ΔP is the pressure drop across the filter, r is the specific resistance of the material, l is the cake thickness and μ is the viscosity of the filtrate.

From Equation 4-3, dt/dV describes the resistance to flow, which increases as the cake thickness builds up. If the volume of filtrate is proportional to the amount of cake deposited on the filter, then V is proportional to l. When the filtration is run a constant pressure, there is a linear relationship between the resistance and the volume of filtrate (Richardson *et al*, 2002). By plotting t/V against V as the material builds up on the filter and fitting a linear relationship, the intercept represents the resistance at the start of filtration, caused by the filter cloth and heel, and the gradient represents the resistance of the filter cake.

In this case study, the filtrate volume can be measured by the liquid level that is recorded in the filter receiver, so the resistance can be estimated by monitoring the receiver level as the batch is transferred. Richardson et al (2002) commented that plotting t/V against V does not produce reproducible results because the inputs to the calculation depend on the exact timings at the start of the operation. However for the

case study at AstraZeneca, when the batch transfer is complete the flow rates of the subsequent water washes may provide information of the overall resistance of the batch and hence information of the expected drying time. Measuring the wash flow rate was found to produce consistent results and is described further in Section 4.2.2.

4.1.3.3 Conclusions from Physical Models

A physical model of the drying process in this case study must consider the heat transfer from the vessel wall and the nitrogen gas, as well as the resistance to gas flow caused by the material in the batch and the heel. The models presented above suggest that temperature, pressure and flow rate are key parameters in the drying process. These variables can be measured or estimated from sensors in the dryers.

Nere *et al* (2012) recommend several settings that can be adjusted to optimise drying, however changes to the process are outside of the scope of this project. The focus for this case study is on empirical modelling to predict the drying time. The potential input variables and modelling methods are presented in the remainder of this chapter.

4.1.4 Modelling Methods for the Prediction of Drying Times

A range of modelling and prediction methods were discussed in Chapter 3 that could be applied to the drying times data. The most straightforward approach would be to take a limited number of measurements from specific points in the drying process as input variables and create a model with the drying time as the response. A linear model would be the most practical to implement on the production plant, since no statistical software will be required to implement the model. However many industrial processes exhibit non-linear characteristics and hence neural network models may provide a better representation of the process, resulting in improved predictions for new unseen data.

A constraint of some linear models is that the input data points are required to be uncorrelated to satisfy the assumptions of the methods. In this study the data is recorded at regular intervals throughout the drying process and hence the profile of the batch can be monitored, rather than summarising the profile in terms of individual data points. Multivariate methods such as principal component analysis (PCA) and partial least squares (PLS) can handle variables with multiple time points, by reducing the size of the data set to a smaller number of latent variables. An alternative approach is case-based reasoning (CBR), based on quantifying the differences between batch profiles and does not require any assumptions about the trends within the data.

The aim of modelling is to predict the drying time with a high level of accuracy, using variables measured early in the drying process. This approach would remove the need

to take LOD samples to detect the drying end point, thereby reducing drying times by preventing batches from being dried for too long and removing the need for multiple LOD samples to be taken before drying is complete. A further benefit of increasing the understanding of the drying process would be the optimisation of the timings for performing a clean to remove the heel on the filter, so that the filters are not cleaned more frequently than necessary but very long drying times are prevented.

4.2 Variables Associated with Drying Time

The goal of the modelling task is to predict drying times early in the process, so data collected close to the start of the drying process should be used. In addition, to build a linear or neural network model, the trends in the data must be summarised to individual data points that are not strongly correlated with each other. A number of process variables are considered which are expected to be related to the drying time of a batch. These variables have been identified from discussions with process technical experts and from previous work undertaken to investigate the drying times. The variables considered are nitrogen flow rate, water wash flow rate and inlet and outlet gas temperatures and pressure.

4.2.1 Nitrogen flow rate

The flow rate of nitrogen during the drying process will indicate how quickly water is being driven off from the product by the circulating gas. However, it is known that as the heel on the filter becomes older, the rate of drying appears to slow. As the powder in the heel becomes more compressed, the flow of nitrogen through the batch is restricted and hence less water is driven off.

Figure 4-3 shows the N_2 flow rates during the drying of two batches, a fast and a slow drying batch. The fast drying batch is following the removal of the heel, so the nitrogen gas flow is fast, resulting in a short drying time of 52 hours. The slow drying batch is the 18^{th} batch on the same heel, so the N_2 flow rate is reduced, resulting in a drying time of 83 hours. The N_2 flow rate is stopped temporarily when the agitator is used to mix the filter cake, and it may resume at a lower rate if the agitator removes cracks in the product, through which gas can pass. The flow rate can also drop towards the end of the drying process as the cake becomes more compact.

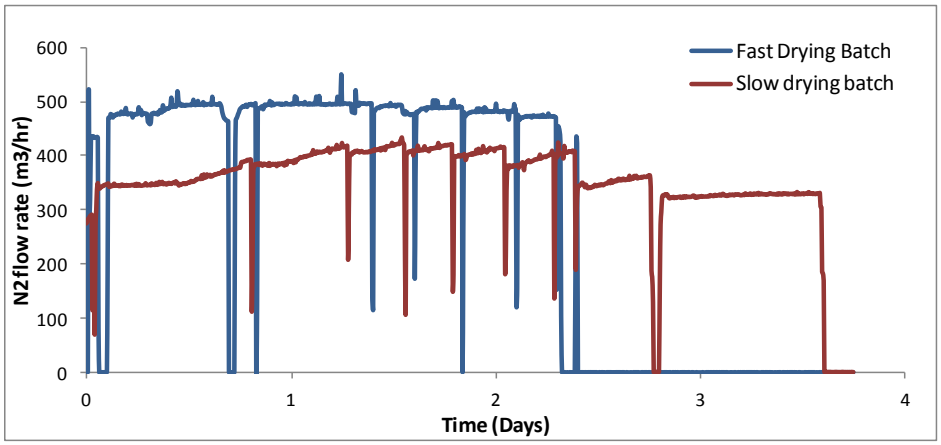


Figure 4-3: N_2 flow rates of a fast and slow drying batch

Measuring the N_2 flow rate at the start of a batch will give an indication of the flow rate that is expected throughout the duration of the batch and will therefore suggest the time that the batch will take to dry. A potential summary of the N_2 flow rate profile is to determine the average flow rate between each agitator run. However, batches with a low flow rate at the start tend to have a similar flow rate for the remainder of the batch. Hence if the flow rates between each agitator use are collected for several batches, correlations would be expected within the data collected for each batch and only one data point could be used as an input to a linear model.

A further constraint of using the N_2 flow rate in a prediction model is that one of the filter dryers, dryer one, does not have an in-range flow meter, and hence an alternative input is required for this dryer.

4.2.2 Wash flow rate

An alternative measure of the resistance of the heel is to assess the flow rate of the water that is washed through the product after the batch is transferred to the filter. At this time, purified water is washed through the product and collected into the receiver. The rate of change of the level in the receiver indicates the flow rate of the water through the cake and how much resistance is caused by the material on the filter. By measuring the flow rate of the filter washes, useful information is gained at the start of the process, before drying commences. In addition, the data of the wash flow rate measurement is more straightforward to collect than the average N_2 flow rate between agitator runs.

Figure 4-4 shows a typical profile of the receiver level as the batch is transferred to the filter and water is washed through. The time taken for each wash to be collected in the receiver can be used to measure the flow rate through the filter. By zooming in around

a particular wash, it can be seen that the change in level over time could be used to calculate the wash flow rate in m^3/hr (Figure 4-5).

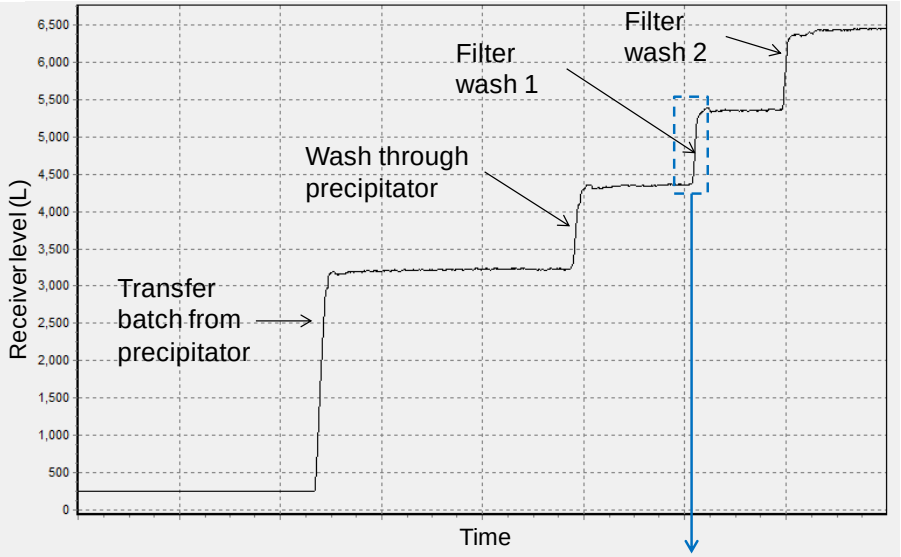


Figure 4-4: Receiver level as batch is transferred to filter

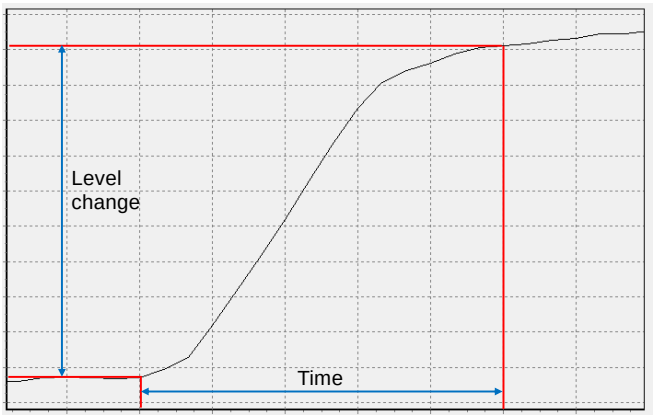


Figure 4-5: Measurements for flow rate calculation

Over the build up of a heel, the wash flow rate drops as more batches are run in the dryer (Figure 4-6). For this particular heel, the drying times are not seen to increase until around 16 batches have been dried, after which the drying times for subsequent batches increase rapidly. When two consecutive batches require more than 80 hours of drying the heel is washed off.

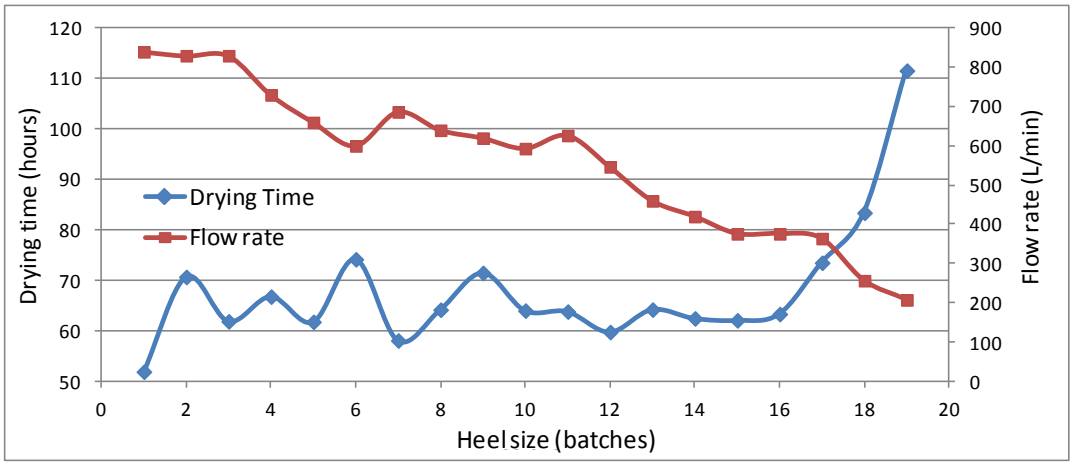


Figure 4-6: Drying times and wash flow rates over one heel

4.2.3 Temperature profile

The temperature inside the dryer is measured by temperature probes on the inlet and outlet gas streams (Figure 4-1). The inlet temperature is generally more constant throughout a batch, although variation is seen in the readings when the gas flow is stopped and the agitator applied (Figure 4-7 and Figure 4-8).

The outlet gas temperature is also affected when the agitator is in use, but overall it increases during the drying of a batch. When the temperature reaches 29.5^{o}C the drying is expected to be complete. Batches that dry more quickly generally have a lower outlet temperature during the initial stage of drying, suggesting that more water is being driven off the batch and cooling the nitrogen gas (Figure 4-7). Conversely batches that take longer to dry have higher temperatures at the start of drying, since less water is being driven off by the nitrogen (Figure 4-8). These are the same batches as shown in Section 4.2.1.

Since the outlet temperature is generally not constant throughout a batch, taking the mean over a time period may not provide a useful summary of the temperature data. An alternative summary is to compare fixed points during drying to describe the temperature profile. The agitator is generally started up at fixed times from the start of the process, so the temperature measurements around these times could be used to compare between batches. For example the temperatures immediately before and after agitation will be considered, along with the minimum temperature between agitator runs.

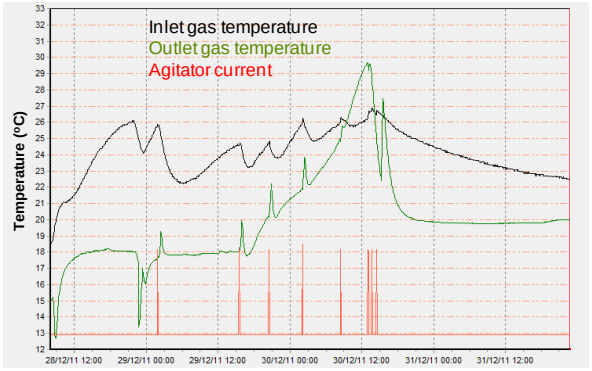


Figure 4-7: Temperature profile of a fast drying batch

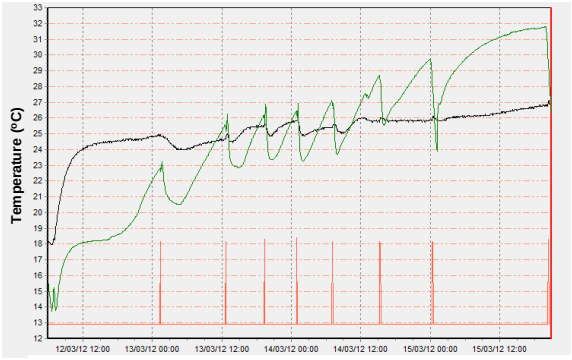


Figure 4-8: Temperature profile of a slow drying batch

4.2.4 Pressure Profile

The pressure drop across the filter is also an important variable in the rate of drying (Richardson et al, 2012). The pressure drop can be estimated from data collected of the inlet and outlet gas pressure. However the pressure drop data shows a high level of variation and does not shows a distinction between fast and slow drying batches. Therefore the pressure data will not be used for subsequent modelling.

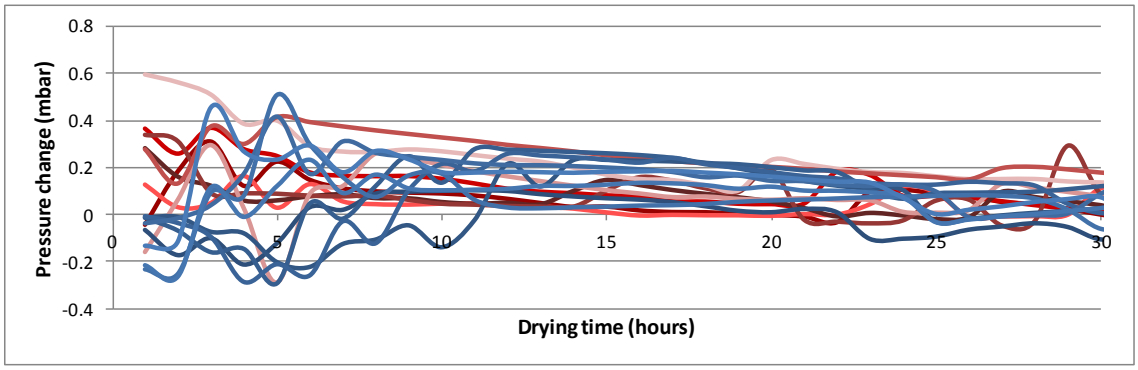


Figure 4-9: Pressure drop across the filter for fast (blue) and slow (red) drying batches, in the first 30 hours of drying

4.2.5 Comparison of Process Variables

A number of measurements have been proposed to be used to predict the final drying time. Some of these measurements, such as readings from the same probe would be expected to be correlated and therefore cannot all be used to build a standard linear model.

The correlation structure between the potential variables is assessed with data of the chosen measurements collected for a set of 16 batches from one dryer. For each batch the data were separated by agitator use and the following information was collected up to the fourth agitation (Figure 4-10):

- Mean N_2 flow rate

- Minimum outlet temperature
- Maximum outlet temperature, excluding the peak after agitation
- Maximum outlet temperature immediately after agitation
- Flow rate of the 1st filter wash

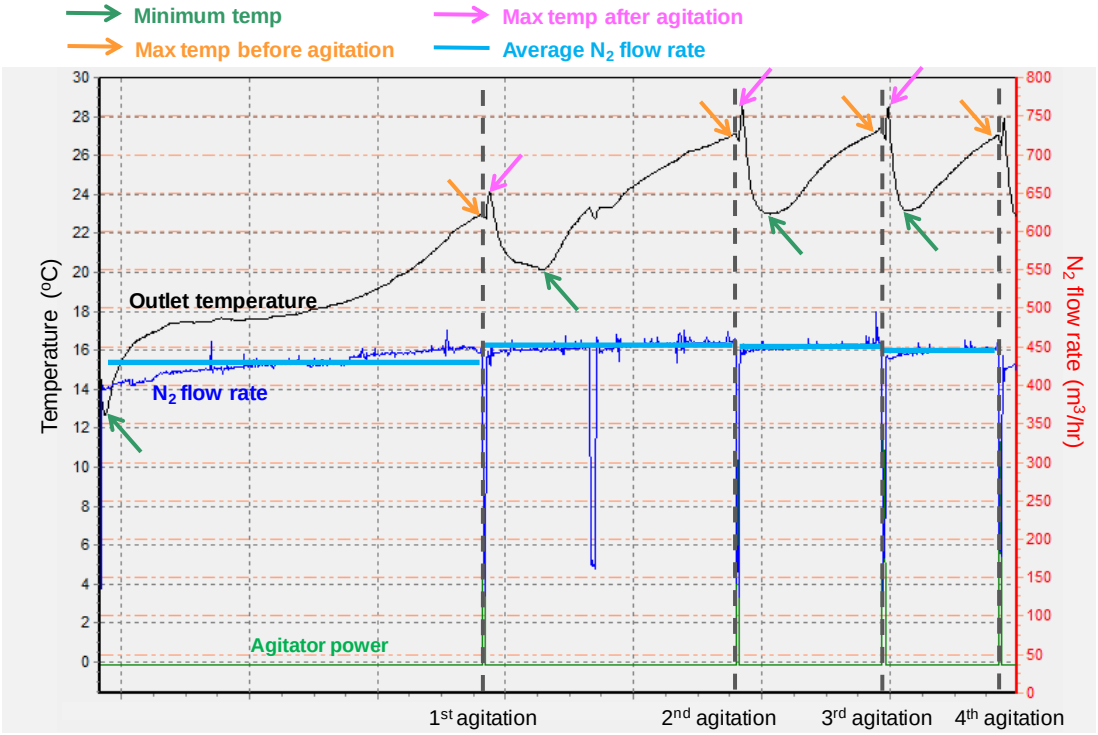


Figure 4-10: Data points to be included in PCA model

A principal component analysis was the undertaken, to assess the correlation structure within the data. The first principal component explained 74% of the total variation in the data. The loadings of the first component show that the N_2 flow rates correlates well with the wash flow rate and the temperature measurements correlate with each other and the drying time, but negatively with the flow rates (Figure 4-11).

Since it was not possible to collect N_2 flow rate measurements for dryer one, the wash flow rate will be used as an input variable to indicate the N_2 flow rate. The maximum temperature variables have similar loadings, suggesting that they all provide the same information. To apply the model early in the drying process, the measurements will be taken from around the first agitation. The temperature peak after the agitator is run is the most straightforward measurement to extract from the temperature profile and hence this measurement will be used as an input to the linear model.

The PCA loadings suggest that there is a negative correlation between the flow rate and temperature measurements. The effect of correlated inputs on linear models can be assessed with variance inflation factors, which will indicate whether the correlation is too high to include both variables (Section 3.1).

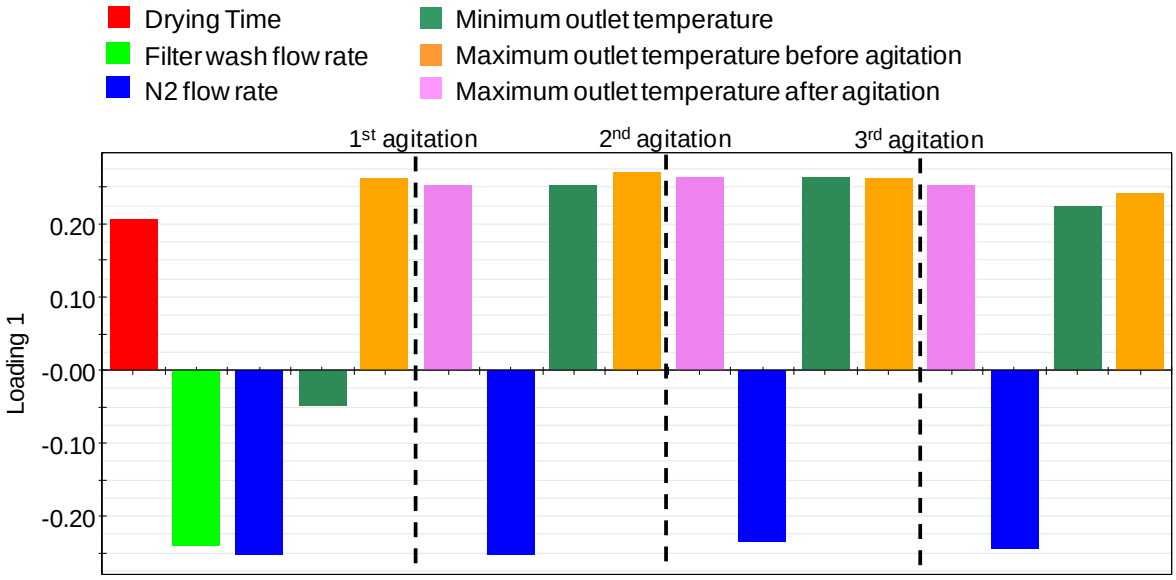


Figure 4-11: Loadings from principal component model

4.2.6 Final LOD Result

A further cause of variation in the drying times is the final water content of the batch. The final loss on drying result can vary from 95.3% to 97.5%, meaning that some batches have drying times that are longer than required. Consequently, if the actual drying times are used as the response variable in predictive modelling, the model will over-estimate the time that is required for the LOD result to reach 95.3%. To prevent the model from over-estimating the drying times, the time that was required for the LOD to reach 95.3% must be inferred for each batch.

The required drying times can be estimated by looking at the profile of the loss on drying results when more than one sample is taken (Figure 4-12). Data collected from batches that passed on the second LOD test shows that the average rate of drying between samples is 0.20%/hr. Therefore the expected time at which the LOD of batch would have been at the specification limit of 95.3% can be calculated as:

$$Adjusted\ drying\ time = Total\ drying\ time - \frac{Pass\ LOD\ result - 95.3}{0.2} \qquad \text{Equation 4-4}$$

Since information is not available on the drying rate of batches that passed on the first LOD, it will be assumed that the rate is similar to the drying rate for batches that passed on the second test.

Batches that passed on the third or fourth LOD tests showed a slower rate of drying up to the sample that passed (Figure 4-13). The dataset was limited because only five batches required more than two samples to be taken. For these batches the average

drying rate was found to be 0.07%/hr. Therefore the adjusted drying time was calculated as:

$$Adjusted\ drying\ time = Total\ drying\ time - \frac{Pass\ LOD\ result - 95.3}{0.07}$$

Equation 4-5

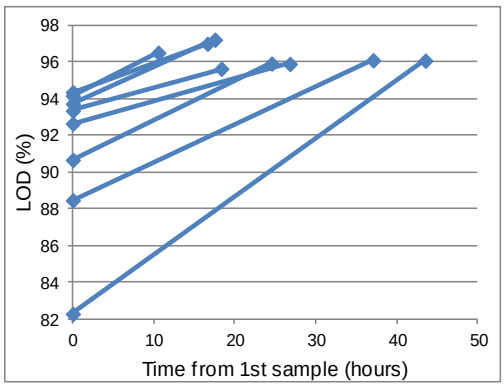


Figure 4-12: Loss on drying results for batches that passed on the second LOD test

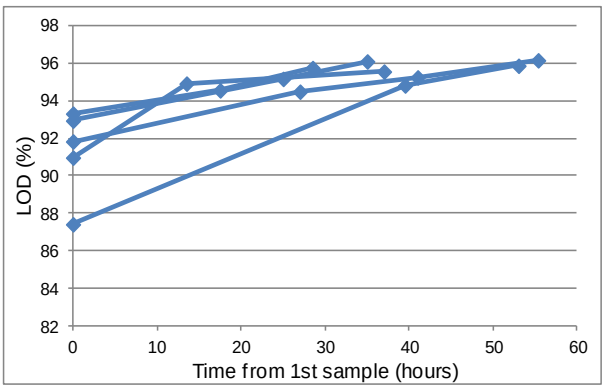


Figure 4-13: Loss of drying results for batches that passes on the third or fourth test

The adjusted drying times were calculated for each of the batches in the dataset, to ensure that the most accurate predictions of the required drying time of a batch could be made. For batches that passed on the first or second LOD test, the adjusted drying time were calculated from Equation 4-4. For batches that passed with three or more LOD tests, the adjusted drying time were calculated from Equation 4-5.

This approach makes a number of assumptions about the rate of drying. Firstly it is assumed the rate of drying at the end of batches that passed on the first sample is the same as the batches that passed on the second sample. In Figure 4-3 two samples are taken before the batch is finished and the flow rate is seen to drop after each sample is taken, suggesting that the rate of drying also slows down. Therefore it is likely that the batches that pass on the first sample may have a faster rate of drying that those that have more than one sample taken. An additional limitation when only two samples are taken is that a linear trend in the LOD results must be assumed between the first and second samples. The data in Figure 4-13 suggests that the trend is close to linear for all but one of the batches.

The rate of drying for fast drying batches could be measured if more samples are taken before the batch has finished drying. Ideally, at least two extra samples would be taken before the end of drying to estimate how the drying rate changes as the LOD reaches 95.3%. Then the drying rate could be compared to other characteristics of the drying process, such as the N2 flow rate, to assess if the drying rate at the end of a batch could be estimated from process data.

4.3 Data set

Following the identification of potential process variables to be predictors for the drying time, an initial dataset was created consisting of drying times, LOD results, heel size, wash flow rate and temperature after the first agitation. From the LOD results, the adjusted drying times were calculated as described in Section 4.2.6. Batches were included from each dryer, over a time span of more than a year that included a range of drying times and heel lengths.

A total of 101 batches were included in the dataset (Table 4-1). The batches were then separated into training and test data sets. The training data (68 batches) was used to build models, and the test data (33 batches) was used to determine how accurate the predictions were for new data. The distribution of drying times in each dataset is shown in Figure 4-14.

For each filter dryer, the majority of batches have drying times of less than 70 hours, with a few batches having longer drying times. The batches to be included in the data sets were chosen to show an even spread over the range of drying times, so that the resulting models were not biased by having more batches with shorter drying times. Fewer batches are available for dryer three because a fault had been recently fixed that was previously resulting in prolonged drying times. Therefore data collected from older batches would not be expected to following the same trends as more recent batches. Only three batches from dryer one were found to have drying times greater than 70 hours, which is seen because on two occasions plant cleans occurred when a number of batches has been run on one heel, but before drying times started to increase above 70 hours. Therefore the heel was not left on the dryer for long enough to observe longer drying times.

Overall there are fewer batches with longer drying times from dryers one and three, so there may not be enough variability in the data for these dryers to develop models that can predict longer drying times. For each batch, the adjusted drying time was calculated using the final loss on drying result, with Equation 4-4 and Equation 4-5.

Dryer	Training batches	Test batches
1	24	13
2	26	13
3	18	7

Table 4-1: Data set sizes for linear modelling

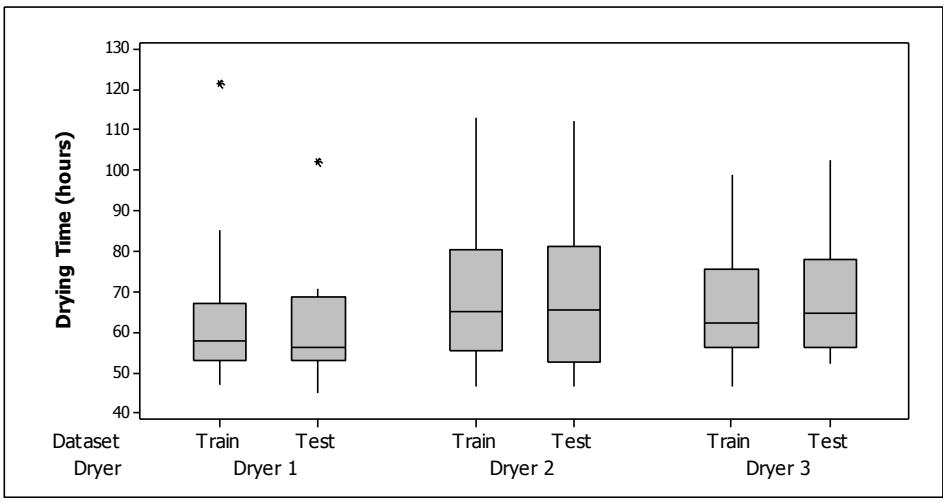


Figure 4-14: Drying times in each data set

4.3.1 Wash flow rate

Comparison of the wash flow rate with the adjusted drying times suggests that the relationship may not be linear (Figure 4-15). Following a clean, the flow rate tends to be high at around 800m^3/hr and drying times are short. The flow rate gradually reduces over consecutive batches, while the drying times remain consistently low (Figure 4-6). However once the flow rate drops to around 500m^3/hr, the drying times start to increase. For the linear models, plots of the residuals will be studied to determine whether a transformation is necessary to fit a linear relationship between the inputs and the drying time.

Figure 4-15 highlights a potential outlier in the data from dryer two exhibiting a very low flow rate. On further investigation it was found that this batch incurred a problem during the purified water washes, which resulted in slow washes but did not affect the overall drying time of the batch. Additionally a batch from dryer three has a very long drying time but not a particularly slow flow rate. During the drying of this batch several pauses in the nitrogen recirculation were noted, which may have reduced the rate of drying. Since the causes of the outliers have been identified, these batches were removed from the data set to prevent them from having a large influence on the fit of the models.

The trend between the flow rate and the drying times is less consistent for dryer three than the other two dryers, with longer drying times being observed for higher flow rates. Consequently there may be difficulties when predicting the drying times for this dryer.

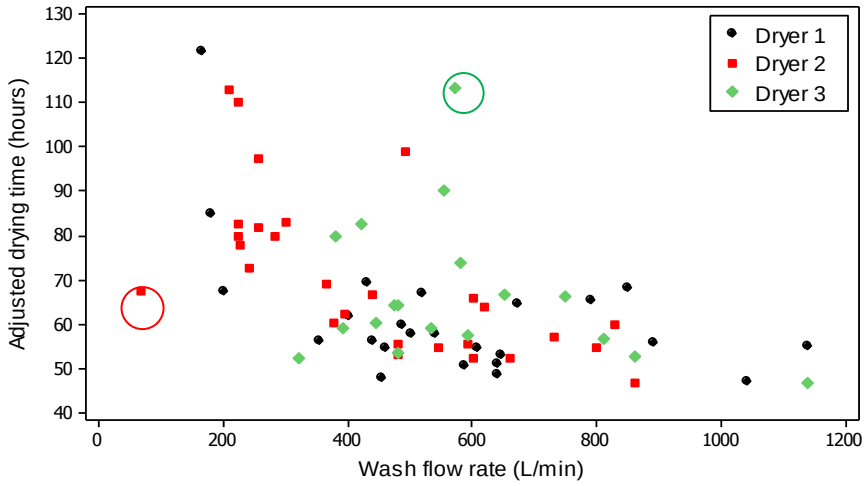


Figure 4-15: Adjusted drying time vs. wash flow rate, for training data, potential outliers circled

4.3.2 Heel age

Although in general drying times are seen to increase as the age of the heel increases, there is variation in the number of batches that are dried before a heel is required to be washed off. Consequently there is not a strong relationship between heel size and drying times (Figure 4-16) and the number of batches on a heel will not be included as a predictor variable in the linear model.

4.3.3 Temperature peak

The relationship between the drying times and the temperature peak after the first agitation appears to be approximately linear for the batches from dryer two (Figure 4-17). However batches from dryer one appear to have short drying times across most of the temperature range, up to 23^{o}C. Conversely batches from dryer three generally show a low temperature after the agitation, but these batches have drying times up to 90 hours. Although the dryers are designed to be identical in set up, changes over time have resulted in differences in dryer behaviour.

The drying times from dryers one and three showed a weak relationship with the temperature after the first agitation (Figure 4-17). On further investigation of the drying profiles, it can be observed that for some batches run in dryer three that have long drying times, the outlet gas temperature does not start to increase until after the agitator has been run for the first time (Figure 4-18). Therefore measuring the temperature around the first agitation does not provide sufficient information to enable a prediction of the drying time.

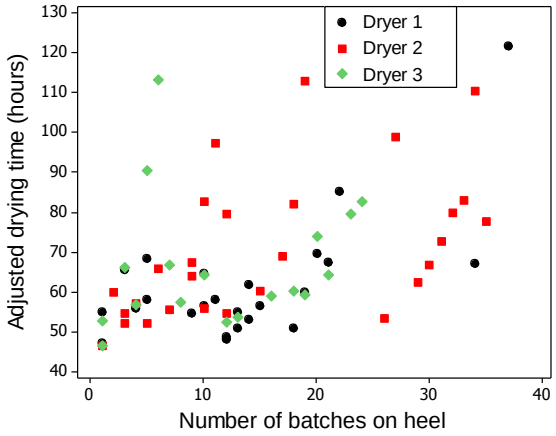


Figure 4-16: Adjusted drying time vs. number of batches since heel wash, for training data

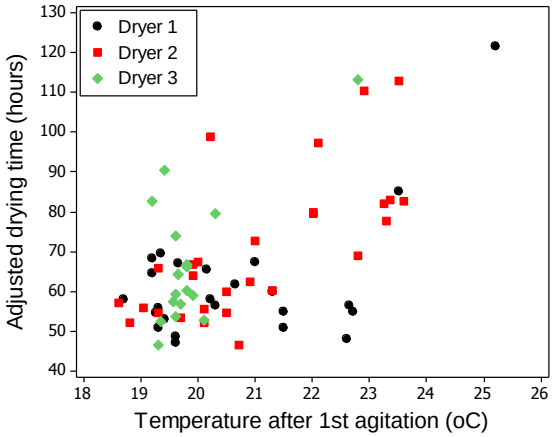


Figure 4-17: Adjusted drying time vs. outlet temperature after first agitation, for training data

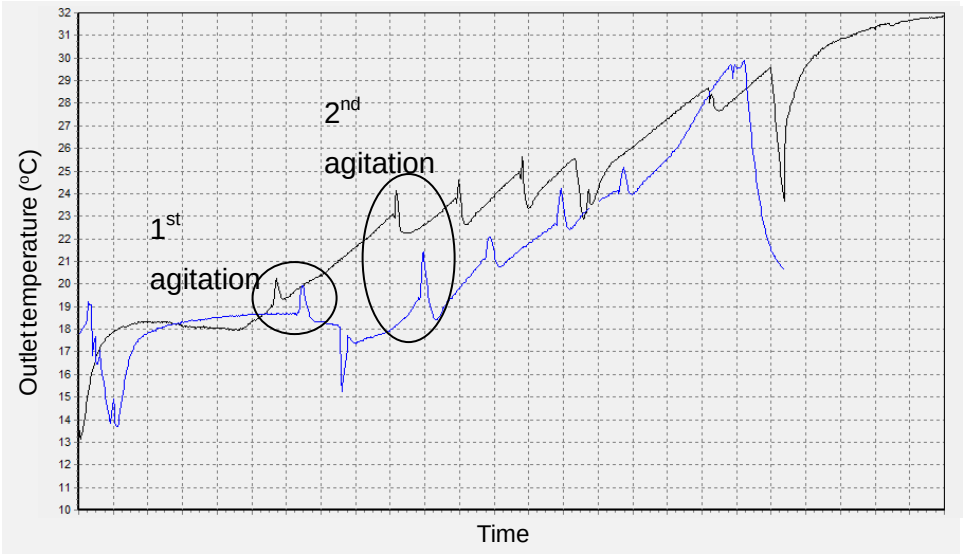


Figure 4-18: Comparison of two batches with fast (blue) and slow (black) drying

For batches dried in dryer three, there is a positive correlation between the drying times and the temperature peak after the second agitation, suggesting that for this dryer a linear model may be appropriate with the inclusion of the additional temperature variable. However for batches that were run on dryer one, the correlation appears much weaker. For this data set, there are only three batches with drying times in excess of 70 hours, so there is not enough variation in the drying times to determine if there is a good underlying correlation. In general, the batches dried on dryer one finished quickly and hence there may not be enough batches with long drying times to make up a suitable data set with which to construct a model.

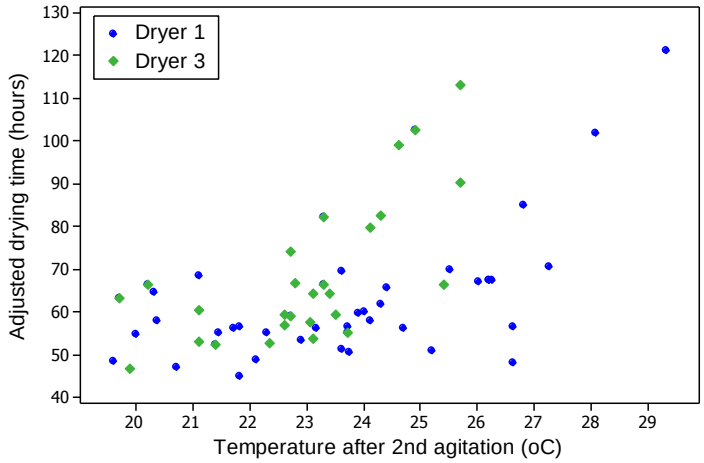


Figure 4-19: Adjusted drying time vs. temperature peak after second agitation, for dryers one and three

Since difference trends have been observed between the three dryers, separate models are required for each dryer, to capture the individual trends and provide optimal predictions of the drying times.

4.4 Linear Models

Linear models were constructed to assess the strength of the linear relationship between the predictor variables and the drying times. The response variable is the adjusted drying time that was calculated as discussed in Section 4.2.6. Models were constructed using Minitab 16.

4.4.1 Model for Dryer Two

The data from dryer two appears to shows a strong relationship between the drying times and the wash flow rate and temperature profile, hence the first linear model was created for dryer two. The initial linear model was created with the flow rate and temperature peak after the first agitation as the predictor variables.

4.4.1.1 Linear model 1

The first linear model shows a good level of fit, with an R^2 of 67%. However the residuals show curvature when plotted against the fitted values (Figure 4-20), with positive residuals observed for the highest and lowest fitted values, suggesting that the assumption of constant variance is not satisfied. It has been noted previously that the relationship between the wash flow rate and the drying times may be non-linear (Section 4.3), so a transformation may be necessary to find a linear model that satisfies the underlying modelling assumptions.

Both predictor variables are shown to have a significant effect on the drying time since the respective p-values are less than 0.05 (Table 4-2). The constant term has a large

p-value, suggesting that the coefficient could equal to zero. However since the drying time would not be expected to be zero when the predictors are zero, the constant term will be kept in the model.

The variance inflation factor (VIF) for the two predictor variables is 2.64 (Table 4-2), which suggests the correlation of the inputs causes a small increase in the variation of the model. However the effect is not large enough to require that one of the predictors should be removed from the model.

Term	Coefficient	SE Coef	T	P	VIF
Constant	-14.4	51.0	-0.282	0.780	
Flow rate	-0.0358	0.0163	-2.19	0.039	2.64
Temperature	4.75	2.127	2.23	0.036	2.64

Table 4-2: Coefficients of linear model 1

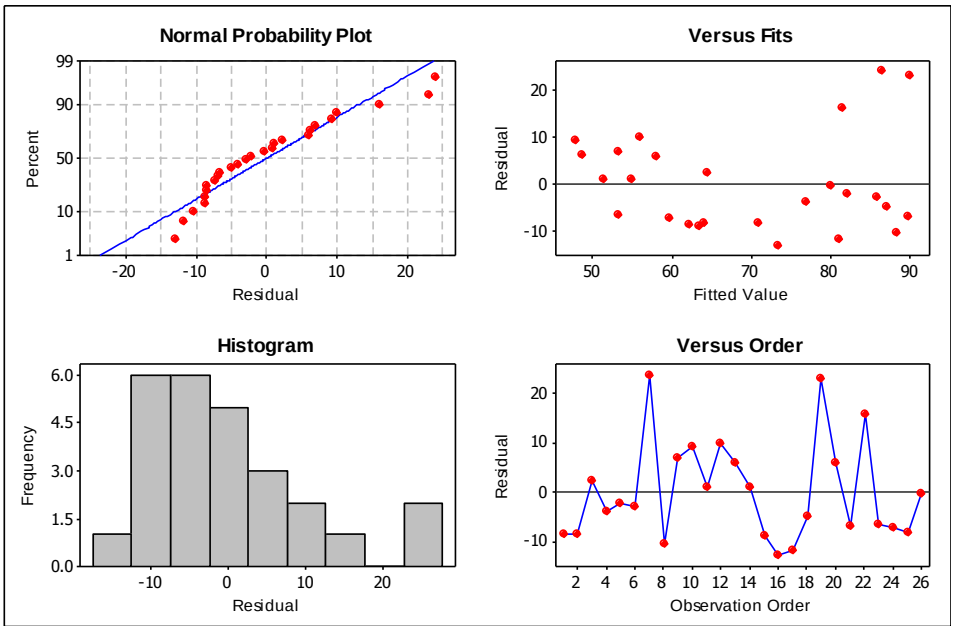


Figure 4-20: Residual plots for linear model 1

4.4.1.2 Linear model 2: Log Transformation of Wash Flow Rate

Figure 4-15 suggests that there could be non-linear relationship between the wash flow rate and the drying time, in particular a log transformation of the wash flow rate data may be appropriate to improve the plots of the residuals. Linear model 2 was constructed with an additional term of the log transformation of the flow rate (Table 4-3). The residuals show an improved fit when plotted against the fitted values (Figure 4-21). The p-value for the temperature term is large (0.64), suggesting that this term may not be required in the model. However when the temperature term is removed the MSE of the test data increases from 74 to 93 and hence this term will remain in the model.

The variance inflation factor (VIF) for the terms 'flow rate' and 'Ln(flow rate)' are 36 and 46 respectively. The high VIF values suggest that there is too high correlation between these two terms and this correlation can increase the variation in the model. However when the 'flow rate' term is removed from the model, curvature is observed in the residuals (Figure 4-22), as in linear model 1. The high correlation of the flow rate variables may be the cause of the high standard error and subsequent low p-value of the temperature term.

Term	Coefficient	SE Coef	T	P	VIF
Constant	444	170	2.61	0.016	
Flow rate	0.107	0.0532	2.01	0.057	36
Ln (flow rate)	-74.2	26.6	-2.79	0.011	45
Temperature	1.09	2.29	0.48	0.638	3.9

Table 4-3: Coefficients for linear model 2

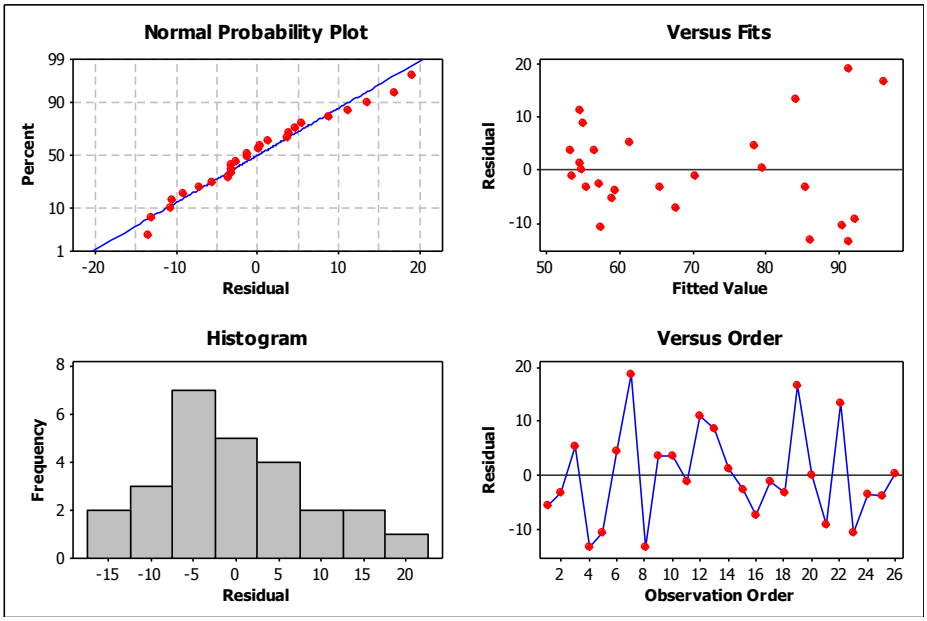


Figure 4-21: Residual plots for linear model 2

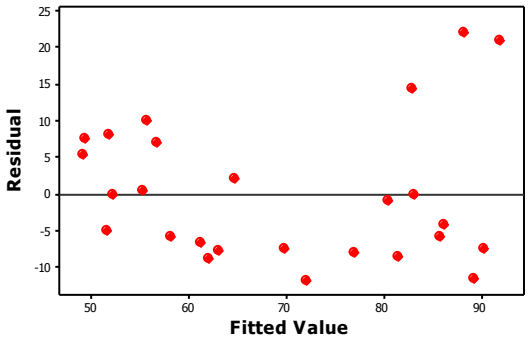


Figure 4-22: Residuals vs. fitted values with 'flow rate' term removed

4.4.1.3 Comparison of linear models 1 & 2

The predictability of models 1 and 2 were assessed by applying them to the training and test data (Figure 4-23 and Figure 4-24). Both of the linear models for dryer two show good predictability when applied to the training and test data sets, with R^2 values of 78% for the test data for both models (Table 4-4). However the high correlation of the input variables in model 2 suggests that model 1 is the preferred model to implement.

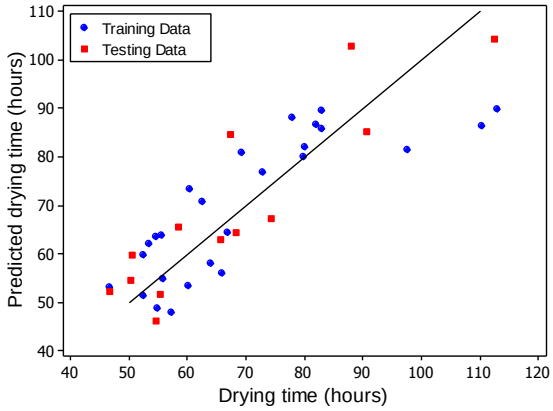


Figure 4-23: Predictions for model 1

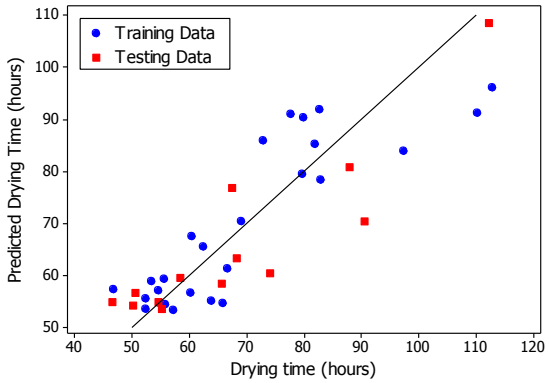


Figure 4-24: Predictions for model 2

	Model 1	Model 2
MSE (train)	99.0	73.2
MSE (test)	74.4	73.5
R^2 (train)	66.8%	75.4%
R^2 (test)	78.3%	78.5%

Table 4-4: Fits and errors of models 1 and 2

4.4.2 Linear model 3 (Dryer Three)

Linear model 3 was created for dryer three, with the temperature peak after the second agitation included alongside the wash flow rate and first temperature peak. An additional two new batches that had recently been manufactured were added to the data set to increase the number of batches with longer drying times, one each were included in the training and test data sets.

For the resulting model, the flow rate has a large p-value and the coefficient for this term is close to zero, suggesting that the flow rate is uninformative as a predictor for the drying time (Table 4-5). It was noted from Figure 4-15 that there is no evidence of a strong correlation between the wash flow rate and the drying time for dryer three.

Removing the flow rate variable produces a model in which all the remaining variables are significant (Table 4-6). The residual plots appear reasonable (Figure 4-25) and the variance inflation factor is 1.1, which suggests that there is no significant correlation between the two temperature measurements.

Term	Coef	SE coef	T	P	VIF
Constant	-248.9	84.9	-2.95	0.01	
Flow rate	0.00	0.01	-0.35	0.73	1.32
Temp peak 1	8.94	3.88	2.31	0.04	1.07
Temp peak 2	6.04	1.68	3.59	0.00	1.32

Table 4-5: Coefficients for linear model for dryer three

The resulting model shows a reasonable level of fit to the training data (Table 4-7). However when the model is applied to the test dataset, none of the variation in the drying times can be predicted (R^2=0). The test dataset is small, with only 8 batches, which may not be large enough to give a reliable indication of the predictability to new data. An alternative method to measure the prediction accuracy for new data is to apply cross-validation (Section 3.1.1.3). Using leave-one-out cross-validation, an R^2 of 35.7% was attained for predictions to the new data (Figure 4-26), which suggests that the model predictions are poor when applied to new data.

The poor predictability of the model for dryer three could be due to the data set lacking enough batches with longer drying times, hence the input variables do not show a strong relationship with the drying time over the range in which they have been collected. Additionally, non-linear models may be more appropriate and are investigated in Section 4.5.2 and Section 4.7.2.

Term	Coef	SE coef	T	P	VIF
Constant	-260.3	74.3	-3.50	0.00	
Temp peak 1	9.12	3.73	2.45	0.03	1.05
Temp peak 2	6.31	1.45	4.34	0.00	1.05

Table 4-6: Coefficients for linear model 3

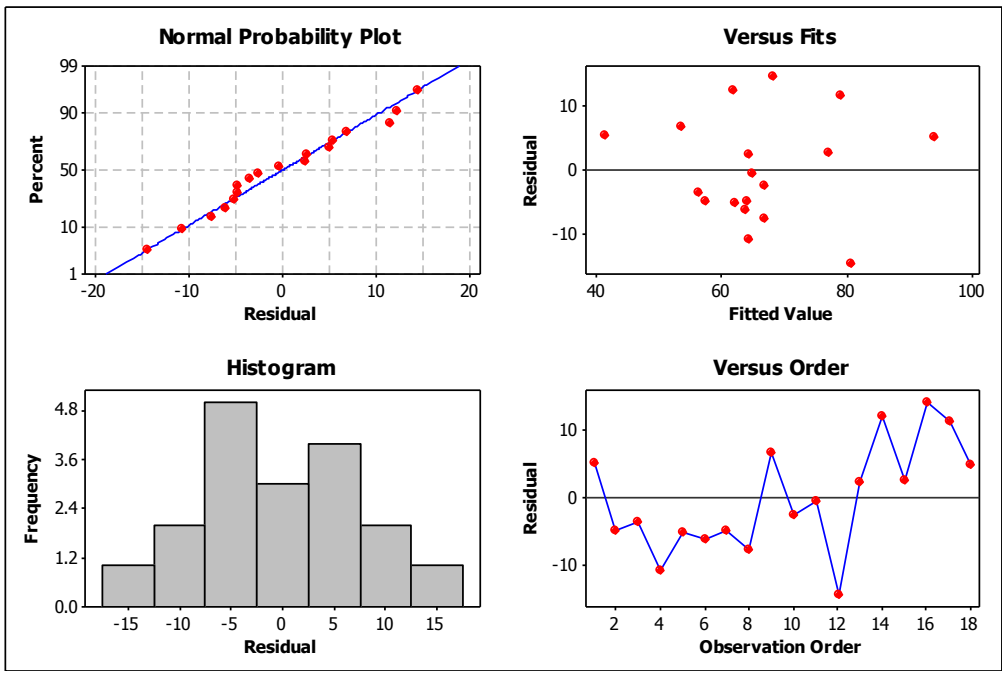


Figure 4-25: Residual plots for linear model 3

	Model 3
MSE (train)	93.9
MSE (test)	183
MSE (CV)	132
R^2(train)	66.9%
R^2(test)	0%
R^2(CV)	35.7%

Table 4-7: Fit of model 3

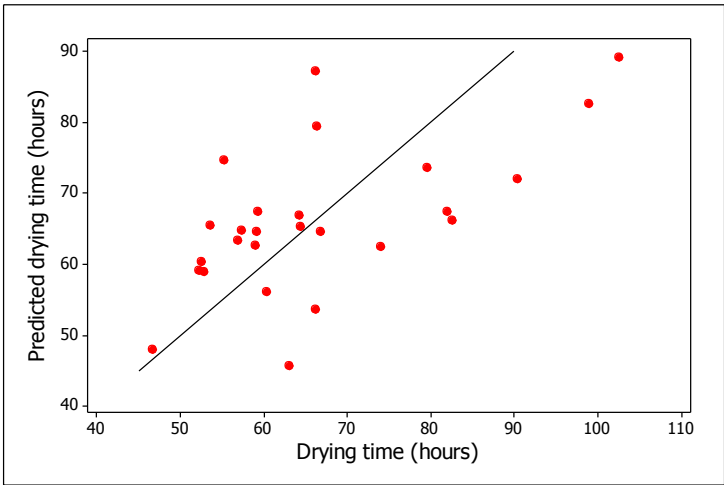


Figure 4-26: Predictions for model 3, with cross-validation

4.4.3 Linear Model 4 (Dryer One)

The process above was repeated with the data from dryer one. A linear model was built with the wash flow rate and the first two temperature peaks as predictor variables. Since there were only three batches with drying times longer than 70 hours there is limited variation in the data, so all of the train and test batches were used to build the model and leave-one-out cross-validation was applied to assess the accuracy of the model when applied to unseen data.

Similar to dryer three, the wash flow rate was not found to be a significant predictor (p-value=0.50) and was removed from the dataset. The resulting model shows strong curvature between the residuals and the fitted value, so a Box-Cox transformation of

the inverse of the drying time was applied to create the final model (Table 4-8 and Figure 4-27).

Linear model 4 shows a poor level of accuracy for predicting the drying time (Table 4-9 and Figure 4-28). The model cannot identify the difference between batches with drying times less than 70 hours, and drying times longer than 70 hours are all under-predicted. As a result, non-linear modelling methods may be more appropriate for the data from this dryer.

Term	Coef	SE coef	T	P	VIF
Constant	-0.0422	0.0051	-8.23	0.000	
Temp peak 1	0.0008	0.0004	2.10	0.043	2.2
Temp peak 2	0.0004	0.0002	1.68	0.103	2.2

Table 4-8: Coefficients of linear model 4

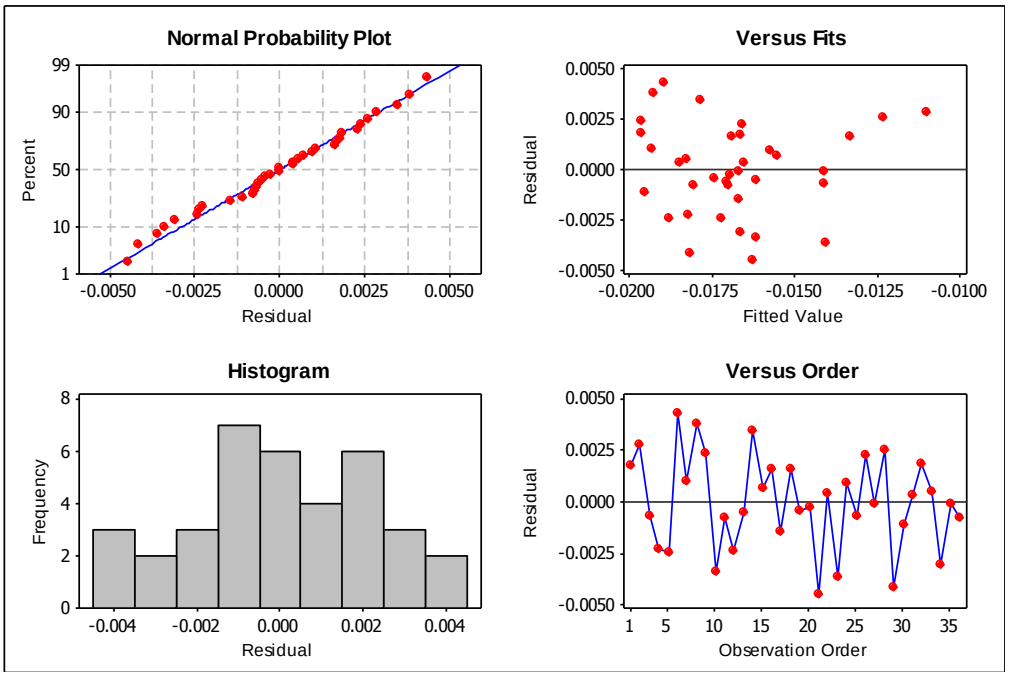


Figure 4-27: Residuals for linear model 4

	Model 4
MSE (train)	93.2
MSE (CV)	123
R^2 (train)	45.1%
R^2 (CV)	44.3%
VIF	2.2

Table 4-9: Fit of linear model 4

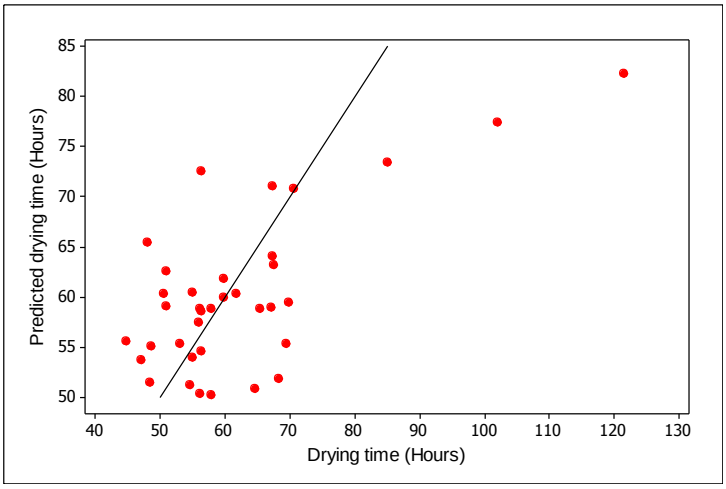


Figure 4-28: Predictions for model 4, with cross-validation

4.5 Neural Networks

The residual plots from the linear models suggest that the residuals do not satisfy the assumptions of independence and identically normally distributed. As an alternative modelling method, neural networks were investigated to represent the potential non-linear relationship between the predictors and the drying time. Stacked neural networks (Section 3.4.2) were considered since they have been found to be more robust than standard neural networks for predicting the responses of unseen data. Neural networks were initially applied to the data from dryer two, since the strongest relationship between the predictors and response were observed for this dryer. Neural networks were constructed using the Matlab (2008b) Neural Networks toolbox, but adapted to generate stacked neural networks (Figure 4-29).

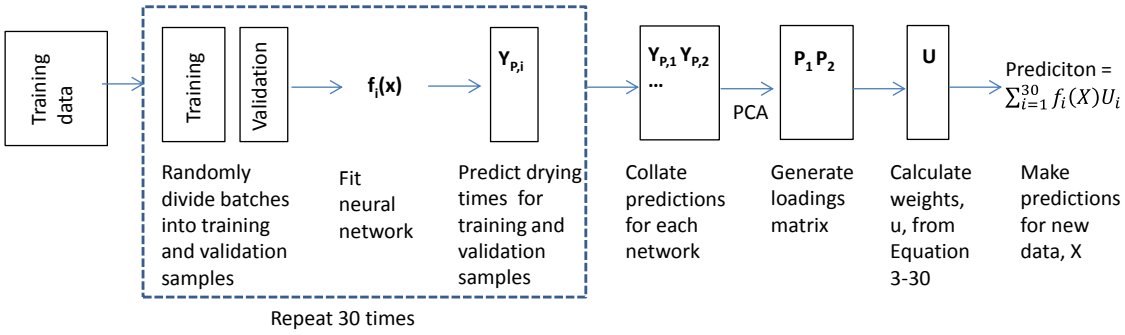


Figure 4-29: Process for fitting stacked neural networks

4.5.1 Dryer Two

The dataset for dryer two comprised the same 39 batches as for the linear modelling analysis, of which 13 batches formed the tests dataset. The 26 remaining batches were split into 17 training and nine validation batches and were selected randomly for each

individual network that was created. Therefore over a stacked network each batch would be expected to appear in both the training and validation datasets.

4.5.1.1 Network Structure

In the development of a stacked neural network, a number of features must be determined: the number of principal components (PCs) to retain, the number of hidden nodes and the number of individual networks to be stacked.

Firstly the number of PCs was determined. A stacked neural network was created comprising 30 individual networks, recommend by Zhang *et al* (1997), and one hidden node, since there are only two inputs. One PC was found to explain 90% of the total variation and hence one PC will be retained in all subsequent models in this section. To determine the optimal number of individual networks to stack, stacked networks were created with 10, 20, 30 and 40 individual networks. 20 repeated networks were fitted of each structure to show the variation between identically set up networks (Figure 4-30). A stacked network comprising 30 individual networks was found to minimise the error when applied to the test data; increasing the size to 40 networks did not reduce the error any further. Finally networks were assessed with one, two or three hidden nodes (Figure 4-31). One node was found to be sufficient to consistently minimise the error for the test data, adding more nodes may over fit the model to the training data and result in poorer predictions when applied to the testing data.

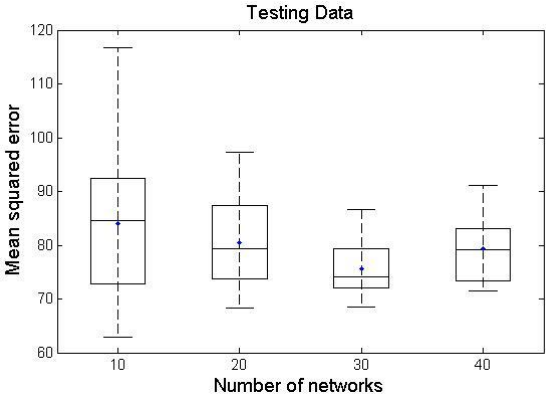


Figure 4-30: MSE vs number of individual networks

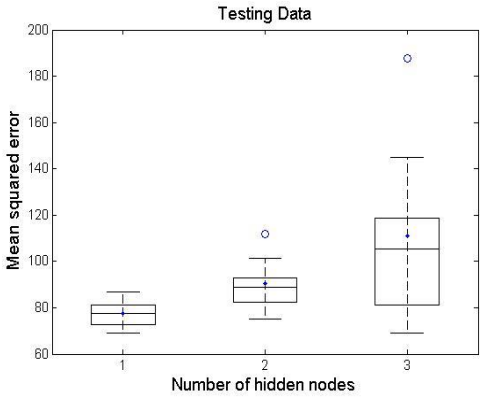


Figure 4-31: MSE vs number of hidden nodes

4.5.1.2 Stacked Neural Network

The final stacked neural network model was created from 30 individual networks, with one retained PC, one hidden node and two predictor variables: wash flow rate and first temperature peak. The model shows a high level of accuracy when applied to both the training and test datasets (Table 4-10 and Figure 4-32). The level of fit is similar to the linear model in Section 4.4.1, with similar R^2 and MSE values, a full comparison of all of the modelling methods is given in Section 4.8.

Dataset	MSE	R^2
Training	72.4	75.7%
Testing	68.6	79.9%

Table 4-10: Fit of stacked neural network, dryer 2

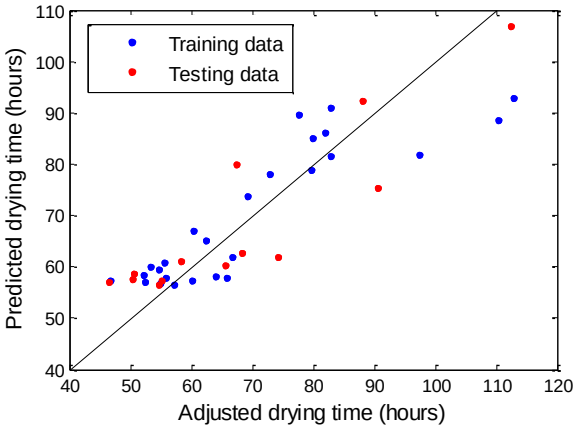


Figure 4-32: Predictions for stacked neural network, dryer two

4.5.1.3 Individual Neural Network

For comparison, an individual neural network was created from the dataset for dryer two, with one hidden node. The results were similar to the stacked network (Table 4-11, Figure 4-33), although a slightly higher level of accuracy was achieved for the test dataset. This is because there is a large amount of variation between individual networks that are created from the same data, so an individual network can be selected that gives the highest possible fit to the test dataset. Conversely the stacked network will contain some networks that have a lower level of fit, so the overall fit is lower. However the stacked network is more likely to represent the range of trends that are seen in the data and hence will be more applicable to future data that is collected.

Dataset	MSE	R^2
Training	70.0	76.2%
Validation	87.5	71.8%
Testing	56.3	83.5%

Table 4-11: Fit of individual neural network, dryer two

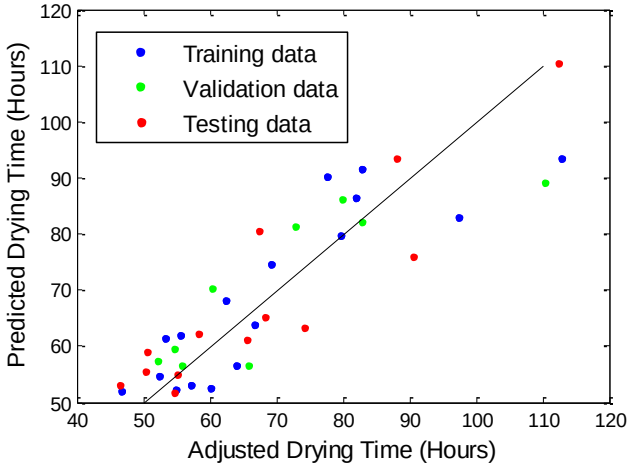


Figure 4-33: Predictions for individual neural network, dryer two

4.5.2 Dryer Three

The above process was repeated for the data from dryer three. The 19 training batches were divided randomly for each network into 12 training and seven validation batches, and the same eight test batches were used as for the linear modelling analysis. 30

stacked networks were used and two PCs were retained that explained 87% of the total variation.

The results from the linear model found that wash flow rate was not a significant predictor for the drying time, so the neural network results were compared with and without the wash flow rate as an input variable (Figure 4-34). The error for the test data was minimised when the wash flow rate was included, suggesting that a non-linear relationship is present that cannot be represented with a linear model. One hidden node was required to minimise the error.

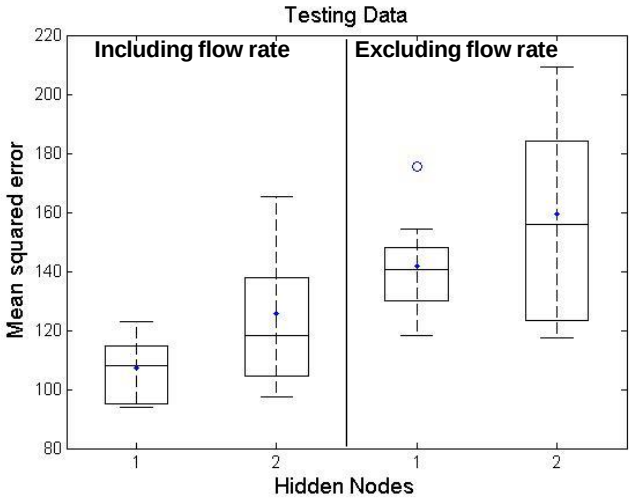


Figure 4-34: MSE vs. number of hidden nodes, with and without flow rate variable

The final stacked neural network for dryer three was fitted with one hidden node and three input variables: wash flow rate, 1st and 2nd temperature peaks. A good fit is observed between the actual and predicted drying times (Table 4-12 and Figure 4-35); in particular the neural network produced a much better fit than the linear model when applied to the test dataset (Section 4.4.2).

Dataset	MSE	R^2
Training	43.8	76.6%
Testing	94.8	60.3%

Table 4-12: Fit of stacked neural network, dryer 3

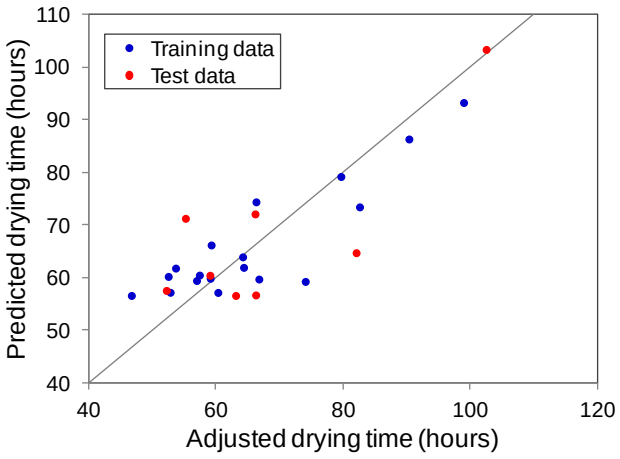


Figure 4-35: Predictions for stacked neural network, dryer 3

4.5.3 Dryer One

The data for dryer one only includes only three batches with long drying times, consequently a linear model was difficult to develop. For the neural network model, the training data was split into 15 training and eight validation batches, with the same 13 test batches that were used in the linear modelling. For a stacked network comprising 30 individual networks, two PCs were retained that explained 91% of the variation.

Similar to dryer three, the flow rate variable was found not to be a significant predictor in the linear model. A comparison of stacked neural networks with and without the flow rate as an input suggested that the flow rate was not required to optimise the predictions for the test dataset (Figure 4-36). In addition, the error appears to be minimised when three hidden nodes were included in the network. This result is unexpected because the neural networks for the other two dryers only required one hidden node, but since the data set is limited for this dryer, the network may be more difficult to fit and hence more hidden nodes are required.

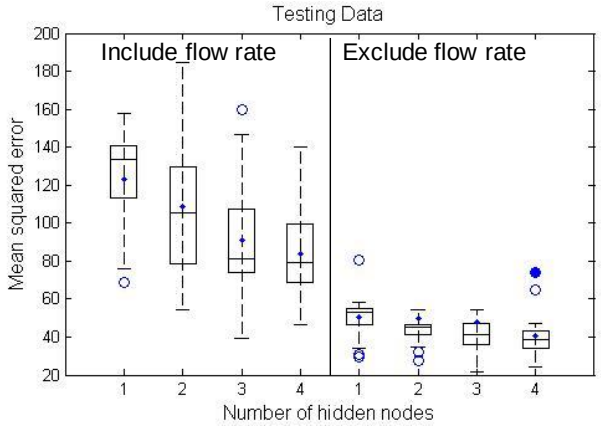


Figure 4-36: MSE vs. number of hidden nodes

The predictions for the neural network model show an improved fit compared to the linear model (Table 4-13 and Figure 4-37), particularly for the batches with drying times longer than 70 hours. However both models are unable to model the batches with drying times less than 70 hours. This may be because there is not enough variation in the input variables for batches with shorter drying times, or other unknown factors may be influencing the drying times.

Dataset	MSE	R^2
Training	50.1	78.6%
Testing	31.9	83.8%

Table 4-13: Fit of stacked neural network, dryer one

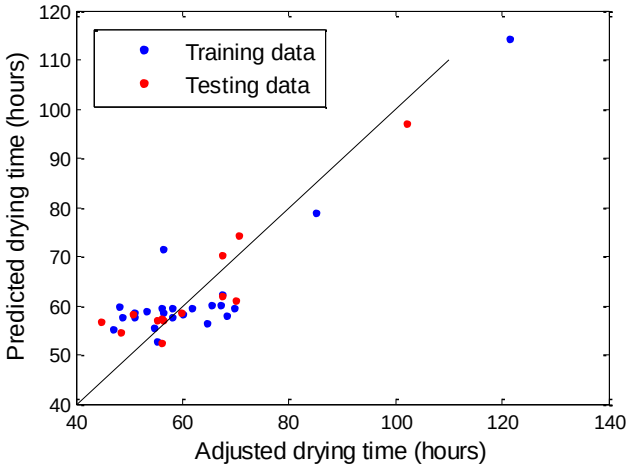


Figure 4-37: Predictions for stacked neural network, dryer one

4.6 Multi-way Partial Least Squares

An assumption of the linear and neural network modelling methods is that the input variables are not correlated and hence individual variables were selected from the measurements that are collected throughout the running of each batch. However, reducing the information to two or three variables creates the risk of losing a large amount of information that is contained within the rest of the data. Alternative approaches that allow input data to be used from multiple time points throughout the batch include multi-way partial least squares (MPLS) (Section 3.2.1) and case based reasoning (Section 3.2.2). These two techniques are applied to the drying data and the results are compared to those from the linear and neural network models, to determine the most appropriate approach for predicting the drying times.

The key variables measured during the drying process are the N_2 flow rate, the inlet gas temperature and the outlet gas temperature. Dryer one does not have an in-range flow meter, and hence this dryer is excluded from the rest of the analysis. Data is collected every ten seconds, resulting in over 10,000 data points for each batch. To extract the data in a manageable amount, the data was extracted as an hourly average for each variable, during the running of each batch. Since the profiles of the variables do not change considerably from hour to hour, no significant information is expected to be lost by averaging. The data was then aligned for each batch (Section 4.6.1.1) and used to develop a MPLS model. The MPLS process is summarised in Figure 4-38.

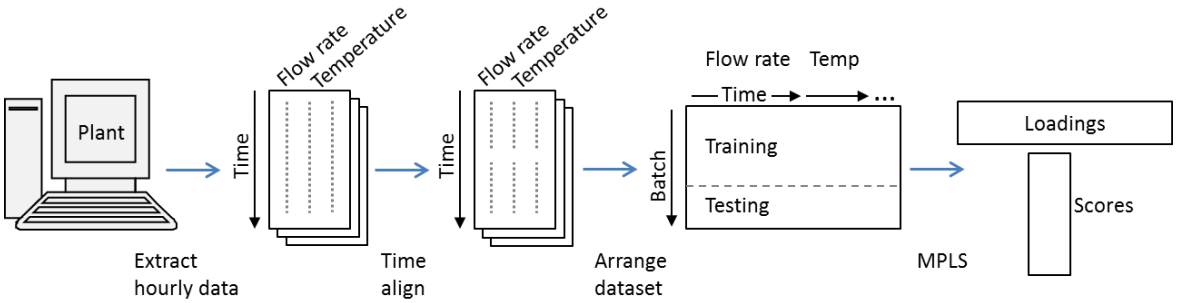


Figure 4-38: Process for MPLS

4.6.1 Dryer Two

For the 39 batches in the dataset for dryer two, the N_2 flow rate and the inlet and outlet gas temperature measurements were collected as hourly averages. Figure 4-39 shows a typical batch trend of the measured variables. The profiles generated from taking the hourly averages are similar to those produced directly by the plant data system (Figure 4-7 and Figure 4-8). Changes are observed around the times that the agitator is used to mix the powder on top of the filter. The flow rate drops when the agitator is operated because the gas supply is stopped and a lower flow rate may result when the gas supply is restarted as the product can become more compressed. Data for each batch was collected up to the fifth agitation, which is scheduled to be 48 hours from the start of the drying process.

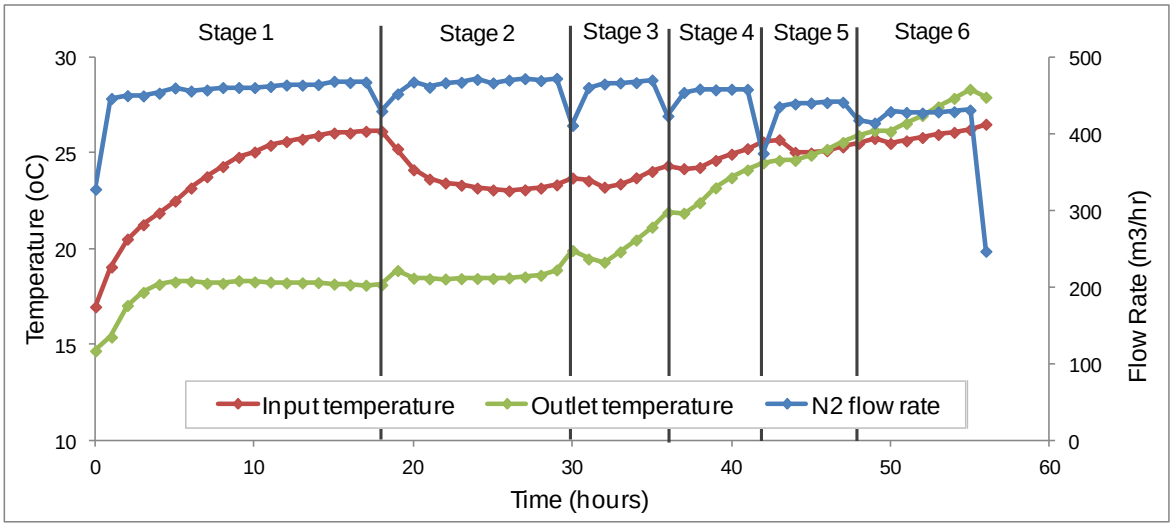


Figure 4-39: Typical batch trend

4.6.1.1 Data Alignment

The agitator is started manually by the operator, so the exact time that it is switched on varies between batches. Since the operation of the agitator can cause changes in the trends of the temperature and N_2 flow rate, it is necessary to align the data from each

batch around the operation of the agitator, to ensure that the same information is being compared across the batches.

To align the data from each batch, the time points were divided in stages, with each stage starting after the agitator was run (Table 4-14). The timings of agitator runs were found in plant's control system. Figure 4-40 to Figure 4-42 show an example of batches before and after alignment, respectively. The data from each stage was lined up so that gaps were left when a stage was run for less time than usual (batch 2) and data removed when a stage ran for more time than usual (batch 3). The alignment process was the same for each input variable. Data alignment was done manually in Microsoft Excel 2007.

Stage	Time points (Hours)
1	0-17
2	18-29
3	20-35
4	36-41
5	42-47

Table 4-14: Time points in each stage of the drying process

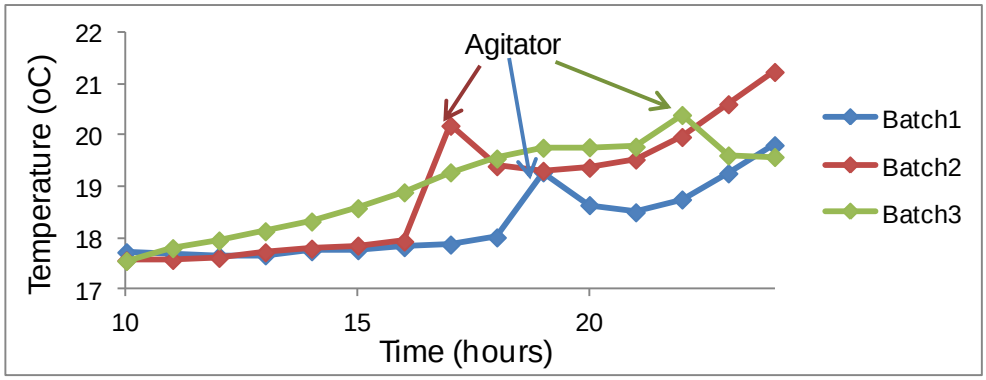


Figure 4-40: Example of data before alignment

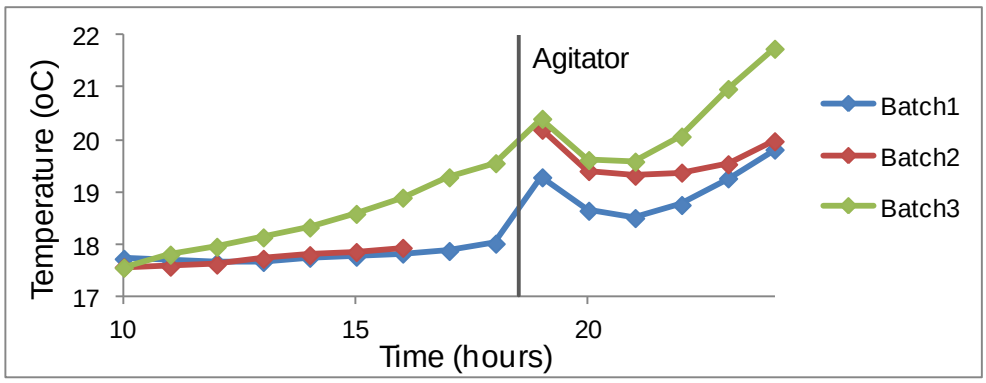
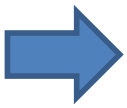


Figure 4-41: Example of data after alignment

Time	Temperature		
	Bx1	Bx2	Bx3
10	17.90	17.75	17.75
11	17.80	17.75	17.90
12	17.80	17.80	18.00
13	17.83	17.85	18.20
14	17.85	17.88	18.40
15	17.88	17.91	18.55
16	18.00	18.04	18.75
17	18.20	20.30	19.00
18	18.50	19.35	19.45
19	19.40	19.40	19.70
20	18.80	19.50	19.75
21	18.70	19.60	19.80
22	18.85	19.80	20.50
23	19.15	20.30	19.45
24	19.50	21.00	19.45
...	...	...	...



Time	Temperature		
	Bx1	Bx2	Bx3
10	17.90	17.75	17.75
11	17.80	17.75	17.90
12	17.80	17.80	18.00
13	17.83	17.85	18.20
14	17.85	17.88	18.40
15	17.88	17.91	18.55
16	18.00	18.04	18.75
17	18.20		19.00
18	18.50		19.45
19	19.40	20.30	20.50
20	18.80	19.35	19.45
21	18.70	19.40	19.45
22	18.85	19.50	19.80
23	19.15	19.60	20.60
24	19.50	19.80	21.30
...	...	...	...

Figure 4-42: Example of time aligning data (coloured by stage)

4.6.1.2 MPLS Model Results

An MPLS model was created using the three input variables each collected over 48 time points and aligned so that the start of each stage is lined up (Figure 4-42). The data set was transposed to have a row for each batch and column for each time point and variable. Modelling was performed using the Simca P+ 12.0 software. The data was scaled so that each variable had a mean of zero and unit variance. Retaining two latent variables was found minimise the MSE and maximise the R2 when the MPLS model was applied to the test data (Figure 4-43). By retaining two latent variables, 71.5% of the variation in the test data can be predicted (Table 4-15). Overall the model showed good predictions of the drying time (Figure 4-44).

The scores and loading plots suggest that the flow rate is an important predictor throughout the drying process; with batches with a high flow rate having a shorter drying time (Figure 4-45 and Figure 4-46). The outlet gas temperature is also important, but large loadings are only observed after the first 12 hours of drying. Higher outlet gas temperatures are associated with longer drying times, since higher temperatures suggest that less water is being driven off the batch. The inlet gas temperature appears to have a weaker relationship with the drying time since larger loadings are only seen in the second latent variable (Figure 4-47).

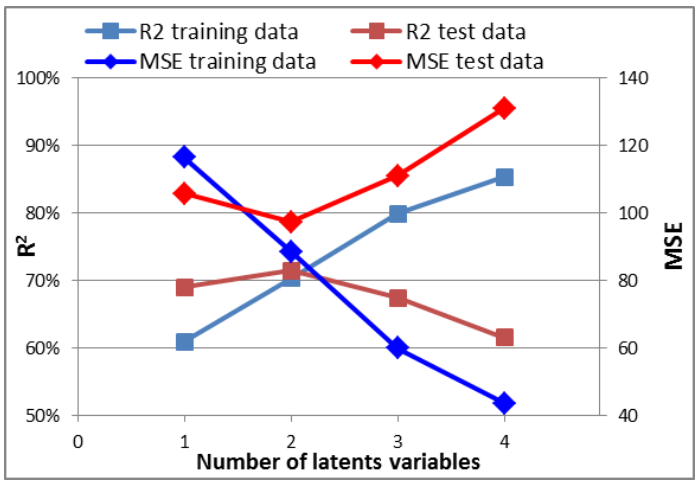


Figure 4-43: Fit to training and testing data as more latent variables are added

Dataset	MSE	R^2
Training	88.5	70.3%
Testing	97.2	71.5%

Table 4-15: Fit of MPLS model for dryer two

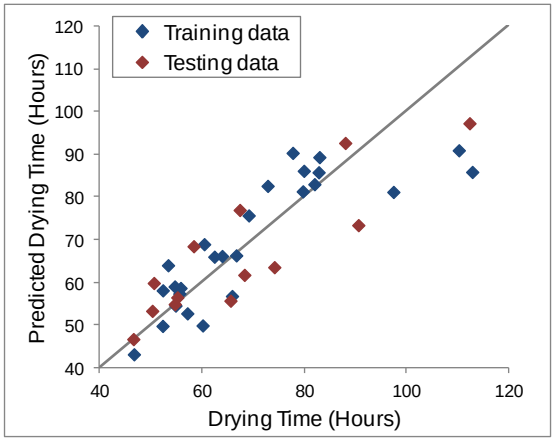


Figure 4-44: Predictions for training and testing data

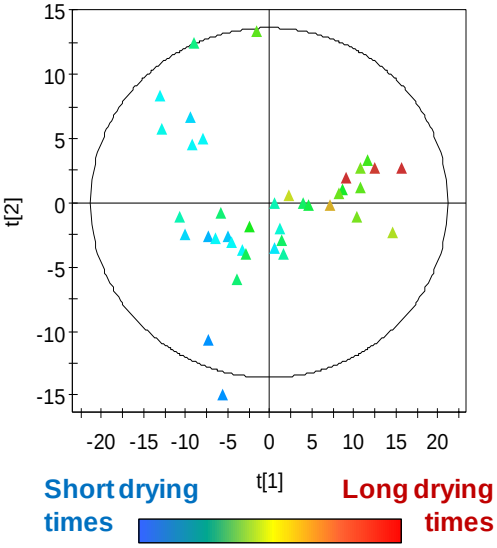


Figure 4-45: Score of first two latent variables

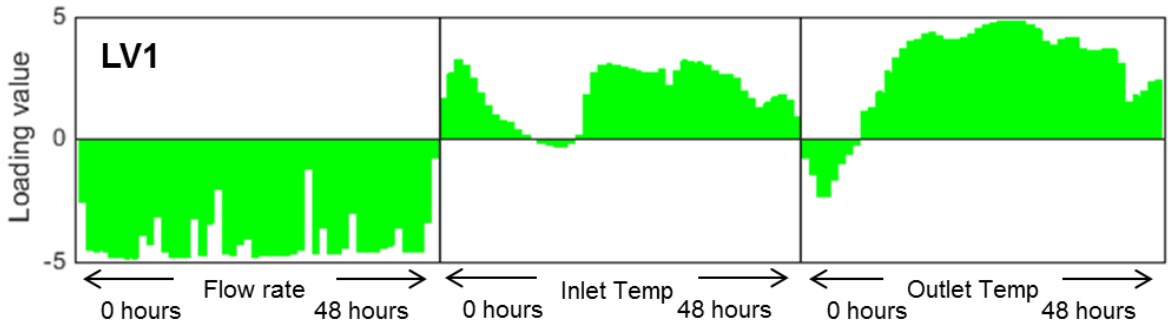


Figure 4-46: Loadings of first latent variable

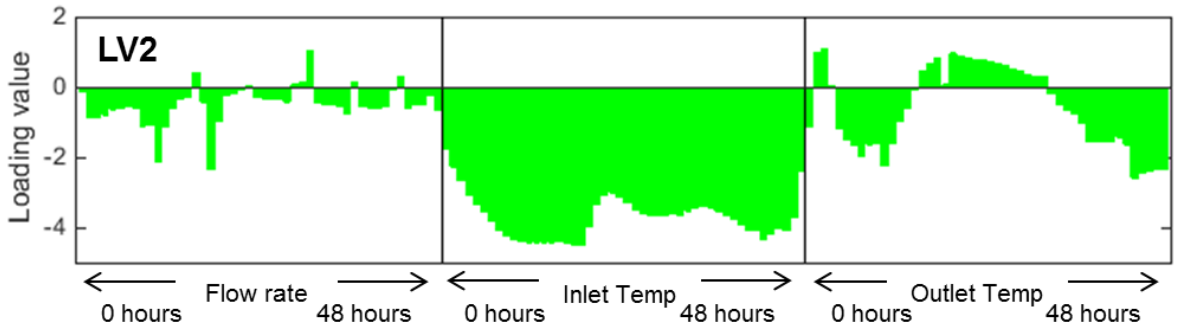


Figure 4-47: Loadings of second latent variable

The presence of non-linear trends can be assessed by comparing the input and output scores of the PLS model (Figure 4-48 and Figure 4-49). In both figures a linear trend is seen between the input and output scores. Along with the normal probability plot of the residuals (Figure 4-50), there is no evidence to suggest that non-linear PLS methods are required.

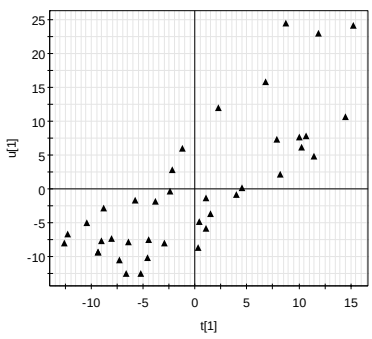


Figure 4-48: Output vs. input scores, first latent variable

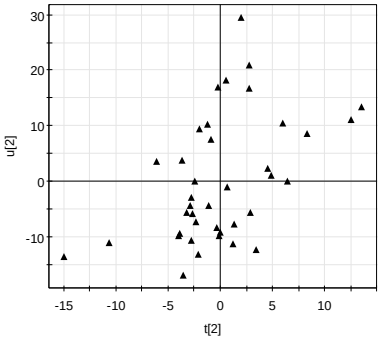


Figure 4-49: Output vs. input scores, second latent variable

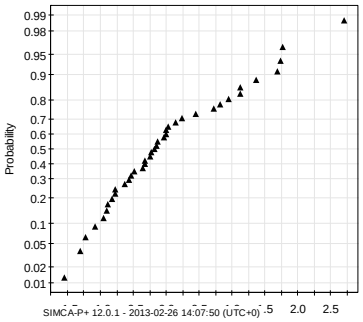


Figure 4-50: Normal probability plot of residuals

4.6.1.3 How many stages of data are needed?

Ideally, the drying time should be predicted early in the batch so that information is available as soon as possible to plan up and down stream processes. The MPLS model created above requires data to be collected for the first 48 hours of the drying process. To determine how much data is required to produce good predictions, the above process was repeated to create five new MPLS models, with data included up to each of the first five agitator runs. The prediction accuracy for new data (Q^2) increases as more data is included in the model (Figure 4-51). Q^2 was calculated directly by Simca P+, using five-fold cross validation. Only a small difference is seen from the first to the fifth agitation, with Q^2 increasing from 64% to 67%, showing that good predictions can be achieved early in the drying process.

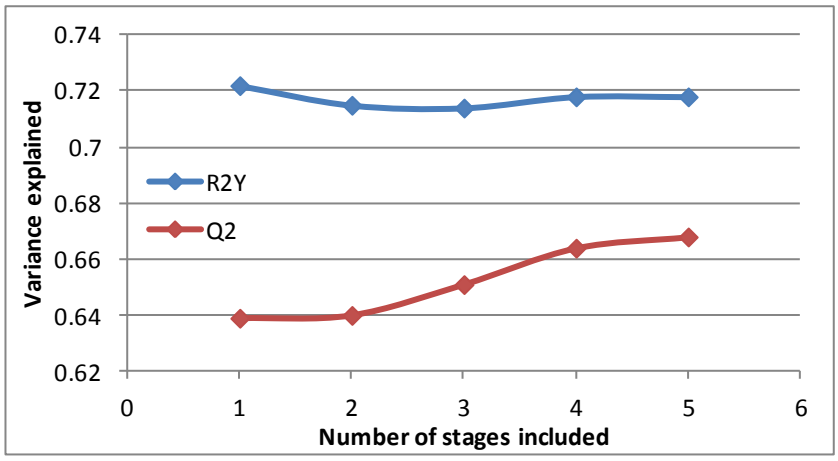
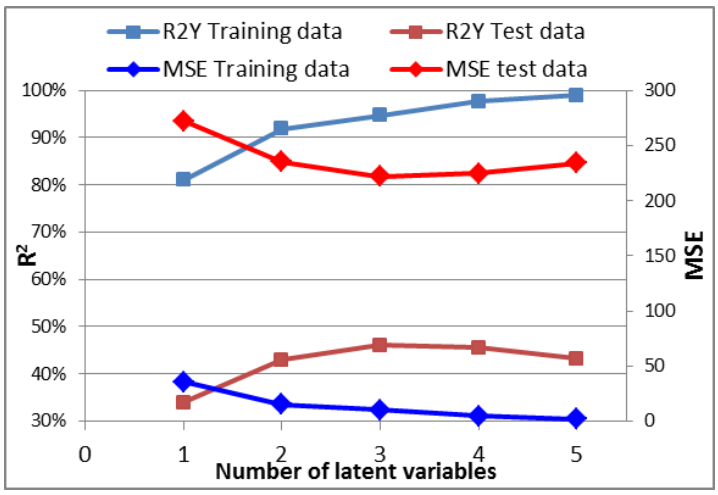


Figure 4-51: Variance explained for models with different amounts of input data

4.6.2 Dryer Three

The MPLS analysis was repeated for batches dried on dryer three. When the models were applied to the test data, it was found that selecting three latent variables minimised the MSE and maximised the R^2 value of the test data (Figure 4-52).

A good level of accuracy was found for the training data, with an R^2 of 95%, however when applied to the test dataset the value of R^2 fell to 46% (Table 4-16, Figure 4-53). The loadings plots suggest that the most important input is the outlet gas temperature around the middle of the drying duration (Figure 4-55), which shows a positive correlation with the drying time (Figure 4-54). This finding agrees with the results from the linear model for dryer three (Section 4.4.2), where the flow rate was not found to be a significant predictor and the temperature after the second agitation, at 30 hours, was required to produce good predictions. Across the three latent variables high loadings are seen for all of the process variables and time points, so the MPLS models uses information from across the drying process (Figure 4-55 to Figure 4-57).



Dataset	MSE	R^2
Training	10	95.0%
Testing	221	46.1%

Table 4-16: Prediction accuracy for MPLS

Figure 4-52: Fit to training and testing data as more latent variables are added

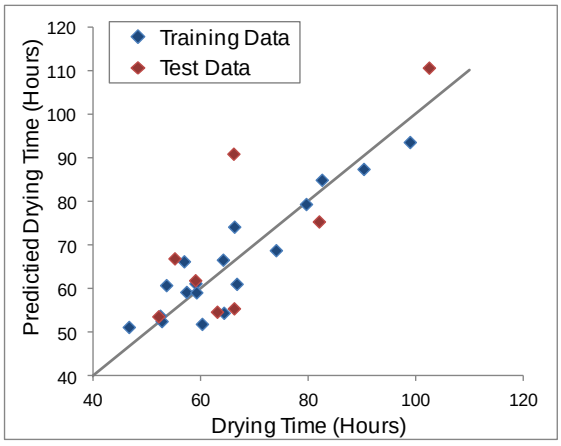


Figure 4-53: Predictions for training and testing data

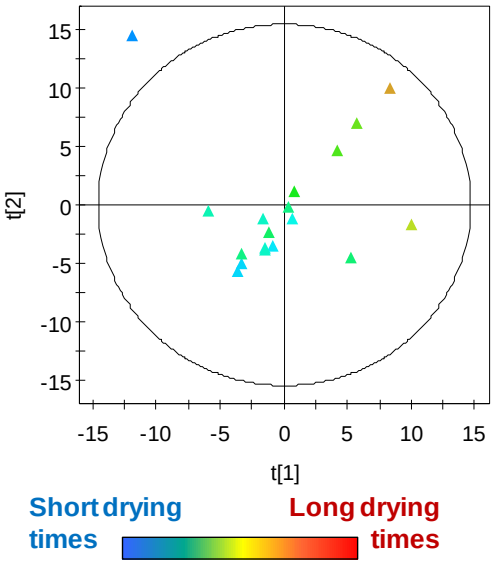


Figure 4-54: Scores of first two latent variables

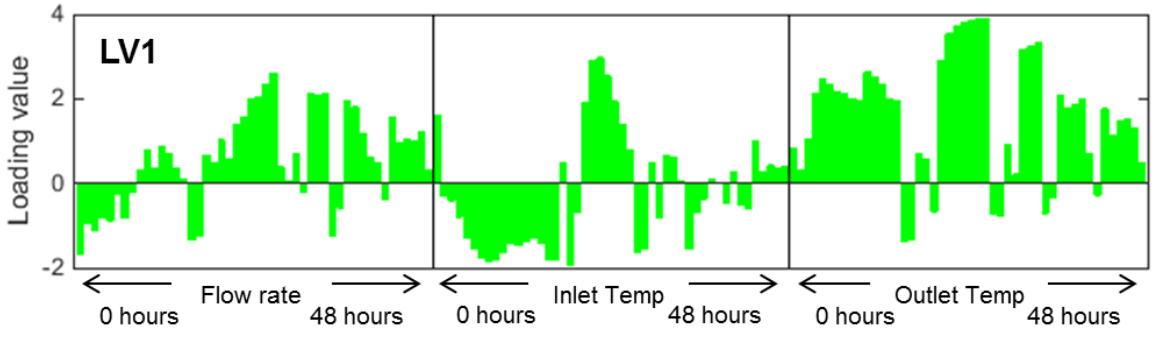


Figure 4-55: Loadings for first latent variable

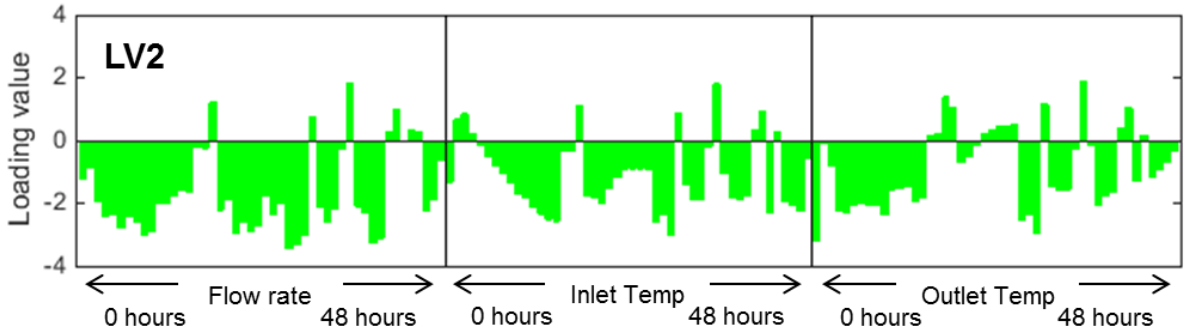


Figure 4-56: Loadings for second latent variable

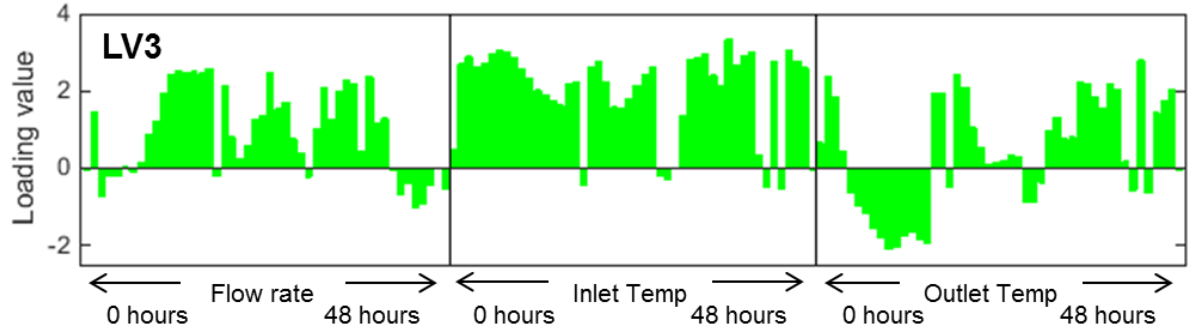


Figure 4-57: Loadings for third latent variable

4.7 Case Based Reasoning

The final method to be assessed for predicting drying times was case based reasoning (CBR). This method makes no assumptions about the shape or trends within the data, with predictions being made by comparing batch profiles and identifying those that are most similar. The methodology was explained in Section 3.2.2. CBR analysis was conducted using Matlab 2008b.

Profiles of a new and historical batch are compared by calculating the difference in variables measured at the same time points and then these are summed across the whole batch (Figure 4-58). Due to the time alignment (Section 4.6.1.1), the variables from the drying process included some missing data, so the difference is only calculated when neither batch has missing data. The difference, D, is calculated from n time points of the new and historical flow rate data (f_{new}, f_{hist}) and m time points of the inlet and outlet temperature data (t_in and t_out:), for which there is no missing data.

$$D = \frac{r}{n}\sum_{i=1}^{n}\left(f_{new,i} - f_{hist,i}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(t\_in_{new,i} - t\_in_{hist,i}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(t\_out_{new,i} - t\_out_{hist,i}\right)^2$$

Equation 4-6

Since the flow rate and temperature data have different ranges, the difference between flow rate measurements will be greater than the differences between temperature measurements. Therefore a weighting, r, is included to ensure the influence of each

type of data is equal. The difference between data points can be calculated as either the squared or absolute difference.

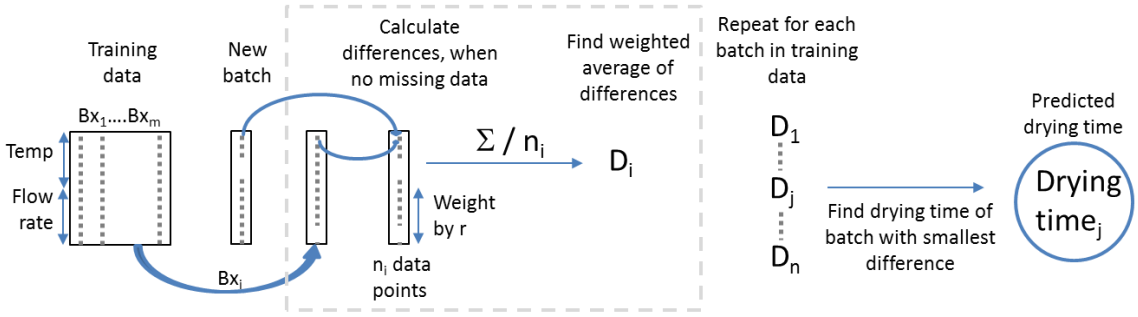


Figure 4-58: Process for CBR

The value of the weighting, r, was calculated by applying CBR to the training data for dryer two and comparing the mean squared error for various values of r (Figure 4-59). The MSE was minimised when r was set to 0.05, and D was calculated from the absolute difference of the data points. Using the squared difference will accentuate large differences between batch profiles that may only be caused by a small number of noisy data points. In contrast the absolute difference places greater emphasis on differences that are occur over a larger number of data points and hence the absolute difference is more likely to represent true differences between batches.

4.7.1 Dryer Two

Case based reasoning was applied to the dataset from dryer two that had been used for the MPLS analysis. Since the N_2 gas flow is stopped when the agitator is run, the flow rate measurement at this time is low and does not contain information about the rate of drying, so the flow rate data was removed for these time points.

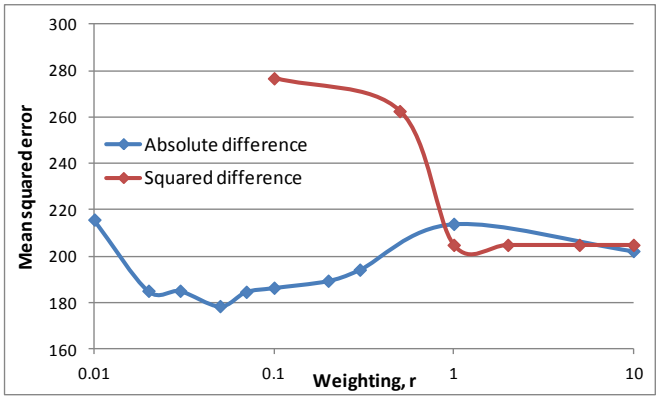


Figure 4-59: MSE for values of r

The ability of the method to predict drying times for new batches was tested using the training data as the historical dataset and predictions were made for the batches in the

test dataset. CBR showed low prediction accuracy when applied to both datasets (Table 4-17).

Dataset	MSE	R^2
Training	170	42.8%
Testing	211	38.1%

Table 4-17: Prediction accuracy for CBR

The low prediction accuracy may be attributed to the amount of noise within the dataset. The drying process, in particular the N_2 flow rate, is influenced by processes that occur around the filter dryer. For example, liquid from the initial water washes is collected in a receiver that is emptied during drying. When the receiver is emptied the nitrogen flow is briefly stopped, causing a dip in the flow rate measurement. The data that had previously been aligned for each stage (Figure 4-60), was also filtered by removing dips in the flow rate that occurred whilst the dryer was running (Figure 4-61).

In addition, the loadings from the MPLS model suggested that the inlet gas temperature has a smaller effect on the drying time compared to the flow rate and the outlet gas temperature. Consequently the inlet temperature data may add noise to CBR method by creating unnecessary variables that do not have a strong relationship with the drying time. Further CBR models were created, to determine whether filtering the data by excluding outliers or removing the inlet temperature would improve the predictions.

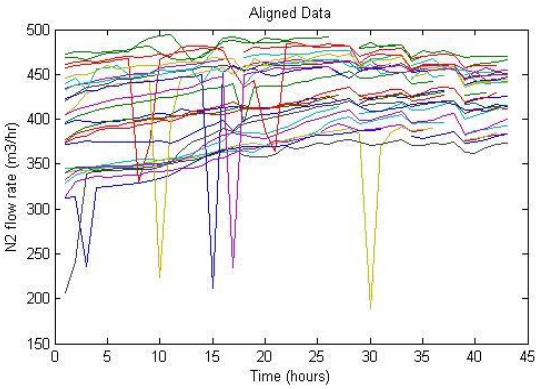


Figure 4-60: Aligned flow rate data

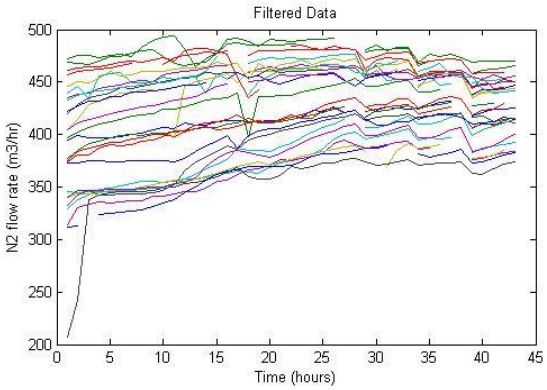


Figure 4-61: Aligned and filtered flow rate data

The highest level of fit to the test data set was found when the filtered dataset was used and the inlet temperature data was removed (Table 4-18). However only half of the variation in the drying times can be predicted (Figure 4-62), so the level of accuracy is too low to give a good indication of the expected drying time.

		MSE train	MSE test	R2 train	R2 test
Unfiltered	Include T_in	170	211	43%	38%
	Exclude T_in	176	219	41%	36%
Filtered	Include T_in	176	222	41%	35%
	Exclude T_in	179	170	40%	50%

Table 4-18: Level of fit for various datasets

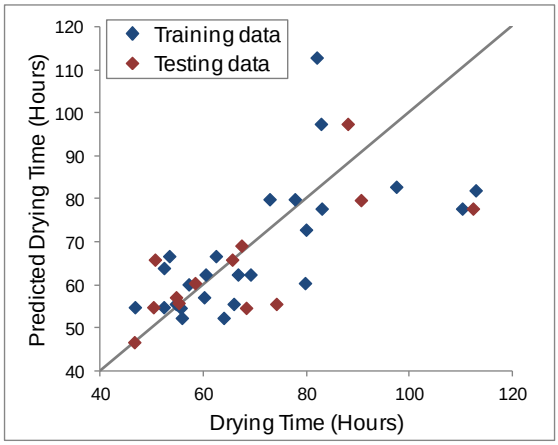


Figure 4-62: Predictions for CBR

The results suggest that even though the data was collected as an hourly average and outliers were removed, the CBR method is unable to handle the amount of noise that remained in the data. Large differences caused by a small number of unusual data points can have a strong influence on the calculated difference between batches, causing similar batch profiles with some unusual data points to be labelled as very different.

4.7.2 Dryer Three

The CBR analysis was repeated for dryer three. The optimal value for the weighting, r, between the N_2 flow rate and the temperature data was found to be 0.0005, suggesting that the temperature measurements have a much stronger relationship with the drying time than the flow rate. However, removing the flow rate data completely resulted in an increase in the prediction error.

The results show that good predictions can be made for the test dataset, although the fit is poor for the training data (Table 4-19), suggesting that CBR will not always produce good predictions for this dryer. In particular the predictions for training batches with drying times less than 70 hours are very poor for the training data (Figure 4-63).

Dataset	MSE	R^2
Training	97.1	48.2%
Testing	57.5	75.9%

Table 4-19: Prediction accuracy for CBR

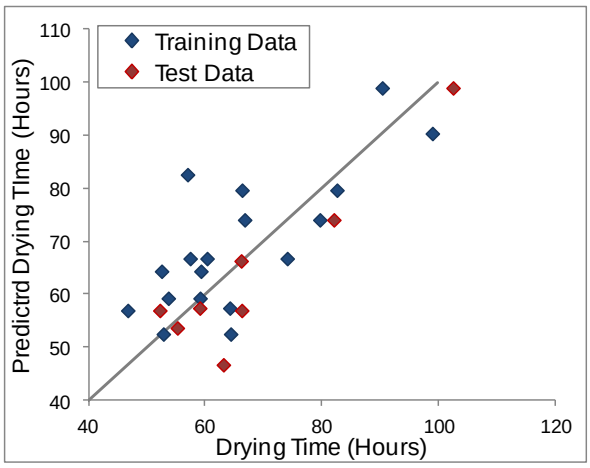


Figure 4-63: Predictions for CBR

4.8 Comparison of Methods

The modelling methods that were applied to the drying data can be grouped into categories: linear or non-linear methods and techniques that use individual variables or techniques that use the whole batch profile.

Comparison of all four modelling methods when applied to the data from dryer two shows that the highest prediction accuracy for the test dataset was found for the linear model and the stacked neural network (Table 4-20). The results show that reducing the data set to two variables, wash flow rate and first temperature peak, does not reduce the amount of information that is captured about the drying time. Although non-linear trends were observed data, they are not so strong that a linear model is not able to provide accurate predictions.

The MPLS model was also found to produce predictions with a good level of accuracy. With this method, the data collected at multiple time points was combined linearly and reduced to two latent variables that summarise the gas flow rate and the drying temperature, providing inputs that are similar to those of the linear model.

The case based reasoning approach resulted in the lowest level of accuracy for the predictions of the drying time. This method may be the least able to handle the noise in the data and therefore cannot identify the differences in batch profiles that indicate the drying time. Although MPLS uses the same input data as CBR, MPLS reduces noise by linearly combining the data from all of the time points into a latent variable, in effect averaging out the noise.

The linear model has the advantage of being the most straight forward method to use on the manufacturing plant and can be implemented with an Excel Spreadsheet into which the values for the two input variables are added.

Model	MSE (train)	MSE (test)	R^2 (train)	R^2 (test)
Linear	99	74	67%	78%
Stacked NN	72	69	76%	80%
MPLS	89	97	70%	71%
CBR	179	170	40%	50%

Table 4-20: Comparison of prediction accuracy for dryer two

For dryer three, the more accurate predictions for the test dataset were produced by the stacked neural network and the CBR methods (Table 4-21), suggesting that significant non-linear trends occur in for the data for this dryer. However for CBR, the accuracy for the training data is low. Of the three dryers, the dataset for dryer three is the smallest with 27 batches, so data from additional batches may be required to fully test the accuracy of the CBR method for making predictions. For the linear model, the R^2 from the test data was zero. Cross validation was applied to determine whether a limited data set was limiting the performance of the linear model. The R^2 improved slightly to 36%, but overall the fit of the model to new data is very poor.

	MSE (train)	MSE (test)	R^2(train)	R^2(test)
Linear	94	(CV) 132	67%	(CV) 36%
Stacked NN	44	95	77%	60%
MPLS	10	221	95%	46%
CBR	97	58	48%	76%

Table 4-21: Comparison of prediction accuracy for dryer three, CV implies cross validation of all batches

The results for the data from dryer one found that the neural network model was able to produce more accurate predictions than the linear model (Table 4-22). For this dryer, non-linear trends in the data were such that the linear model could not represent the relationship between the inputs and the drying time.

Model	MSE (train)	MSE (test)	R^2(train)	R^2(test)
Linear	93	(CV) 123	45%	(CV) 44%
Stacked NN	50	32	79%	84%

Table 4-22: Comparison of prediction accuracy for dryer one, CV implies cross validation of all batches

4.9 Conclusions and Further Work

The analysis has shown that the data collected throughout the drying process can be summarised in terms of two or three variables without losing information required to predict the drying times. The relationship between the summary variables and the drying time has shown the presence of non-linear trends, although a suitable linear model was developed for dryer two.

Differences are seen in the trends for the three dryers. The linear models for dryers one and three required the temperature after the second agitation to be included as a predictor, but this variable was not required for dryer two. The neural network for dryer one suggested that the wash flow rate did not have a strong relationship with the drying time for this dryer. Although all of the dryers were set up in the same way, different parts have been replaced over time which could cause differences in the way that each dryer is run and in the measurements that are taken. In addition, the length of pipe work that the nitrogen gas travels through is different for each dryer, which may cause difference to the gas flow rate and temperature. Further investigation could be undertaken by the process technical team to understand the differences between the dryers.

The poorest predictions were observed for batches made on dryer one. This dataset was limited with only three batches having drying times in excess of 70 hours. A future project could be to collect further data from batches with longer drying time with the intention of improving the models for this dryer. At present the availability of the required data is limited because few batches are observed with long drying times.

The major benefit from this piece of work has been the implantation of a linear prediction model on dryer two that allows the plant managers to estimate the expected drying time early in the drying process. The linear model is straightforward to implement and is manually run in an Excel Spreadsheet. The input data is obtained from graphs of the outlet temperature and filter receiver level produced by the plant's data recording system. The prediction model has been particularly useful to determine when to schedule a clean of the filter.

Since a good level of fit was found for the linear model on dryer two, it may be possible to construct better linear models for dryers one and three if more batches with longer drying times could be added to the datasets. Linear models are the preferred models to use since they are the most simple to implement, requiring just an Excel spreadsheet. Therefore a recommendation was made to improve the linear models by collecting more data from batches with longer drying times. If a suitable level of fit can be found

for dryers one and three, then these models can also be implemented. Data collection and analysis could be supported through AstraZeneca's Six Sigma programme.

An original aim of this study was to predict the drying time with enough accuracy so that the loss on drying samples are not required to determine the drying end point. The outcome was in part limited by the available data that could be used as inputs to the model. A further measurement that has the potential to provide useful information is the humidity of the gas leaving the filter. At the start of drying, the humidity would be expected to be high as a large volume of water is removed from the product. As the product dries the humidity of the outlet gas should reduce, until the level is low enough to indicate that the product is suitably dry. An extension to this piece of work for AstraZeneca would be to install a humidity probe to the outlet gas line of each dryer, and determine whether the humidity provides a more accurate indication of the drying end point than the current method of measuring the temperature profile.

Following the drying process, the material is transferred to the milling facility to reduce the particle size of the solid powder. In Chapter 5, the PLS and stacked neural networks techniques presented in this chapter, along with principal component analysis, are applied to data from the mill to understand the variation in the particle size distribution of the milled product.

5 Multivariate Analysis of Particle Size Distribution Data

5.1 Introduction

When an active pharmaceutical ingredient (API) is manufactured it typically takes the form of a solid powder. The particle size distribution (PSD) of the API will affect a number of important properties of the product during the formulation stage and will also impact on the quality of the final product (Iacocca *et al*, 2010). For example, smaller particles may produce greater uniformity of the drug content within tablets, whereas a small number of large particles in one tablet can result in a high dose that is out of specification (Orr, 1982). Additionally a greater overall surface area of the API particles will lead to more rapid dissolution and a higher release rate of the drug into the body (Simões *et al*, 1996).

A number of methods are available to measure the particle size distribution of a powder, based on either the physical properties of the powder or the interaction of the particles with light (Section 5.2). Measurement methods differ in terms of complexity and accuracy, and range from sieving to microscopy and laser diffraction.

The data that is produced from a particle size distribution measurement will consist of a number of size measurements and the frequency density of the particles at each size. Although information can be summarised in terms of the mean or percentiles of the distribution, these methods do not use all the information that is available. The PSDs of several samples can be compared graphically by overlaying the curves of each distribution and identifying differences. However to compare a number of samples effectively, the information from each distribution can be summarised using multivariate analysis techniques.

This chapter consists of two sections. The first section introduces and compares methods for measuring the particle size distribution of a pharmaceutical powder (Section 5.2). The benefits and limitations of each method are considered, along with the assumptions underpinning the interpretation of the data. The second section focuses on an analysis of PSD data from an API product that is manufactured by AstraZeneca. Multivariate analysis methods were applied to understand the variation in the data and to assess the differences between batches produced on different processing plants (Section 5.3). Prediction models are then created to determine how the variation in the production and milling processes may affect the PSD of the final product (Section 5.4).

5.2 Measurement of Particle Size Distribution

The particle size distribution (PSD) of a solid powder describes the range of sizes of the individual particles that make up the powder. The size can be quantified in terms of the length, surface area, volume or mass of the particles. The shape of the particles is also important and can be described by factors such as the surface area to volume ratio or the geometric shape of a particle (Allen, 1997). This section reviews a number of the PSD measurement methods that are commonly used for the characterisation of pharmaceutical powders.

The most straightforward methods, such as sieving (Section 5.2.3), separate out the particles by their size and then the quantity of particles in each size range is measured. Alternatively microscopy (Section 5.2.4) is used to create an image of the particles, which is then used to estimate the particle sizes. Microscopy is the only technique that allows the particles to be individually viewed and measured. Laser diffraction (Section 5.2.5) uses the interaction between the particles and laser light to determine the PSD; this method is complex but produces rapid and highly repeatable results. The choice of method depends on the size of the particles to be measured, the level of agglomeration in the powder and the cost, accuracy and speed of the measurement technique.

The PSD measurement that is produced will depend on the response of the instrument that is used and hence measurements of the same powder will vary between instruments. To produce a reliable PSD measurement, several measurement techniques could be compared to find a good representation of the particle size (Shekunov *et al*, 2007).

The simplest way to describe the PSD is to calculate the mean or median particle size. However these measures do not provide an indication of the range of sizes that are present. By also including percentiles, such as the 10^{th} and 90^{th} percentiles, more information relating to the fine and coarse ends of the distribution can be attained. Alternatively the PSD can be presented as a frequency density distribution. In this case the particle sizes are divided into groups and the proportion of particles within each group is calculated and presented as a graph of frequency densities.

5.2.1 Equivalent Particle Diameter

There are several challenges faced when determining and describing the particle size distribution of a powder. In practice particles typically have irregular shapes, so a measurement of the length of the particles will depend on the orientation in which they are measured. Many methods of particle size measurement assume that the particles are spherical and the calculation of the PSD from the measurements is based on this

assumption. Therefore data from non-spherical particles can involve complex interpretation to produce a representative particle size distribution (Shekunov *et al*, 2007).

Determination of the PSD of non-spherical particles through the application of a method that assumes that the particles are spherical involves the use of the equivalent particle diameter (Allen, 1997). To find the equivalent particle diameter, a property such as the length or volume of the particle is found and the equivalent particle diameter is given as the diameter of a spherical particle with the same length or volume. The properties of the particles that can be compared also include the mass, surface area or sedimentation rate. The type of measurement that is used for comparison is selected to best represent the important properties of the powder that is being studied, so particles that have the same equivalent particle diameter will have the same property of interest.

In pharmaceuticals, the assumption of spherical particles is rarely satisfied, since crystallisation and milling processes are commonly used and these processes do not produce spherical particles (Iacocca *et al*, 2010). Therefore methods of particle size measurement which assume that particles are spherical will typically not produce a representative PSD of a pharmaceutical product.

5.2.2 Sampling Error and Dispersion

Measurement of the particle size distribution usually involves taking a sample of the powder, so it is essential that the sample is representative of the whole population and reflects the PSD of the product. When the particles are non-spherical, the orientation of the particles relative to the measuring device will affect the measurement that is taken, so a suitably large sample is required so that the particles will be measured from all angles and all orientations will be captured (Allen, 1997).

The particle size distribution of a powder can also be affected by particles bonding together. Primary particles are the smallest particles present and are held together by molecular bonding. These particles can then become attached to each other to form aggregates or agglomerates (Tinke *et al*, 2008). Aggregates are structures that consist of primary particles held tightly together by atomic or molecular bonding at their crystal faces. Agglomerates are more loosely bonded structures where the primary particles are attached to each other by weaker Van der Waals forces. Since aggregates are held together tightly, a lot of energy is required to break them back up into the primary particles. However agglomerates are more loosely bonded and less energy is needed to break them up.

When making a particle size measurement, it is important that the sample is well dispersed to ensure that individual particles are measured rather than agglomerates (Tinke *et al*, 2009). Dispersion methods can be used to break up agglomerates, for example by shaking, stirring or the ultrasonic treatment of suspensions. De Villiers (1995) highlighted the importance of good sample dispersion by comparing three drug powders with different initial levels of agglomerates. The samples were dispersed either with a dry powder disperser or an ultrasonic bath for liquid suspensions and the PSDs were measured before and after dispersion. For all three powders a smaller mean particle size was measured after the sample had been dispersed.

Conversely, tightly bonded aggregates may remain together during the formulation of a drug product and should therefore be included in the calculation of the PSD. These aggregates should be preserved during sample preparation so that they can be measured to give a true representation of the PSD of the final product (Iacocca *et al*, 2010).

5.2.3 Sieving

Sieving is a simple and widely used method for particle size analysis that is suitable for particles greater than 20 µm in diameter (Allen, 1997). It is considered to be a low cost and reliable method that gives reproducible results (Rhodes, 1998). Particles are separated by their diameters by passing the product through a series of sieves with decreasing hole sizes. The sieves can be hand shaken or mounted onto a vibrator (Coulson, 2007). The diameter of spherical particles that pass through each layer of the sieve can be measured using standards of known particle sizes. The quantity of particles that is collected on each layer is then used to calculate the distribution of the particle sizes in a sample. The resolution of the measurements will be determined by the difference in the sizes of the consecutive sieves.

Sieving is a low cost straight forward method that is easy to set up and implement. However the accuracy of the results is determined by the resolution of the sieves, hence results from sieving will not produce as detailed a particle size distribution as can be obtained from other methods.

5.2.4 Microscopy

An alternative approach for assessing the PSD is through microscopy. In this method an image is generated, enabling the variation in size and shape of individual particles to be viewed. Optical microscopy and scanning electron microscopy (SEM) are commonly used for the analysis of pharmaceutical powders in research and development (Shekunov *et al*, 2007). Optical microscopy can be used for particles in

the range of 1 to 100 µm, whilst SEM can be effective for particles as small as 0.001 µm (Coulson, 2007). The particle size distribution is calculated through image analysis, which measures the size of individual particles observed by the microscope.

Scanning electron microscopy has a much larger resolution and depth of focus than optical microscopy (Allen, 1997). To take a measurement, a fine beam of electrons is scanned across a sample, which interacts with the particles. The signal from the detected electrons depends on the size of particles in the sample and can be interpreted to display an image of the particles (Section 5.2.4.1).

The use of microscopy for particle size analysis is important because it is the only method that allows the size and shape of individual particles to be observed and analysed (Allen, 1997). Microscopy can be used to validate the results from other techniques, since the particles can be measured individually and the accuracy of the alternative method determined (Tinke *et al*, 2008).

As discussed previously, a number of methods for particle size analysis rely on the assumption that the particles are spherical and are well dispersed. To investigate these assumptions a qualitative view of the sample is required (Tinke *et al* 2009). A microscope image allows the variation in the particle size, shape and dispersion to be viewed and the presence of aggregates and agglomerates to be determined.

For microscopy measurements, a sample of particles is dispersed in a liquid to create a suspension. A dilute sample is required to ensure that the particles do not overlap on the microscope slide, therefore enabling each particle to be measured individually (Iacocca *et al*, 2010). This limits the number of particles that can be measured at one time and hence the sample may not be representative of the whole distribution. The measurement time can also be slow (Bosquillon *et a*l, 2001), limiting the use of microscopy for on-line analysis.

When a microscopic image is produced, it shows a 2-dimensional representation of the particles. If the largest surface of the particle is parallel to the slide then the shortest dimension will not be measured, which can lead to an overestimation of the particle size (Allen, 1997).

5.2.4.1 Image Analysis

The microscopy measurements are converted to a PSD through the use of image analysis. From a microscope image, the particles first need to be distinguished from the background (Sarkar *et al*, 2009). A 2-dimensional image consists of a number of pixels,

each with a fixed intensity. The edge of a particle can be detected by observing a large change in the pixel intensity over a short space. When the particle edges have been identified, the area of the particle is filled in and the region around the edge left blank to represent the space between particles.

When the individual particles have been identified the required size properties can be measured and collected to give a distribution. For spherical particles, the diameter of each particle can easily be measured. However for non-spherical particles there are a number of methods for characterising the size of a particle. A possible measure is to find the projected area of the particle and then calculate the diameter of a circle that has the same area (Yu and Hancock, 2008). An alternative measure is Feret's diameter, which is the distance between two tangents in a fixed direction on either side of the particle. For the application of Feret's diameter, it is assumed that the particles are randomly orientated across the microscope slide. The choice of metric to quantify the particle size will affect the resulting PSD, so the same metric must be used when comparing different samples.

5.2.5 Laser Diffraction

Laser diffraction (LD) is generally the preferred method for particle size analysis in the pharmaceutical industry (Iacocca *et al*, 2010). A measurement is taken by passing laser light through a sample of powder, which is scattered by the particles and the angle of the light scattering is measured and interpreted to give a particle size distribution (Section 5.2.5.2). Laser diffraction can be used to measure particles in the range 0.1 to 3600 μm, but a single instrument is not capable of measuring the whole range (Allen, 1997) and hence an instrument must be selected that is appropriate for the particles to be measured.

LD has the advantage of a short analytical time and gives robust and precise measurements, which makes it possible to use for on-line process monitoring (Ma *et al*, 2000). This method can analyse a broad range of particle sizes and is considered to be easy to use (Tinke *et al*, 2008). Furthermore, LD can be used for liquid, spray or dry powder samples (Shekunov *et al*, 2007).

The calculation of the particle size assumes that the particles are spherical, which can limit the accuracy of this method for non-spherical particles. However LD measurements can be validated by comparing the results to a more accurate technique such as microscopy followed by image analysis. The difference between LD measurements and the results from other methods can also provide information on the

shape and structure of the particles, since the method errors may be a result of certain characteristics such as non-spherical particles (Kaye *et al*, 1999).

The results from laser diffraction can vary between different instruments and manufacturers, since each will have a unique algorithm for interpreting the particle size from the light scattering results. These differences suggest that the resulting PSD that is calculated may not be completely accurate, but by validating the results with microscopy, a reliable representation of the PSD can be produced (Iacocca *et al*, 2010). In addition, the high precision of LD allows for the detailed comparison of samples measured with the same instrument.

5.2.5.1 Laser Diffraction Instrumentation

For the measurement of the PSD with laser diffraction, a Helium-Neon laser is used to produce monochromatic light at a set wavelength (Allen, 1997). The light is passed through a stream of dispersed particles that are held in suspension. The light is scattered by the particles and a lens is used to focus the light onto a photosensitive silicon detector, which comprises several concentric rings. The detector measures the angle and intensity of the scattered light; with smaller particles scattering light at a wider angle.

Typically a number of repeated measurements are taken and the results are averaged to give the PSD (Ma *et al*, 2000). For each measurement different particles may be detected because the particles move within the suspension. For example, if there are a few large particles present, they may not be measured in every sweep and their presence may not be captured when the results are averaged.

5.2.5.2 Data Interpretation

When the light interacts with a particle that is held in suspension, the light is both scattered and absorbed by the particle (Beekman *at el*, 2005). The light can be scattered by diffraction, reflection or refraction (Figure 5-1). The intensity and angle of the scattered light depends on the size and refractive index of the particle.

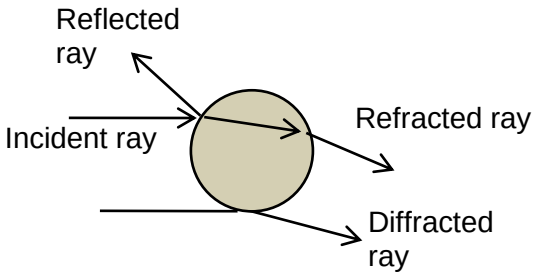


Figure 5-1: Light scattering by a particle (Adapted from Allen, 1997)

A light scattering pattern can be converted into a size distribution using one of two theories, Mie or Fraunhofer theory. Both are based on the assumption that the particles are spherical and require knowledge of the refractive index of the particles (Allen, 1997). Mie theory is more complex than Fraunhofer. The Fraunhofer method can only be used when the particle size is significantly larger than the wavelength of the incident light and is generally suitable for particles larger than 25 µm in diameter (Beekman *et al*, 2005).

Mie theory was developed in 1908 and applies the Maxwell electromagnetic equations, which describe light scattering through a set of partial differential equations (Ma *et al*, 2000, Beekman *et al*, 2005). The equations describe the light scattering effect that is expected for a given spherical particle size. Therefore derivation of the PSD from the light scattering requires the Maxwell equations to be inverted. Given the measured angles and intensities captured by the detectors, the PSD of the sample is calculated for the equivalent spherical particles. This is an iterative process that may not produce a unique solution. The algorithm that is used by a particular instrument will be held by the manufacturer and may not be readily available for comparison with other instruments (Iacocca *et al*, 2010).

5.2.6 Comparison of Methods

Different measurement methods make use of different physical properties of the particles to produce a PSD and hence the results will vary between methods (Haskell, 1998). Consequently it is beneficial to compare several methods of PSD measurement to gain a good understanding of the PSD of the powder that is being studied.

Tinke *et al* (2008) compared the particle size distributions generated for eight powders using both laser diffraction and static image analysis of optical microscopy images. The same samples were analysed using both methods, to avoid sampling error. The powders ranged from spherical particles to long rectangular shaped particles, with sizes between 10 and 500 µm in length. The results were compared by overlaying the particle size distributions for each sample and there was clear evidence of a strong

correlation between the two analytical methods. However the laser diffraction method produced a slightly smaller PSD for particles with a large length to width ratio, reinforcing the fact that greater differences are seen when the assumption of spherical particles is not satisfied.

In the study by Tinke *et al* (2008), wet and dry dispersion techniques were also compared for a sample of spherical particles. Although the main peak of the distribution was the same for each method, the dry dispersion resulted in some of the particles touching each other and hence they were measured as single large particles, causing a second peak at a larger particle size for both the LD and microscopy methods. Overall it was concluded that the laser diffraction results correlated well with the image analysis measurements and hence the LD results could be considered to be representative of the particle size distribution.

Bosquillon *et al* (2001) compared the particle size measurement of microscopy and laser diffraction for the analysis of dry powders with varying levels of aggregation. For each method the median particle diameter was calculated. Light microscopy and electron microscopy were used as reference techniques since the particles could be observed individually. These two microscopy methods were able to separate out the individual particles from the aggregates that were present, so the particle size measurements were not considered to be influenced by the presence of aggregates. Laser diffraction measurements were taken of the particles in both a dry state and when suspended in water, to determine if either method was capable of dispersing the particles and removing aggregates.

For the less aggregated powders good agreement was observed between the median particle diameter estimates for all the methods. However for the most cohesive powder, only the LD method in a wet state was able to produce a measure close to that attained by light microscopy. The LD analysis in the dry state overestimated the particle sizes due to the presence of aggregates. The most aggregated powder could not be analysed by electron microscopy because there were too few individual particles compared with the number of aggregates. In summary this study showed that aggregation can have a significant effect on particle size measurements, thus sample dispersion should be considered when selecting a method.

5.2.7 Conclusions

The particle size distribution of a pharmaceutical product can influence a number of properties of a drug, including its behaviour during formulation, content uniformity and dissolution rate. A number of methods have been described that can be used to

measure the PSD of a powder, which vary in how the sizes are measured and the PSD calculated. Many techniques make the assumption that the particles are spherical and therefore may not produce accurate results for non-spherical particles.

The presence of agglomerates and aggregates can also lead to an over estimation of the particle size, so appropriate dispersion techniques may be required to break them up. However if tightly bonded aggregates are expected to remain within a product then their presence should be measured and included in the PSD.

Sieving is the most straight forward and inexpensive method to measure the PSD, but it lacks the speed and accuracy required to produce a detailed particle size distribution. Microscopy is an important method because it allows the individual particles to be viewed and measured. With microscopy the assumptions of spherical particles and sample dispersion can be assessed. Scanning electron microscopy can be used to measure particles as small as 0.001 µm, with image analysis then applied to calculate the PSD from the microscope images.

Finally laser diffraction was considered, it is a fast and precise method that is suitable for use online. For LD, the interaction of laser light with the particles is measured and interpreted to produce a PSD. The response can vary with different instruments but this method can be validated by comparing the results with the PSD generated from microscopy and image analysis. It has been shown that there is generally good agreement between image analysis and LD results, but the presence of aggregates can result in the over estimation of the particle size by laser diffraction.

The PSD analysis techniques described in this section were applied to a product manufactured at AstraZeneca, to gain an understanding of the PSD of the product and the factors that may cause variation in the particle size. The results are presented in the subsequent sections,

5.3 Particle Size Distribution Study

At AstraZeneca, a project was initiated to characterise the particle size distribution of a particular API product. The project was undertaken to measure the PSD of the product and then to understand the variation in the PSD by assessing the differences between plants and processes. A large amount of PSD data was collected, to which multivariate analysis methods were applied to interrogate the data and present the results.

For the manufacturing process under consideration, a solid powder is formed and then isolated on a filter dryer. The powder is then milled to reduce and homogenise the

particle size. Following the API manufacture, the product is transferred to a different site for formulation into tablets. The product is produced in one of two processing plants, and can be re-worked in a recovery processes (Section 5.3.1).

The first objective of this project was to create a baseline of the current particle size distribution of the milled product. Previously, the particle size had been measured using a sieve test to quantify the amount of material that passed through a fixed size sieve, to ensure that the required quantity of particles were below this size. The particle sizes were measured in more detail using scanning electron microscopy and laser diffraction measurements taken for a number of samples.

Understanding more about the particle size distribution of the product will allow AstraZeneca to determine the range of particle sizes that are currently manufactured and which have been shown to be suitable for formulation into tablets. Consequently, if future changes are to be made to the process then the PSD of new material can be compared against the known PSD of the current material to determine if the changes have had an impact on the PSD. Samples were taken from both processing plants and from the main and recovery processes, hence from the analysis of the data it will be possible to determine whether there are differences between the product manufactured on the different plants and processes.

Batch to batch variability in samples produced by the same plant and process will be a consequence of the variation in either the process to produce the solid API product or in the milling process itself. Variables collected during these processes were related to the final PSD of the product through partial least squares (PLS) and neural network models (Section 5.4). By understanding how the process can influence the final PSD, the company will have tighter control of the particle size, by either reducing the variation in the final product or by working to change the PSD profile if it is determined that a different PSD may be preferable the formulation into tablets.

A future aim of this work is to link the PSD to the behaviour of the product during the formulation process. The particle size will affect how well the API mixes with the excipients used to make up the tablets. If the particles are too large or too small, then the API and excipients will not mix well, resulting in tablets being made with a product content that is outside of the specification limits. If data can be provided of the variation in the product content for the tablets manufactured from each batch of API, then it may be possible to determine the ideal range of particle sizes that will produce tablets with consistent product content, therefore increasing the yield of the tableting process.

5.3.1 Manufacturing Process

The manufacture of the API being studied is a batch process that takes place in one of two processing plants. Primarily plant 1 is used for manufacture, but a second plant is brought into operation when greater capacity is required. The solid product is formed from a precipitation reaction, before being isolated on a filter dryer and then transferred to the milling facility.

Product can be recovered from the liquors from the filter dryers and reprocessed in the recovery process, which can be run on either plant. There are significantly fewer batches manufactured on the recovery process compared with the main process and hence the majority of the data that has been collected is from the main process on plant 1.

The main process is the same in both plants, but the batch sizes in plant 1 are approximately 1.5 times larger than those in plant 2 and hence the quantities of all the raw materials are scaled accordingly. The recovery process contains the same steps but begins with a different starting material, consequently different solvents are added at the start. From the precipitation stage, where the particles are formed, the main and recovery processes are the same

5.3.2 Milling Process

The aim of the milling process is to reduce and homogenise the particle size of the API so that the product is suitable for formulation into tablets. A schematic of the mill is shown in Figure 5-2. The material is transferred to the milling facility in an Intermediate Bulk Container (IBC), which is positioned at the top of the mill. Material is discharged from the IBC into the feed hopper which is attached to a set of weigh scales. The weigh scales are used to control the flow of product into the screw feeder. When the weight drops to a specified level, the valve from the IBC is opened and more material is discharged into the feed hopper. From the weigh scales, the loss in weight is used to calculate the overall weight throughput of material flowing into the mill.

The material is carried into the mill in a stream of cooled nitrogen gas, which is used to provide an inert atmosphere. In the nitrogen gas flow, the product enters the milling chamber where impaction at the beater causes the particles to reduce in size. The milled material is then carried to the reverse jet filter, where the nitrogen gas is separated off and the product is discharged into approximately 15 kegs, depending on the batch size.

For each batch that is milled, a sample of material is taken as the product is discharged into the kegs. Samples are taken approximately 25kg into the batch discharge and 50 kg before the end, this is usually from the second and penultimate kegs. The two samples are mixed together to create a 'blend' sample that is assumed to be representative of the whole batch. The sample is used for release testing by the Quality Control (QC) department.

For the purpose of the PSD study, this sampling approach may not capture the full range of the PSD for a specific batch. Changes to the milling conditions during a batch, for example to the product feed rate, could cause variability to the PSD in different parts of the batch. Ideally a sample would be taken from every keg, or from a number of kegs, from the same batch and analysed separately to determine the variation in PSD across a batch. However any extra samples taken would reduce the overall yield of that batch, and the high cost of the material prevents the company from allowing extra samples to be taken.

5.3.3 Particle Size Distribution Data

The first stage in the analysis of the particle size distribution was to take samples from a number of batches to create a baseline dataset to determine the current PSD of the product. Most of the samples were from the main process on the first plant, but plant 2 and recovery process batches were also analysed. The samples analysed consisted of the standard QC samples that are taken from every batch. All of the batches that were analysed for the baseline study were suitable for release.

The particle size distribution measurements were taken with a Sympatec Dry Dispersion Laser Diffractor. Samples were initially analysed at the AstraZeneca's Site A, to provide a baseline dataset of the current PSD of the product being manufactured. Following this work, a new Sympatec Helios Laser Diffractor was installed at Site B, where manufacture takes place, and a new set of samples were analysed. The two instruments are similar, but different algorithms are used to calculate the PSD from the laser diffraction measurements, so some differences may be expected.

An initial comparison was undertaken of the material before and after milling, prior to a subsequent analysis that focused on samples taken after the batches were milled. The results of the analysis are presented in Section 5.3.4 and Section 5.3.5.

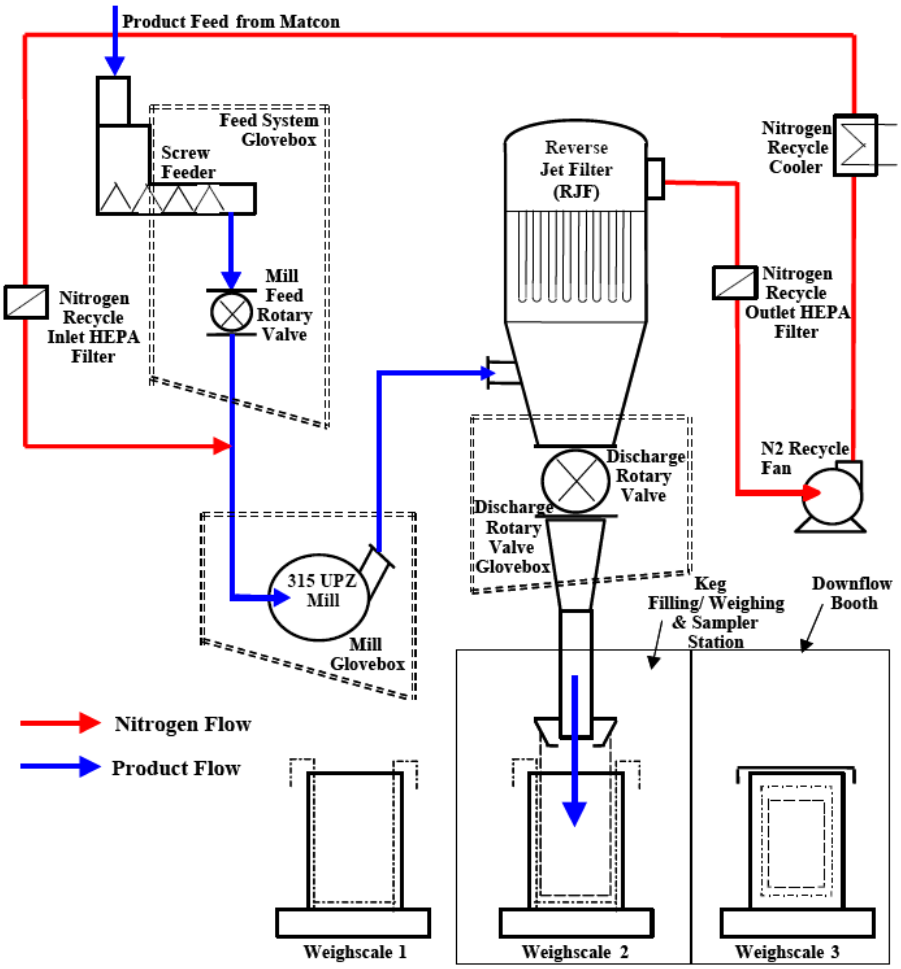


Figure 5-2: Mill Schematic

5.3.4 Milled and Unmilled Material

In the original characterisation study that was conducted at Site A, a comparison was performed of the API material before and after the milling process. Figure 5-3 compares the PSD profiles of a sample of unmilled material and a sample of milled material from the same batch. Figure 5-3 clearly shows that the milled material comprises a higher percentage of finer particles than the unmilled material, shown by a shift to the left of the PSD for the milled material. All particle size values are removed to protect commercial confidentiality.

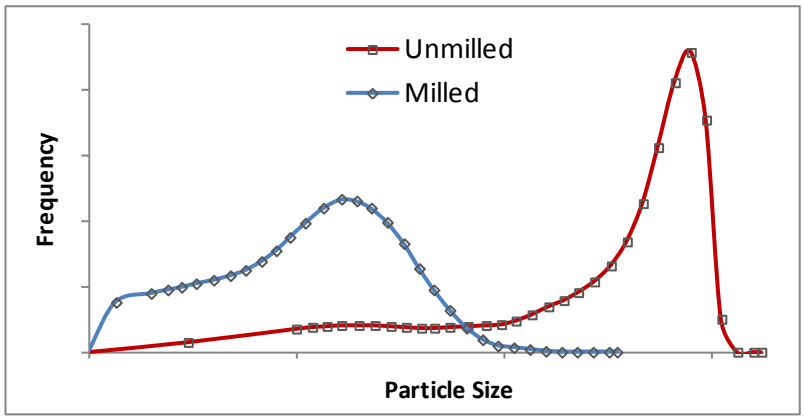


Figure 5-3: Unmilled and milled PSD of a plant 1 main process batch

A number of samples of unmilled material that were produced on plant 1 and plant 2 were analysed. The samples from plant 1 included product from both the main and the recovery process, which showed an approximately consistent profile (Figure 5-4). However the batches that were produced on plant 2 have a greater proportion of fine particles and a multimodal distribution (Figure 5-5). These two figures have the same x-axis scale. Although the manufacturing processes are the same in each plant, the precipitation vessels are not identical. For the main process, the batch sizes are larger for plant 1 and the speed of the agitator in the vessels differ between the two plants, causing the rate of the precipitation reaction to differ. It is expected that the precipitation process happens more quickly in plant 2, leading to a greater number of fine particles forming.

Scanning electron microscopy (SEM) images were taken of samples from each plant (Figure 5-6 and Figure 5-7), the two figures have the same magnification. These support the results from the laser diffraction measurements, showing that plant 2 unmilled material contains some large particles that are similar to those found in plant 1 material from the recovery process, but there are also a greater proportion of fine particles.

Figure 5-8 and Figure 5-9 show the PSD of milled material from plant 1 and plant 2 respectively, with the same x-axis scale. The material from plant 2 still has a slightly higher proportion of fine particles, but these graphs suggest that the mill is effective in increasing the similarity of the PSD from the two plants and reducing the overall particle size. Note that the x-axis scales of Figure 5-4 and Figure 5-5 differ to those of Figure 5-8 and Figure 5-9.

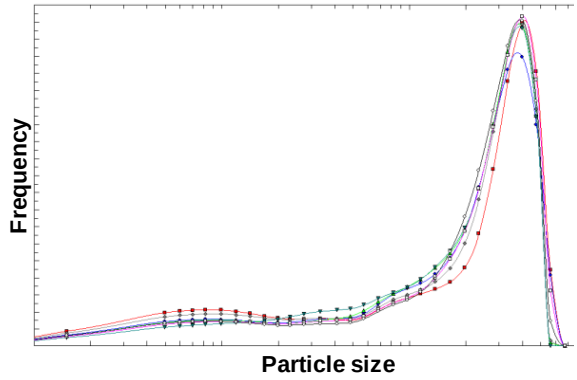


Figure 5-4: Plant 1 unmilled PSD profile, each sample shown in a different colour

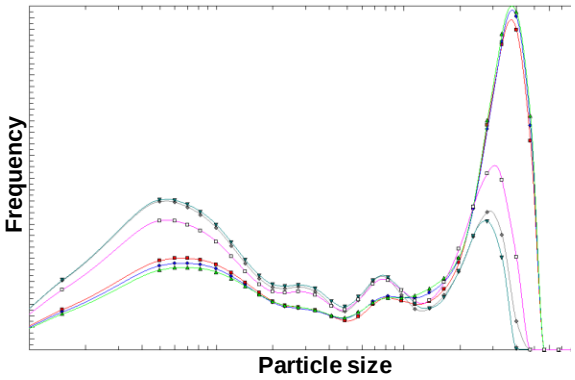


Figure 5-5: Plant 2 unmilled PSD profile

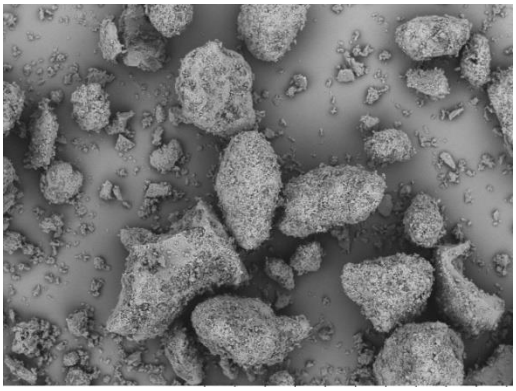


Figure 5-6: SEM of plant 1 recovery process unmilled material

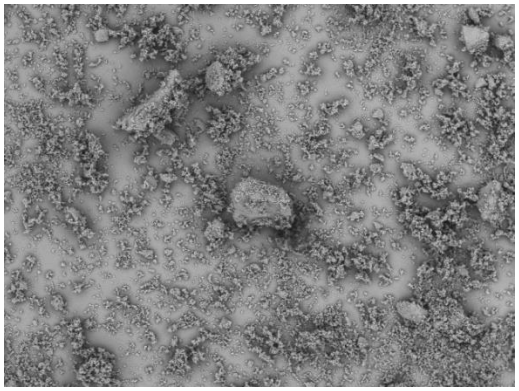


Figure 5-7: SEM of plant 2 main process unmilled material

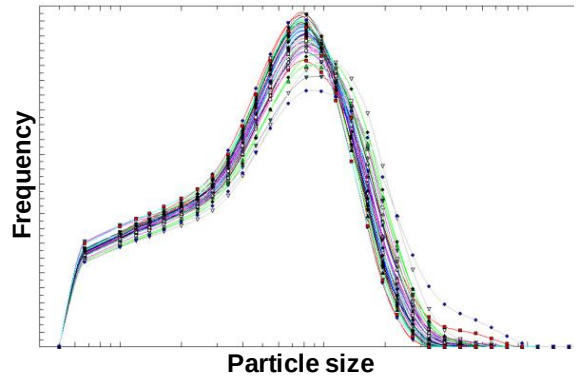


Figure 5-8: Plant 1 milled profile, note the x-axis scale is different to the graphs of unmilled material

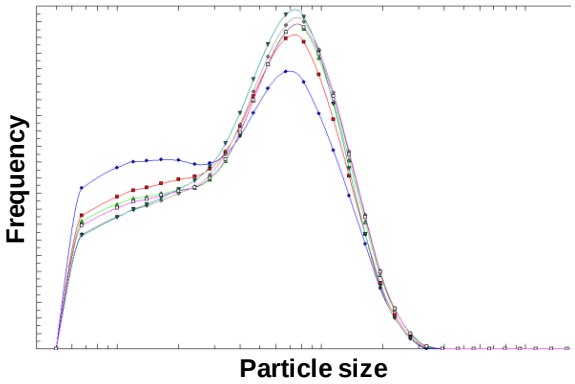


Figure 5-9: Plant 2 milled profile

5.3.5 Principal Component Analysis

The data that is collected from the PSD measurements can be analysed through the application of multivariate analysis techniques. Strong correlations will exist within the PSD data and hence principal component analysis is an appropriate technique to summarise the data into a number of PCs, enabling the assessment of trends and the identification of potential outliers or unusual behaviour.

Samples were initially analysed at Site A, with samples from more recent batches analysed at Site B. The majority of the batches were from the plant 1 main process, but

batches manufactured on plant 2 and from the recovery process were also included, depending on where batches were being manufactured at the time of analysis (Table 5-1).

Plant / Process	Site A	Site B
Plant 1 Main	35	32
Plant 1 Recovery	1	3
Plant 2 Main	6	1
Plant 2 Recovery	0	5

Table 5-1: Number of batches analysed for each plant, process and location

Frequency data was collected for 32 particle sizes. However for sizes larger than the 23rd measurement the frequencies were zero for many of the batches, hence there was not enough information in the data to be able to compare batches and these particle sizes were removed from the analysis. The resulting data set contained 23 variables that reflected a range of particle sizes.

5.3.5.1 PCA Model 1: Site A Data

A principal component analysis representation was created of the PSD data from the 35 plant 1 main process batches analysed at Site A, using the software SIMCA-P+ 12.0.1 (Umetrics AB, Umeå, Sweden). The data was scaled to be mean centered with unit variance. A PCA model comprising the first three principal components captured 99% of the variation in the data (Table 5-2). The Q^2 value, 0.97, is close to the R^2X, suggesting that the model is not over fitted and will be applicable to new data. Therefore PSD data from future batches can be applied to PCA model 1 to assess if there is a difference in the PSD.

Number of PCs	R^2X (cumulative)	Q^2
1	0.80	0.75
2	0.93	0.83
3	0.99	0.97

Table 5-2: Model fit for PCA model 1

From the PCA representation, the loadings identify which variables (particle sizes) have the greatest impact on the variation in the data, for the individual principal component. The loadings for the first principal component (PC) separate out the sizes above and below the main peak in the distribution (Figure 5-10), an example of an individual PSD is shown in Figure 5-11. Samples with a high PC score for this component will comprise a high proportion of fine particles. The second component

represents the height of the peak in the distribution, at around the 14th size measurement. The loadings from the third component identify further regions of the PSD where there are differences between batches, particularly around the tails of the distribution.

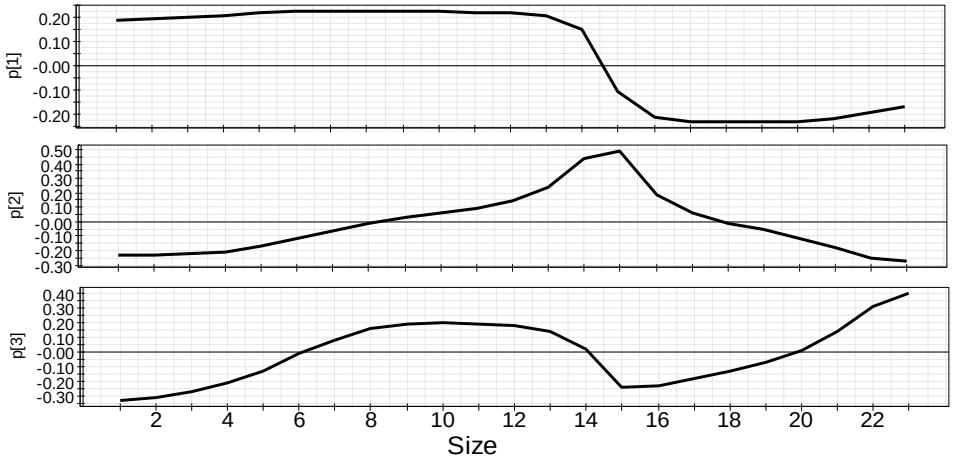


Figure 5-10: Loadings for PCs 1, 2 and 3 (sizes are nominal values)

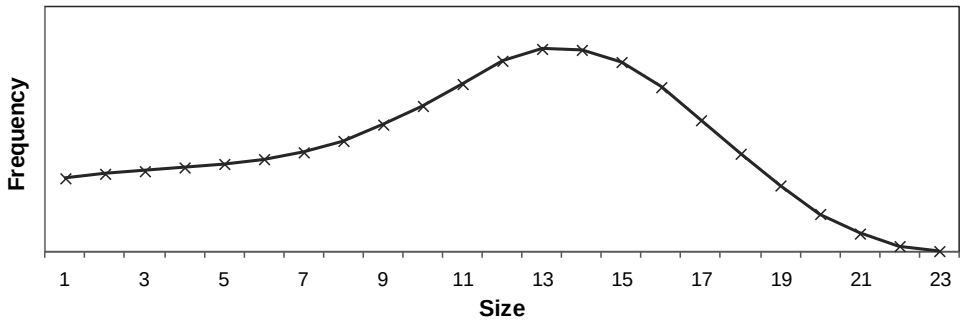


Figure 5-11: Example of a PSD for a batch used to develop PCA model 1

The principal component scores plots show the similarities in the profiles of all the batches in the dataset (Figure 5-12 and Figure 5-13). Two batches stand out as exhibiting unusual behaviour, with each batch outside of the 95% confidence bounds for at least one of the principal components. These batches are highlighted in red and green.

The batch highlighted in red has a higher proportion of the most coarse particles compared to the rest of the batches and a lower peak height (Figure 5-14), identified through the scores in the first and second components. During the milling process for this batch it was noted that there were several spikes in the feed rate of the product into the mill, which may have caused the milling to be less effective, resulting in more coarse particles. The batch highlighted in green has a high proportion of fine particles

and low peak height compared to the other batches, which was identified from the second and third PCs, for which this batch had low score values.

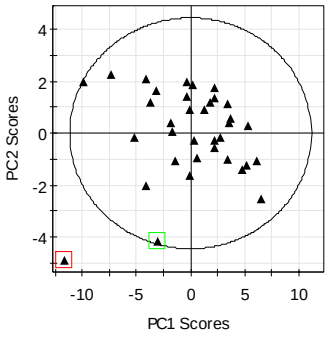


Figure 5-12: Scores plot for the first two PCs, highlighting two outliers

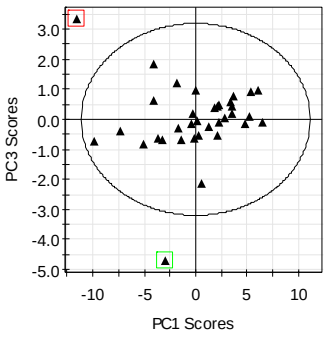


Figure 5-13: Scores plot for PC1 and PC3, highlighting two outliers

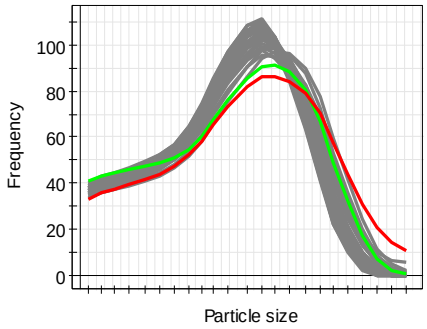


Figure 5-14: Profiles of plant 1 main batches, highlighting two outliers

Within the dataset collected at Site A there was data from six batches from the main process on plant 2 and one from the plant 1 recovery process. These batches were applied to the PCA model as a prediction dataset and the scores values predicted (Figure 5-15). The Hotelling's T^2 and DModX values for the batches produced on plant 2 all lay outside of the 95% confidence intervals (Figure 5-16 to Figure 5-17). The plant 2 batches all had high scores in PC1, which corresponds to a high level of fine particles, and low scores in PC2 corresponding to a low peak height. These differences were also observed when the original PSDs were overlaid (Figure 5-18). It was shown previously that there are large differences between the unmilled material manufactured on the two plants (Section 5.3.4), although the milling process increases the similarity of the two products, differences in the particle sizes is still evident.

For the batch from the plant 1 recovery process, the scores lie within the rest of the data from plant 1. However the DModX value falls just above the 95% confidence limits, suggesting the there is a slight difference in the PSD profile for the batch manufactured on the recovery process. The plot with the distributions overlaid shows that the peak for the recovery batch is slightly to the right, indicating that the material is more coarse than the rest of the plant 1 batches.

It is important to note that despite the differences in PSD between plant 1 and plant 2 material, all of the batches included were suitable for release, indicating that there is a range of acceptable particle sizes.

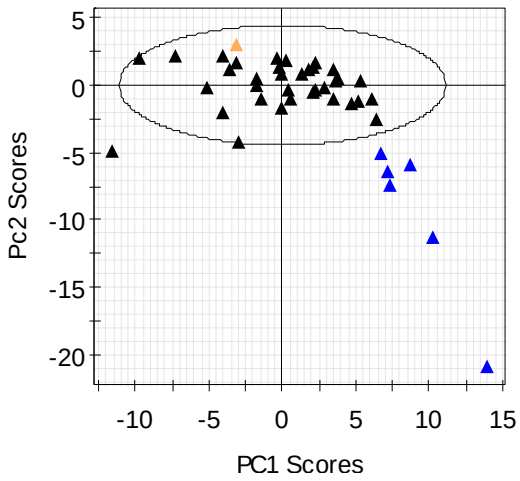


Figure 5-15: Scores plot for the first two PCs, with the prediction set

▲ Plant 1 main
▲ Plant 1 recovery
▲ Plant 2 main

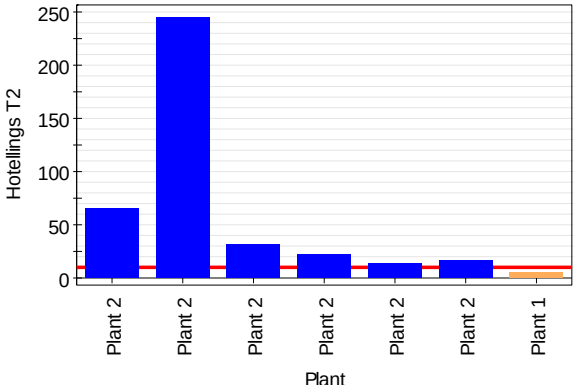


Figure 5-16: Hotelling's T^2 for the prediction data set, 95% confidence level

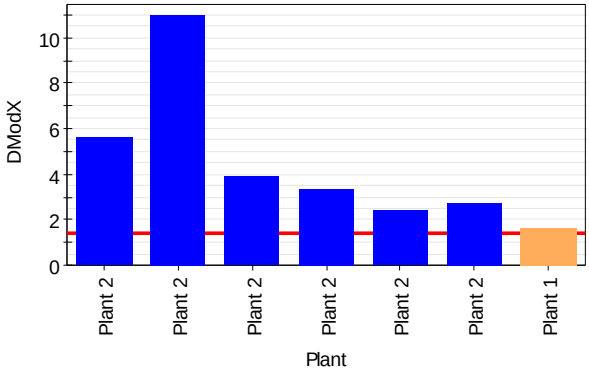


Figure 5-17: DModX (PC1) for the prediction data set, 95% confidence level

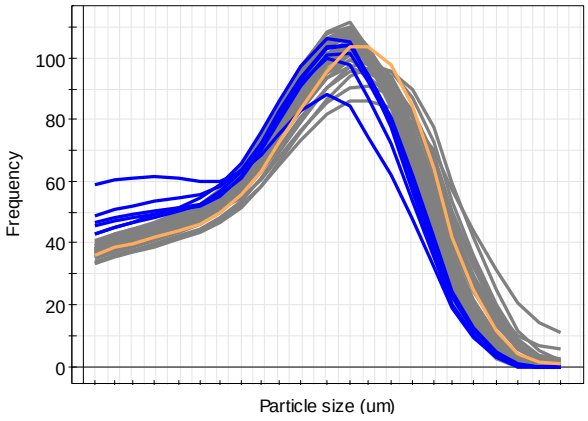


Figure 5-18: PSD profiles of all batches analysed at Site A

5.3.5.2 PCA Model 2: Site B Data

From the dataset collected at Site B, the 32 batches from the main process on plant 1 were used to create a second PCA representation. Three principal components were retained in the model, explaining 99% of the variation in the data (Table 5-3). For this dataset, no unusual batches were observed from the scores plots (Figure 5-19 and Figure 5-20) and the loadings exhibited similar trends to model 1 (Figure 5-21).

Number of PCs	R^2X (cumulative)	Q^2
1	0.88	0.86
2	0.97	0.96
3	0.99	0.99

Table 5-3: Model fit for PCA model 2

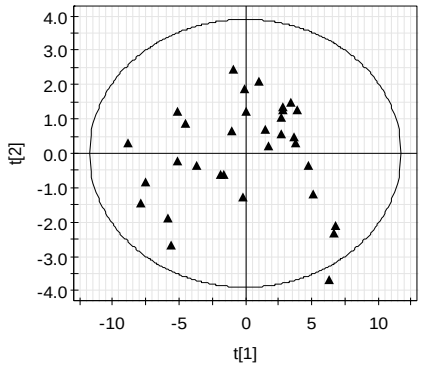


Figure 5-19:Scores for the first two PCS, with 95% confidence bound

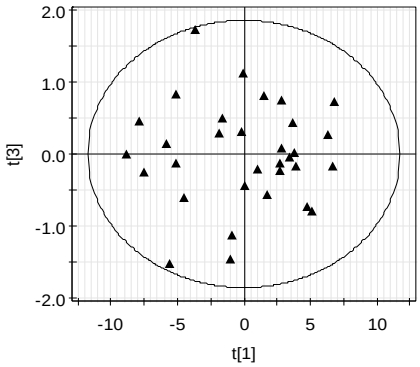


Figure 5-20: Scores for PC1 and PC3, with 95% confidence bound

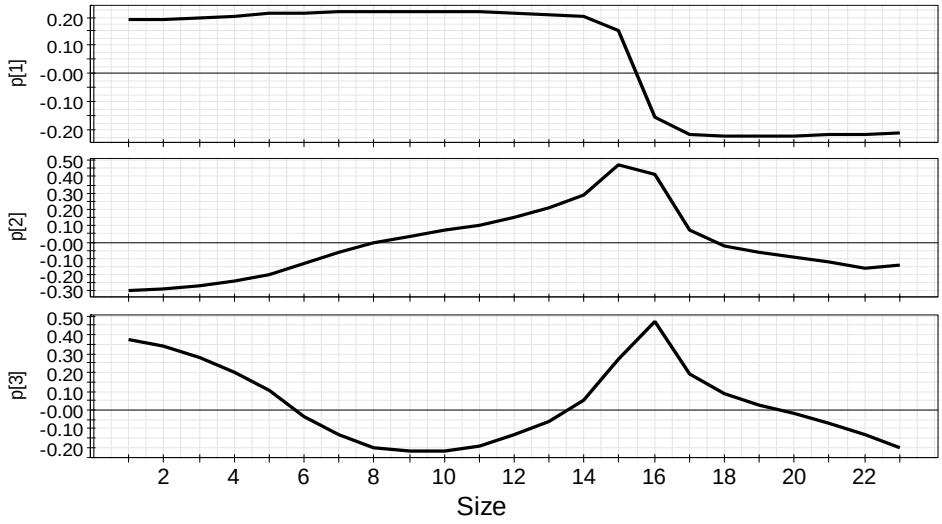


Figure 5-21: Loadings for PCs 1, 2 and 3, model 2 (sizes are nominal values)

The dataset collected at Site B includes batches from the plant 1 recovery process and both processes on the second plant, these batches were applied as a prediction set to PCA model 2. Similar to the data collected at Site B, the scores from the plant 1 recovery batches align fairly closely with the plant 1 main batches that were used to build the model, but form a cluster to one side (Figure 5-22). The Hotelling's T^2 and DModX values are all close to the 95% confidence limits and hence the differences are close to significant (Figure 5-23 and Figure 5-24)

However the difference between batches produced on the two plants is much more pronounced. The Hotelling's T^2 and DModX values for these batches are much larger than the 95% confidence limits; these batches have a lower peak height and higher

proportion of fine particles (Figure 5-25). The results show that the plant where the product is manufactured has a greater influence on the particle size distribution than the process that is operated prior to the solid particles being formed.

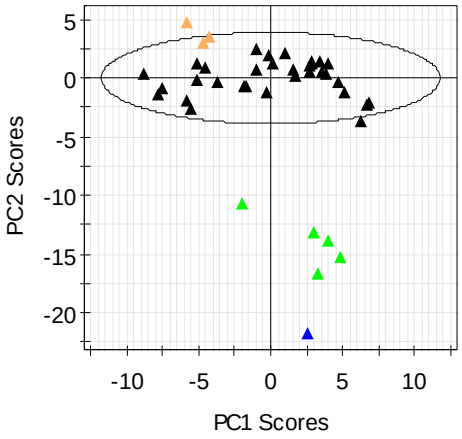
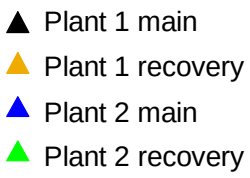


Figure 5-22: Scores plot of first two PCs, with the prediction data set



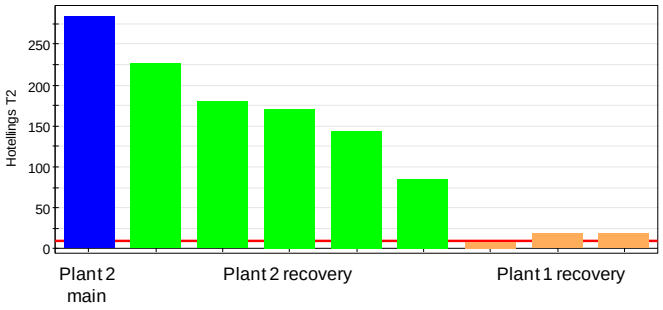


Figure 5-23 Hotelling's T^2 for the prediction data set

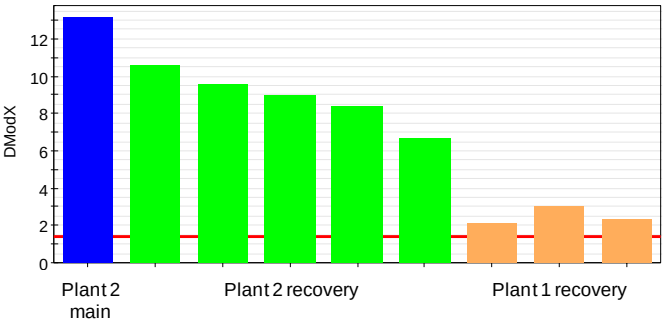


Figure 5-24: DModX (PC1) for the prediction data set

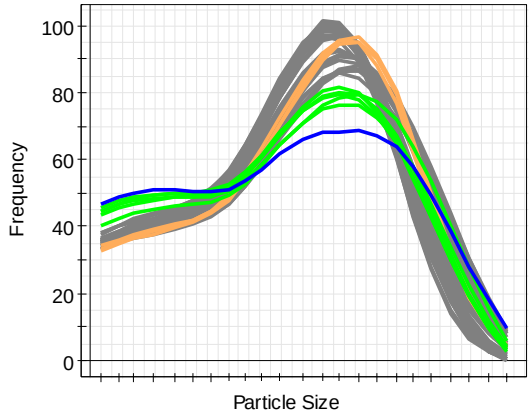


Figure 5-25: PSD profile of all batches analysed at Site B

5.3.5.3 Comparison of Site A and Site B Data

Different instruments were used to measure the particle size of the samples at the two sites, which may cause differences in the results. Separate PCA representations were created for the two datasets, so any differences in results can be investigated by applying each dataset as a prediction set to the model from the other site. The

Hotelling's T^2 and DModX statistics show how closely the data in the prediction data set fits the data used to create the initial PCA model.

It is important to note that different batches were analysed at each site, and the batches analysed at Site B were manufactured after those analysed at Site A, so the differences could be attributed to either instrument or batch differences. Trending the scores for each dataset over time does not show a drift in values (Figure 5-26 and Figure 5-27), suggesting the particle size has not changed over time and hence any differences in the results from the two sites are due to either instrument differences or a step change in PSD.

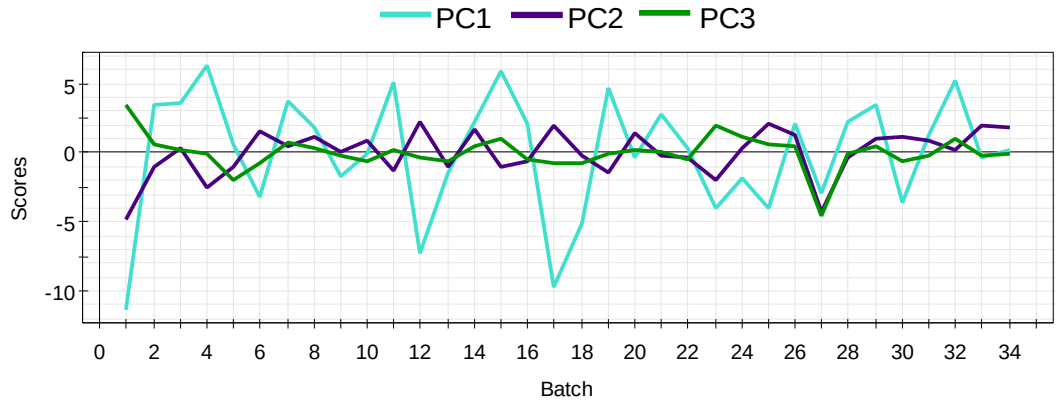


Figure 5-26: Scores from Site A model and batches, in make order

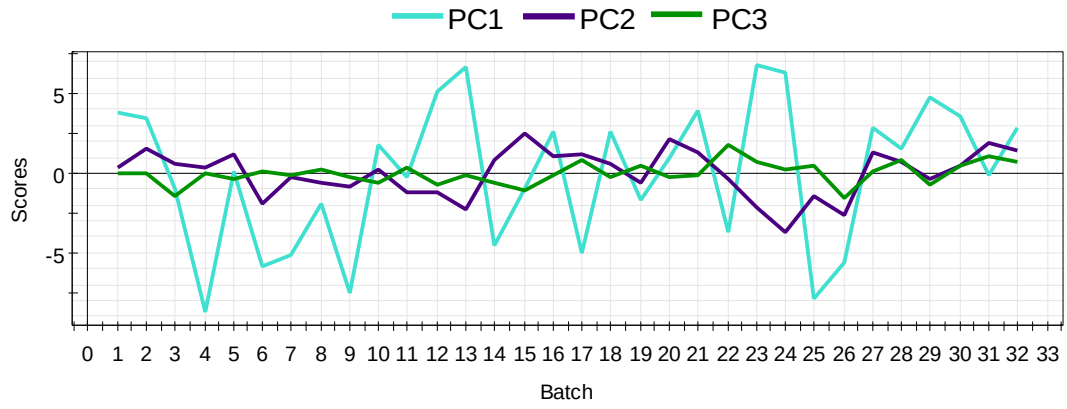


Figure 5-27: Scores from Site B model and batches, in make order

The majority of the batches analysed at Site A have a high Hotelling's T^2 when applied to the Site B model (Figure 5-28). However for the Site B data, many of the batches fit the Site A model, with these batches generally having a lower Hotelling's T^2 values. This difference suggests that there may be greater variation in the Site A dataset. The DModX values are large for both datasets when applied to the opposite model, suggesting that the underlying shape of the distribution may be different for each site

(Figure 5-29). Overlaying the plant 1 main process data for each site shows that samples analysed at Site B showed a higher proportion of coarse particles than those analysed at Site A (Figure 5-30).

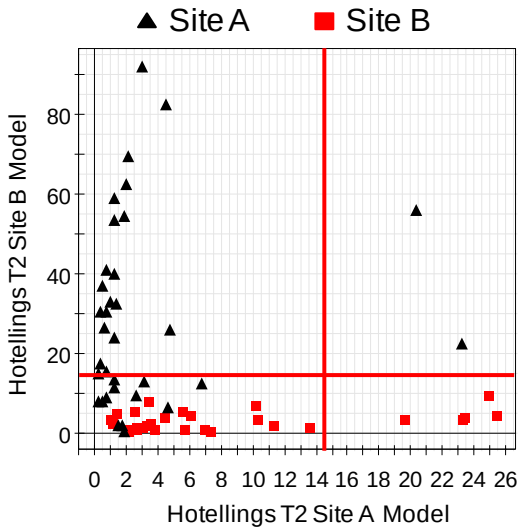


Figure 5-28: Hotelling's T^2 for Site A and Site B data applied to model 1 and model 2, 95% confidence levels shown

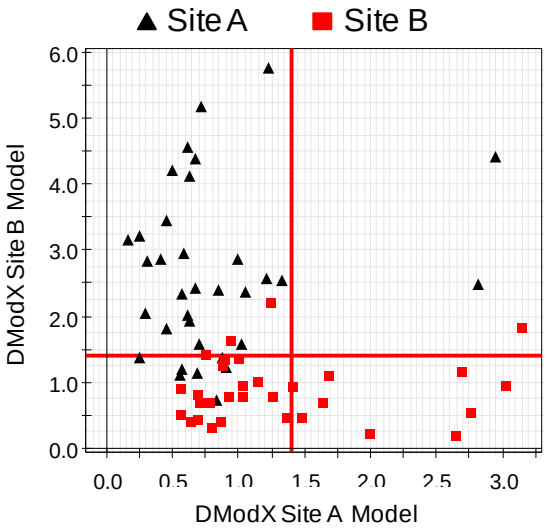


Figure 5-29: Distance to Model for Site A and Site B data applied to model 1 and model 2, 95% confidence levels shown

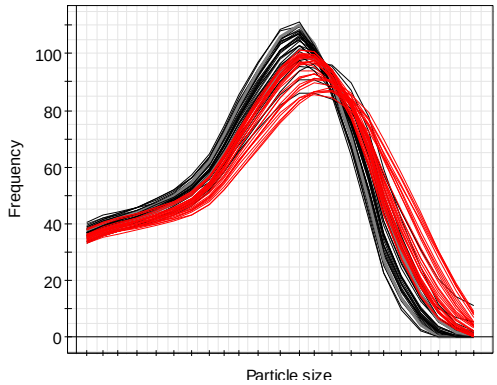


Figure 5-30: PSD of plant 1 main process batches analysed at Site A (black) and Site B (red)

5.3.6 Percentiles

An alternative method for analysing particle size distribution data is to compare the percentiles of the distribution. The 10th (D10), 50th (D50) and 90th (D90) percentiles are often used to summarise the information from a particle size distribution (Figure 5-31).

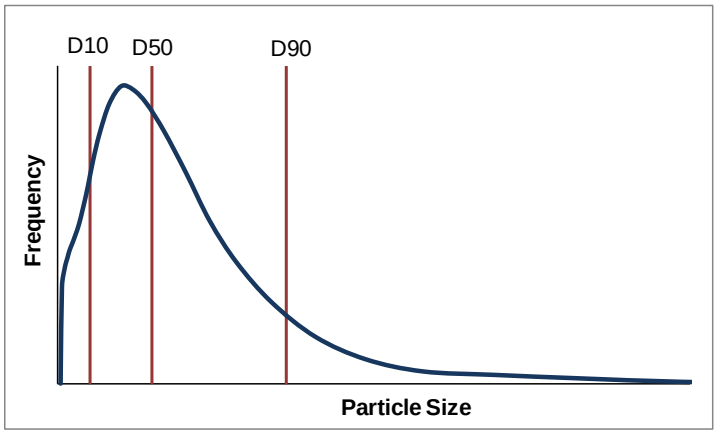


Figure 5-31: Percentiles of a particle size distribution

The measured PSDs of batches manufactured on plant 2 showed a higher proportion of fine particles, which is reflected in the low D10 value, more specifically the lower 10% of the PSD is more fine for plant 2 than plant 1 (Figure 5-32). The 50th and 90th percentiles highlight the difference between the data generated at Site A and Site B (Figure 5-33 and Figure 5-34), the batches analysed at Site B were measured to have a higher proportion of coarse particles and hence higher D50 and D90 values. It is important to note that it is not clear whether the differences are due to the measurements of the two laser diffraction instruments that were used or genuine differences between the material in the batches.

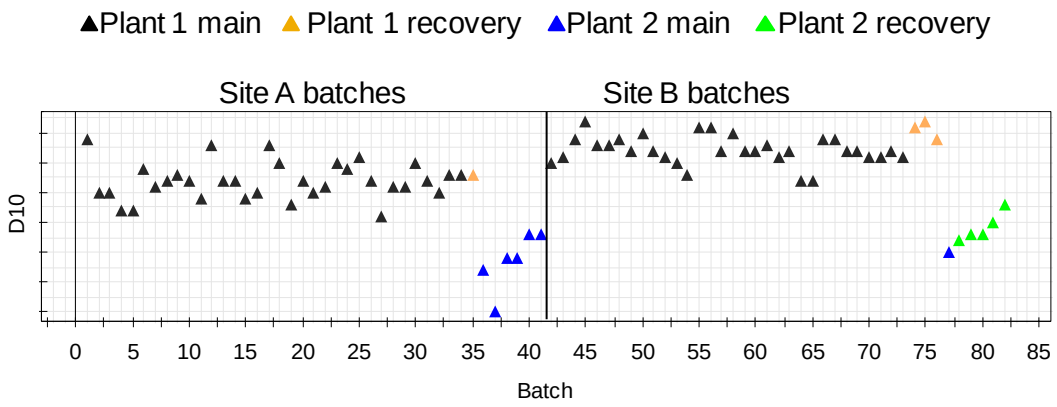


Figure 5-32: 10th percentile of particle size distribution data, for each batch

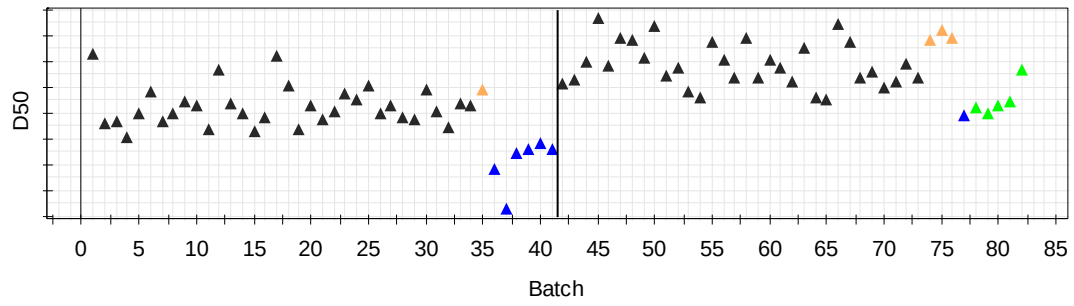


Figure 5-33: Median particle size for each batch

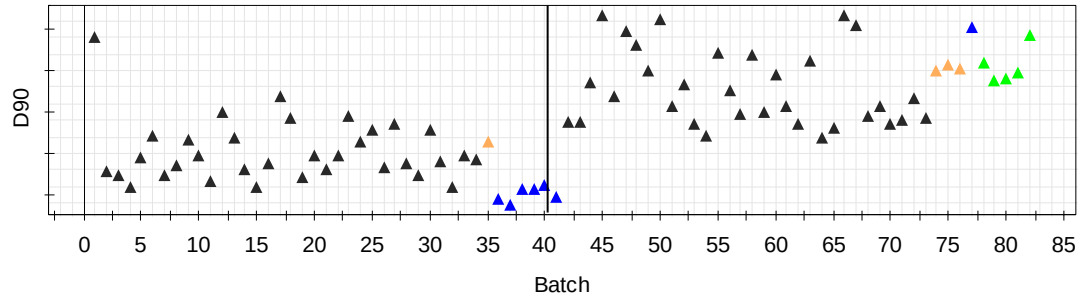


Figure 5-34: 90th percentile of particle size distribution data, for each batch

5.3.7 Measurement Error

Further understanding of the batch to batch variation can be achieved by analysing the variation in the PSD measurements. To determine the effect of changes in the manufacturing process, small changes in the PSD will need to be detected and hence the magnitude of the measurement error is required to be known to determine the size of a change in the PSD that can be identified.

When the particle size measurements were recorded, each sample was analysed three times and the mean calculated. The repeatability of the particles size measurements was investigated by comparing the individual measurements from each batch, for the data from the plant 1 main batches analysed at Site B. Ideally the repeats within each batch should be similar so that differences between batches can be identified. If the variation in the repeatability of the measurement is too large, the batch to batch variability may not be fully captured and true differences in the PSD will not be identified.

5.3.7.1 PCA Model 3

The repeatability of the PSD measurements was quantified by creating a PCA representation using the 35 batches each with three repeats, generating a dataset of 105 samples and 23 PSD variables. From the resulting PCA model, 85% of the variability in the data was captured by the first principal component (Table 5-4). Figure

5-35 shows the scores for this PCA model, colour coded for each batch. The three repeats for each batch tend to be grouped close together, suggesting that the analytical method has good repeatability.

The variation in the data is caused by a combination of batch to batch variability and measurement repeatability. The contribution of each source of variation can be estimated by calculating the components of variation, the calculations for which are shown in Appendix 3. To measure the repeatability, the components of variance were found for the first set of scores from PCA model 3 (Table 5-5). The repeatability of the method contributes 2.4% of the total variation in the data, which is small enough to suggest that the measurement system is able to identify differences between batches (AIAG, 2002).

Number of PCs	R^2X (cumulative)	Q^2
1	0.85	0.83
2	0.97	0.96
3	0.99	0.99
4	1.00	0.99

Table 5-4: Fit of PCA model 3

Source of Variation	Variance Component	% of Total
Repeatability	0.48	2.4
Batch to Batch	19.36	97.6
Total	19.84	100

Table 5-5: Components of variation for the scores of PCA model 3

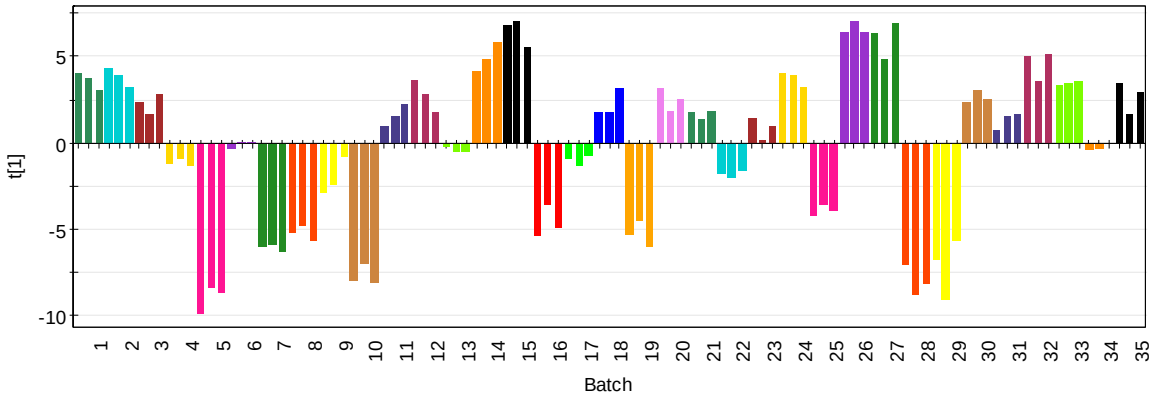


Figure 5-35: Repeatability of scores for PCA model 3, coloured by batch

5.3.8 Conclusions of Principal Component Analysis

An assessment of the particle size distribution of an API product has been undertaken, focusing on how the PSD may be affected by the manufacturing and milling processes. Three approaches have been considered for presenting the PSD data: (i) overlaying of the PSD curves, (ii) comparing the percentiles and (iii) performing principal component analysis. Overlaying the PSD profile provides the most detail about the distribution of each sample, but it is difficult to identify individual curves when there are a large number of samples. The percentiles of the distribution provide a summary of the

information contained within the profiles, but only a small amount of information is displayed on each graph. A PCA representation is able to capture a large amount of detail of the PSD data and display the results through scores and loadings plots. Two PCs have been found to explain up to 96% of the variation in the data. Therefore PCA can be used to summarise the data into as few data points as would be produced from calculating percentiles, but the majority of the information from the data is retained in the PCA representation.

Comparison of the principal component scores highlights the differences between samples and inspection of the loadings indicates the areas of the PSD where differences occur. However, it is always desirable to refer back to the original overlay plot to confirm the relationship between samples. Through the use of PCA, differences between samples can be quantified in terms of Hotellings' T^2 and the distance to model (DModX) statistics.

Two different laser diffraction instruments were used for making the PSD measurements and there appears to be a difference in the PSD results that were generated. In general batches analysed at Site B were measured to have a slightly higher proportion of coarse particles than those analysed at Site A. The difference could be caused by the measurement instruments using different algorithms to calculate the PSD from the measured data. In order to determine whether the differences are due to different instruments or different samples, samples from the same batches need to be tested on both instruments. These samples should be well blended and then separated to be tested on each instrument. If the differences are found to be caused by the PSD measuring instruments, it may be possible to work with the manufacturer to understand how the algorithms differ and cause a change in the calculated PSD. In future experimental work, if the PSDs of experimental batches are to be compared to the current PSD, only the dataset produced at Site B will be used to estimate the current PSD of the product.

From the initial analysis of the data collected at Site A it has been shown that there are differences between the unmilled materials manufactured on the two plants; the material from plant 2 has a much higher level of fine particles. The milling process is effective in both reducing the overall PSD of the product and increasing the similarity between the material from the two plants. However the milled material that was produced on plant 2 does have a slightly higher level of fine particles. It was also concluded that there is little difference between the milled material from the main and recovery processes, suggesting that it is the plant where the particles are formed that

affects the PSD, rather than the manufacturing process that is used before the precipitation stage.

The Hotelling's T^2 and DModX statistics both detected a significant difference between the material from the two processing plants. However all of the material was suitable for release, so using these metrics to detect a change to the PSD does not necessarily indicate that the material is unsuitable for release.

5.4 Batch to Batch Variation

Following the variation in the measurement system, the remainder of the variability in the PSD data is a result of differences between batches. This variation may be caused by factors upstream of the milling process, potentially during the precipitation and drying stages of the process, or by the milling process itself. Although the milling process has been shown to result in a large change in the PSD of the material, the particle size of the unmilled product will influence the final milled material. For example, the material produced on plant 2 has a higher proportion of fine particles and there remains a higher proportion of fines after milling.

Gaining further knowledge on how the manufacturing and milling processes affect the PSD of the milled product will allow AstraZeneca to increase its level of understanding of the process, thereby aligning with the process analytical technology framework (Section 2.2.3). If changes are to be made to the process, for example as a result of a continuous improvement project, it will be possible to determine whether the changes are likely to have an impact on the PSD of the product. A future opportunity of this work is to link the PSD to the properties of the product that are seen during the formulation process. If an optimal PSD can be determined that shows the most desirable characteristics during formulation, then the manufacturing and milling processes may be modified to produce a milled product with the desired PSD, allowing for improved performance during the formulation process.

Determination of the relationship between the manufacturing process and the final particle size distribution was based on data collected on a number of variables from both the manufacturing and milling processes (Section 5.4.1) and relating these to the PSD through the application of the modelling techniques partial least squares (Section 5.4.2) and artificial neural networks (Section 5.4.3).

Process data was collected for the initial set of batches that were analysed at Site A. Since differences have been identified between the PSDs of batches manufactured on

the different plants and from the recovery process, only batches from the main process on plant 1 are considered in this analysis.

5.4.1 Process Data

From the validation of the process, a number of critical and key process parameters have been identified that are expected to have the greatest influence on the quality of the product. These parameters are controlled to a level that has been found to ensure that the final product will be suitable for formulation and subsequent release.

5.4.1.1 Manufacturing Process Data

In total, 15 variables were collected from the manufacturing process (Table 5-6), starting from the formation of the solid particles during the precipitation stage. During the precipitation stage, calcium chloride ($CaCl_2$) solution is added to the batch, causing the product to precipitate and form an amorphous solid.

During precipitation it is essential that an amorphous rather than a crystalline solid is formed. Crystalline material is built up of a rigid structure and the crystals that form are hard solids. Conversely amorphous solids are much less structured, so the powder that forms is lighter and more easily soluble. The formulation process, where the powder is compressed into a tablet, is designed for the properties of amorphous material. Similarly the dose level of a tablet is selected for the dissolution rate of an amorphous solid and hence crystalline material would not produce the correct registered product.

The temperature of the batch during the precipitation reaction must be well controlled. If the temperature exceeds the upper specification limit the batch will begin to melt and form the wrong product. A batch temperature below the lower specification limit has been found to result in variation in the quality of the product.

The precipitation reaction happens rapidly, so it is important that the addition of calcium chloride happens continuously and is completed within a set time. It has been observed that a break in the addition of calcium chloride can cause the batch to crystallise and the material to thicken, causing blockages in the pipe that transfers the batch to the filter dryer. If the batch is allowed to crystallise then the material will be of the wrong polymorphic form and the correct API cannot be manufactured.

Following the precipitation stage the batch is transferred to a filter dryer. Three dryers are operated in parallel. During the drying process, warm nitrogen gas is passed through the product to drive off the excess water. The batch temperature must be controlled so that the material does not melt, particularly during the beginning of the

drying process when the water content is very high. The filter jacket and headspace temperature are monitored to ensure that the batch does not exceed the upper temperature limit for the drying process. When the drying process has finished the water content is measured with a loss on drying (LOD) test, to ensure that the batch is suitably dry. The drying times vary between batches and when the drying times for a particular filter become too long, the filter is cleaned.

Variable	Units
Water volume for calcium chloride	kg
Weight of calcium chloride	kg
Batch size at precipitation	L
Minimum temperature during precipitation	^{o}C
Maximum temperature during precipitation	^{o}C
Calcium chloride addition time	Mins
Minimum isolation temperature	^{o}C
Filter dryer number	1, 2, 3 or 4
Number of batches since filter clean	Count
Maximum drying temperature (first 12 hours)	^{o}C
Maximum drying temperature (remainder)	^{o}C
Maximum drying jacket temperature	^{o}C
Loss on drying result	%w/w
Drying time	Hours
Batch weight after drying	kg

Table 5-6: Manufacturing process variables

5.4.1.2 Milling Process Data

The details of the milling process were described in Section 5.3.2. The milling process will determine the final particle size distribution of the product, so the critical milling parameters could potentially show a relationship with the PSD. Five process variables have been identified as critical process parameters:

- Product weight throughput (kg/hr) – the rate of product being fed into the screw feeder
- Screw feeder speed (rpm) – the setting determined by the operator
- Nitrogen flow rate (m^3/hr) – the gas flow rate carrying the product into the mill
- Mill speed (rpm) – the rotational speed of the mill
- Product temperature (^{o}C) – inside of the mill

Data for each measurement is captured every 10 seconds throughout the milling process, with the average milling time being 12 hours. To assess the relationship with

the PSD, the milling process data must be summarised to a smaller number of variables that will represent the useful information captured in the data. Data was collected for each of the key milling parameters, for the 34 batches in the dataset. For each parameter, the data was captured as an average of every five minutes while the mill was running.

From each batch, two samples were collected from the powder being discharged from the mill, one near the beginning of the process and one close to the end. The two samples were blended before the PSD analysis was performed. It was therefore hypothesised that only the milling conditions towards the start and end of the process will have an effect on the PSD that is measured. The final milling data was reduced to the first 1.5 hours of the process for the initial sample and 4 hours towards the end of the process to capture the milling conditions of the second sample. More time points were included for the second sample because the material may be held up within the mill and hence the milling conditions from earlier in the process may have affected the PSD of the sampled product.

The milling variables were collected at regular intervals during the process, so it may be appropriate to unfold the data and apply multi-way PCA or PLS (Section 3.2.1). This approach assumes that the data describes the evolution of a batch over time and that each batch follows a particular trend. However graphs of the data from the milling process (Figure 5-36 to Figure 5-39) show that changes to the milling conditions can occur at any time and the measurements do not follow a trend over time. A high or low value of one of the measurements could affect the PSD of the product, irrespective of the time that it occurs. Therefore alternative summaries of the data were investigated.

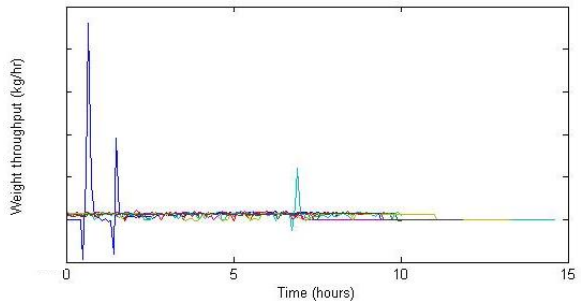


Figure 5-36: Example of product weight throughput

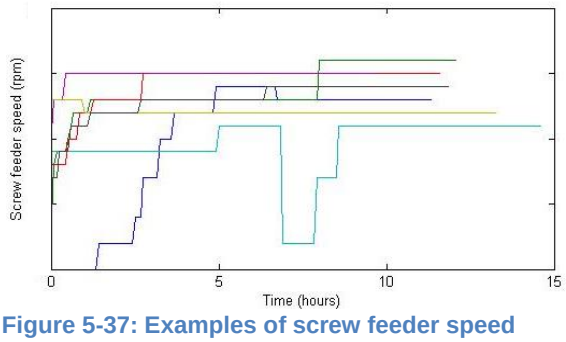


Figure 5-37: Examples of screw feeder speed

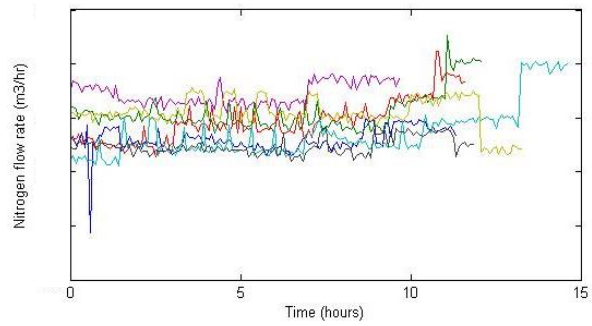


Figure 5-38: Examples of nitrogen flow rate

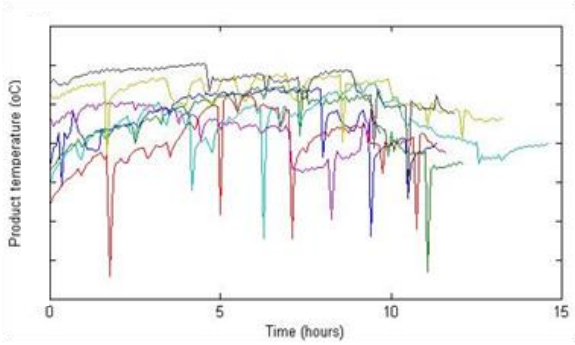


Figure 5-39: Examples of product temperature

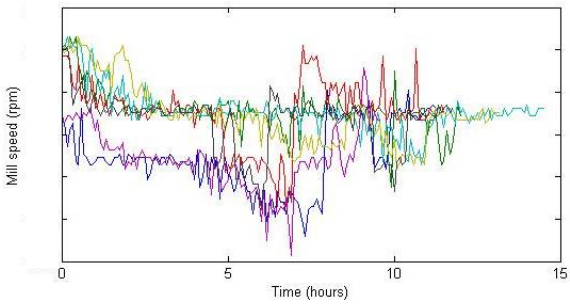


Figure 5-40: Examples of mill speed

The weight throughput of product is approximately constant throughout the duration of a batch, but spikes are observed when a large amount of material drops into the system at one time (Figure 5-36). Weight throughput measurements are also negative when the weigh scale reads incorrectly after a drop of product. The spikes in the data may indicate process conditions that have an effect on the PSD and hence the maximum throughput for each batch should be captured. To provide more detailed information, the 90th, 75th, 50th, 25th and 10th percentiles of the weight throughput data were determined, along with the mean and minimum. Data summaries were collected separately for the start and end portions of the milling process.

The screw feeder speed is generally constant throughout a batch, since the speed is set by the operator (Figure 5-37). The speed is seen to ramp up at the start of a batch and may be reduced if the weight throughput is too high. To summarise the screw feeder speed throughout a batch, the mean and mode will provide an indication of the average and most common speed that was applied during a batch.

The nitrogen flow rate and temperature measurements show some variation during the running of a batch, but differences in the profiles can be identified between batches (Figure 5-38 and Figure 5-39). Therefore the median flow rate and temperature may provide good summaries to compare batches. At times during milling a high spike on the flow rate or low dip in the temperature is observed. These points occur immediately after the mill is paused during milling, for example when a keg is replaced for product to be discharged into. At these times product is not fed into the mill, so the spikes in measurements are not expected to indicate a milling condition that may affect the PSD.

The mill speed is seen to vary minimally; with readings generally lying between 5710 and 5730 rpm (Figure 5-40). The median mill speed was used to capture the speed within each batch.

5.4.2 Assessing the relationship with PLS Models

The relationship between the process variables and the resulting particle size distribution was investigated by creating a PLS model, using the software SIMCA-P+ 12.0.1 (Umetrics AB, Umeå, Sweden). The input variables were the 15 manufacturing variables and 26 mill variables, and the response was the 23 PSD distribution measurements. Data was collected for 34 batches. The data was standardised to have a zero mean and unit variance, to prevent variables with large magnitudes having greater influence on the structure of the model.

5.4.2.1 PLS Model 1

The first PLS model contained all the process variables. Retaining one latent variable, the model was able to explain 54% of the variability in the particle size distribution (Table 5-7). However the Q^2 value is very low, suggesting that the PLS model may be over fitted as a consequence of the large number of input variables.

Number of latent variables	R^2X (cumulative)	R^2Y (cumulative)	Q^2 (cumulative)
1	0.10	0.54	0.212
2	0.22	0.66	-0.03

Table 5-7: Model fit of PLS model 1

The Hotelling's T^2 values (Figure 5-41) suggest that the first batch in the dataset could be an outlier. This is the same batch that was identified in Section 5.3.5.1 as displaying an unusual PSD profile with a high proportion of coarse particles. The distance to model value for this batch is not large, suggesting that the batch is drawing the model plane towards itself (Figure 5-42). The contribution plot for this batch shows that the high Hotelling's T^2 value is caused by variables relating to a high weight throughout and long drying time (Figure 5-43). Since there is only one batch in the dataset with a high weight throughput and a coarse PSD it is not possible to determine causality. This batch appears to be very different to the rest of the dataset and does not reflect the general behaviour of the process, hence it was removed from the model.

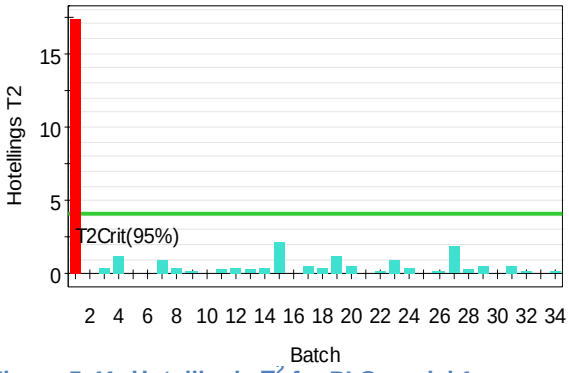


Figure 5-41: Hotelling's T^2 for PLS model 1

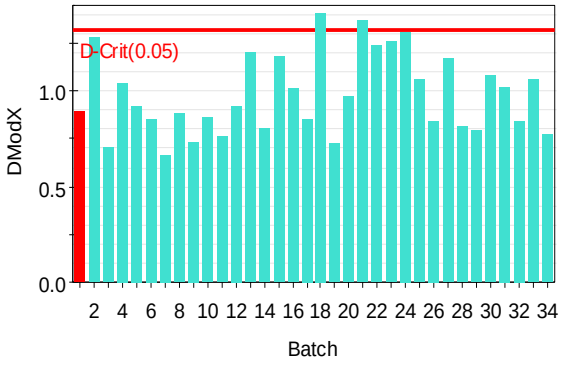


Figure 5-42: DModX for PLS model 1

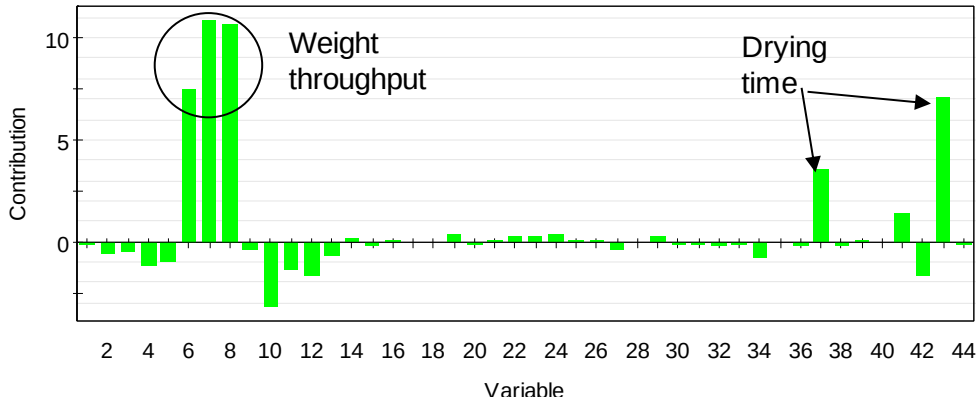


Figure 5-43: Contribution plot for batch 1

5.4.2.2 PLS Model 2

For the second PLS model, the first batch in the dataset was removed and the model refitted. This revised model was able to explain approximately half of the variability in the response data with one latent variable (Table 5-8). However less than 20% of the variation could be predicted when cross-validation was applied, suggesting that the model is again over fitted. More specifically, since there are a large number of input variables in the model, it may be possible to find a relationship between the input and response data, but this relationship does not hold when the model is applied to new data.

Number of latent variables	R^2X (cumulative)	R^2Y (cumulative)	Q^2 (cumulative)
1	0.07	0.56	0.17
2	0.24	0.59	-0.06

Table 5-8: Model fit for PLS model 2

The data set may contain a number of variables from the process that do not have an effect on the PSD and hence these variables will add noise to the model. SIMCA allows the importance of each variable to be compared through a VIP plot. The VIP for each

variable is the sum of the squared loadings, weighted by the amount of variation in the response that is explained by each latent variable. Eriksson *et al* (2006) suggested that variables with a VIP value greater than one are the most important for explaining the response variables.

Figure 5-44 and Table 5-9 show that a number of variables have a low VIP in the PLS model and can therefore be removed from the dataset. In particular, variables taken from the drying and precipitation processes appear to have a greater influence on the PSD than the variables measured during milling. By removing the variables with lower importance, it may be possible to find a PLS model that is less over fitted.

5.4.2.3 PLS model 3

Using the information from the VIP plot, variables that were not expected to have a relationship with the PSD were removed. The remaining variables are summarised in Table 5-9. The resulting PLS model shows slightly improved predictability from the previous model, with a Q^2 value of 0.36 for one latent variable (Table 5-10). However, since less than half of the variation can be predicted in new data, the model fit is still too low to have confidence of a strong relationship between the process variables and the PSD.

Figure 5-45 shows the loadings for the variables in PLS model 3, for the first latent variable. The largest loadings for the predictor variables relate to the drying process. The drying time has a negative loading, along with the maximum dryer and jacket temperatures, which will be correlated since batches that are dried for longer reach higher temperatures. Furthermore the loss on drying result, which indicates the batch dryness, has a positive loading because batches that dry most quickly have greater LOD results. There is a negative correlation between the drying time and the scores from the first latent variable (Figure 5-46), which suggests that batches with a higher proportion of large particles have longer drying times. However it is not possible to determine whether batches with long drying times form larger particles, or larger particles cause batches to have longer drying times.

Precipitation	
Variable	**VIP**
Purified water for CaCl2 solution	0.05
Weight of calcium chloride	0.05
Precipitation batch size	1.03
Min temperature during precipitation	1.13
Max precipitation temperature	0.54
Calcium chloride addition time	1.16
Min isolation temperature	1.78
Drying	
Variable	**VIP**
Dryer 1	1.72
Dryer 2	0.45
Dryer 3	0.19
Dryer 4	1.68
Number of batches since filter clean	0.01
Max drying temp (1st 12 hours)	0.42
Max drying temperature	0.93
Max drying jacket temperature	1.78
Loss on drying result	2.84
Drying Time	2.94
Batch weight after drying	1.03

Milling start	
Variable	**VIP**
Max weight throughput	0.53
P90 weight throughput	0.96
P75 weight throughput	0.64
Median weight throughput	0.05
P25 weight throughput	0.15
P10 weight throughput	0.09
Min Weight throughput	1.12
Mean Weight throughput	0.96
Mean screw feed speed	0.16
Mode screw feed speed	0.38
Median N2 flow rate	0.50
Median mill speed	0.03
Median product temperature	0.02
Milling End	
Variable	**VIP**
Max weight throughput	0.89
P90 weight throughput	0.86
P75 weight throughput	0.60
Median weight throughput	0.92
P25 weight throughput	0.33
P10 weight throughput	0.44
Min weight throughput	0.40
Mean Weight throughput	0.71
Mean screw feed speed	0.08
Mode screw feed speed	0.23
Median N2 flow rate	0.45
Median product temperature	0.68
Median mill speed	0.48

Table 5-9: VIP values from PLS model 2, variables to be removed in grey

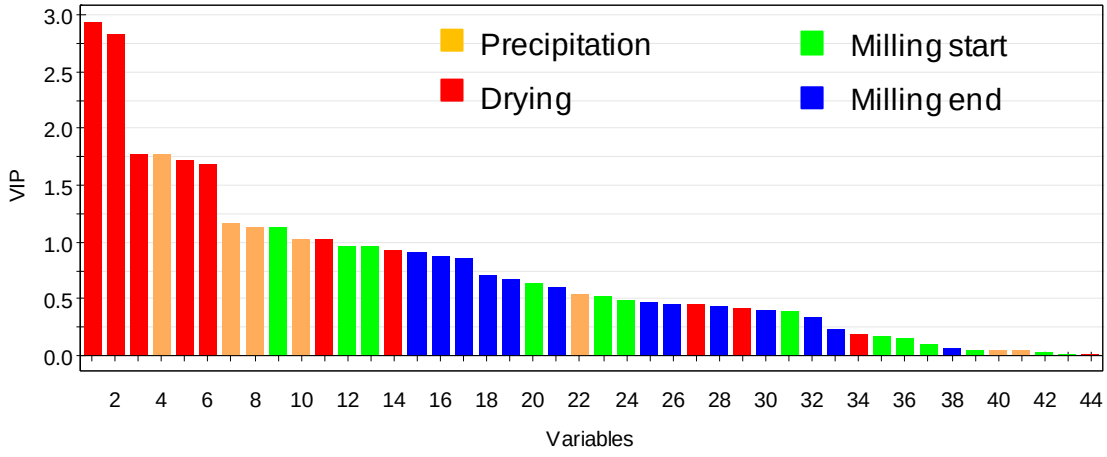


Figure 5-44: VIP plot for PLS model 2, coloured by process stage

Number of latent variables	R^2X (cumulative)	R^2Y (cumulative)	Q^2 (cumulative)
1	0.13	0.56	0.36
2	0.28	0.61	0.30

Table 5-10: Model for PLS model 3

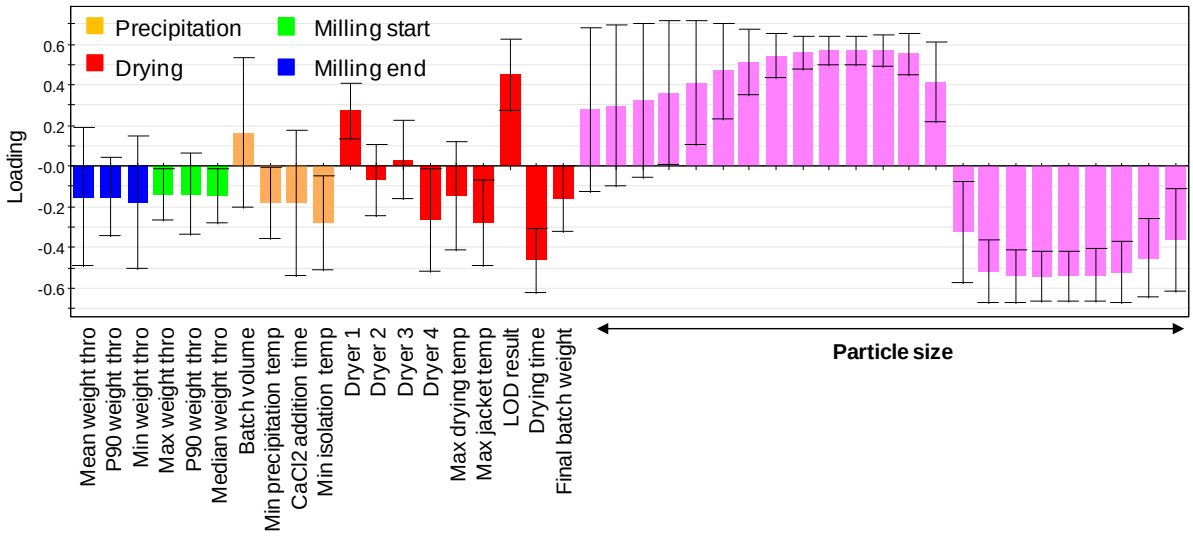


Figure 5-45: Loadings for PLS model 3, latent variable 1

The minimum temperatures during and after the precipitation reaction also have large loadings in the PLS model. However the variation in the temperature measurements is small, around 1°C and there is no evidence of a relationship with the PSD (Figure 5-47 and Figure 5-48). The precipitation time ($CaCl_2$ addition time) also has a large loading, but further investigation of the data shows that the trend is caused by one batch with a particularly long precipitation time (Figure 5-49). The mill weight throughput measurements at the end of milling have the largest loadings of the mill variables, however no strong relationship is seen with the particle size (Figure 5-50).

Assessment of whether non-linear relationships exists between the process variables and the PSD was performed by examining the scores for the input (t) and response (u) variables (Figure 5-51). The straight line pattern indicates that there is a linear relationship between the inputs and outputs of the model, suggesting that a linear PLS model is suitable. However, since no strong relationships have been identified, it may be beneficial to investigate any potential non-linear relationships in the data (Section 5.4.3).

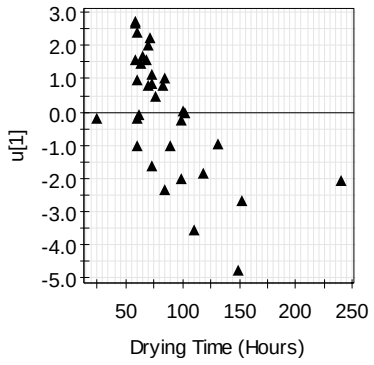


Figure 5-46: LV1 (response) vs. drying time

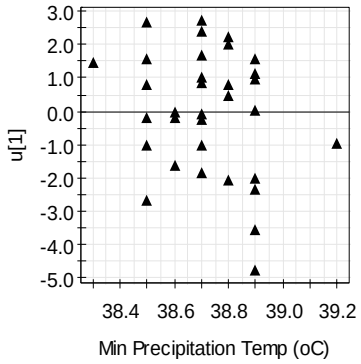


Figure 5-47: LV1 (response) vs min precipitation temperature

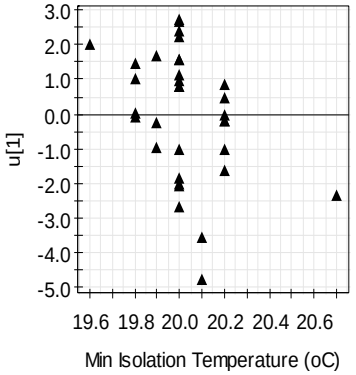


Figure 5-48: LV1 (response) vs min isolation temperature

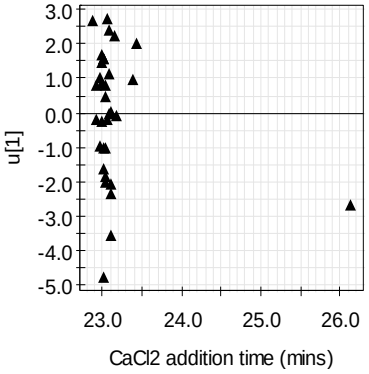


Figure 5-49: LV1 (response) vs CaCl2 addition time

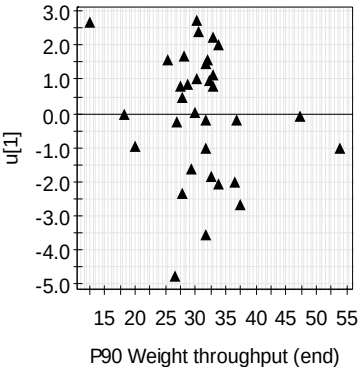


Figure 5-50: LV1 (response) vs P90 weight throughput

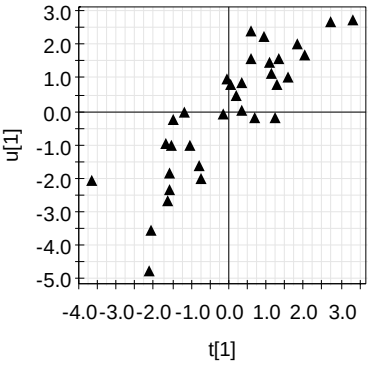


Figure 5-51: Response vs input scores

5.4.3 Assessing the Relationship with Stacked Neural Network Models

Stacked neural network models were developed to investigate whether a non-linear relationship exists between the process variables and the particle size distribution. Principal component analysis was applied to reduce the PSD data to one set of PC scores that represented 82% of the variation in the PSD data. These scores were used as the response variable.

In Section 5.4.2, 16 process variables were identified as having the strongest relationships with the PSD. The variable “dryer” was removed since it is a categorical variable and numerical inputs were required. Additionally, the maximum and P90 weight throughput (end) variables are highly correlated with a correlation coefficient of 0.79, so only the maximum was included in the model.

Cross-validation was used to determine the structure of the stacked neural network model. The batches were divided into six groups of five or six and neural networks were repeatedly fitted using the cross-validation group as the test data set. The training data comprised of 19 batches, with the remaining eight or nine as validation samples. Using stacked neural networks consisting of 30 individual networks and retaining three principal components, one hidden node was found to be optimal at providing the smallest mean squared error for the test data set (Figure 5-52).

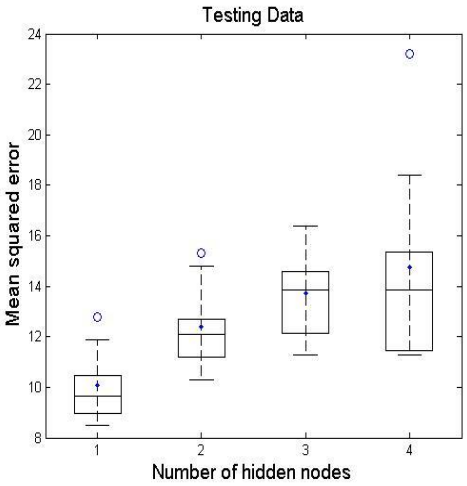


Figure 5-52: MSEs for 30 stacked neural networks

To determine which variables have the strongest relationship with the PSD, each variable was removed individually and a stacked neural network model was created. A low MSE when a variable was removed indicated that the particular variable does not add any information to the model and can therefore be excluded from the analysis. Conversely, an increase in MSE when a term is removed indicates that the variable should be included to minimise the model error.

For computational efficiency, all of the batches were used as either training or validation batches and were randomly split into each group with a ratio of 70:30. Three repeated stacked networks were fitted for each variable that was removed and an average taken of the resulting mean squared errors (MSEs). With no variables removed from the dataset, the MSE was 2.9. Five variables were identified that resulted in a particularly low MSE when they were removed from the dataset (Figure 5-53): Minimum and P90 of the weight throughput (start of milling), batch volume at precipitation, minimum temperature during precipitation and maximum temperature during drying.

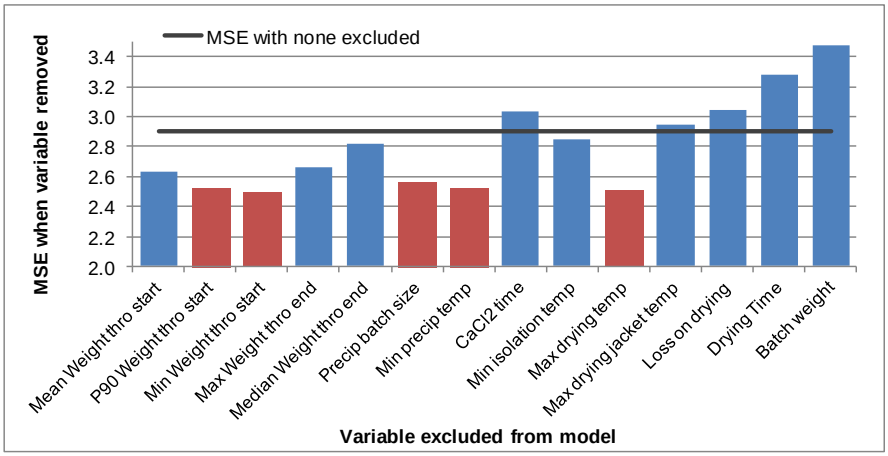


Figure 5-53: Mean squared errors with variables removed from datasets, highlighting five variables to remove from the data set

The stacked neural network model was refitted with these five variables removed, resulting in the mean squared error being reducing to 2.5. As before, each variable was removed individually and a further model fitted. An increase in MSE was seen when any of the remaining variables were removed from the dataset (Figure 5-54), suggesting that all of the remaining terms are useful for predicting the PSD.

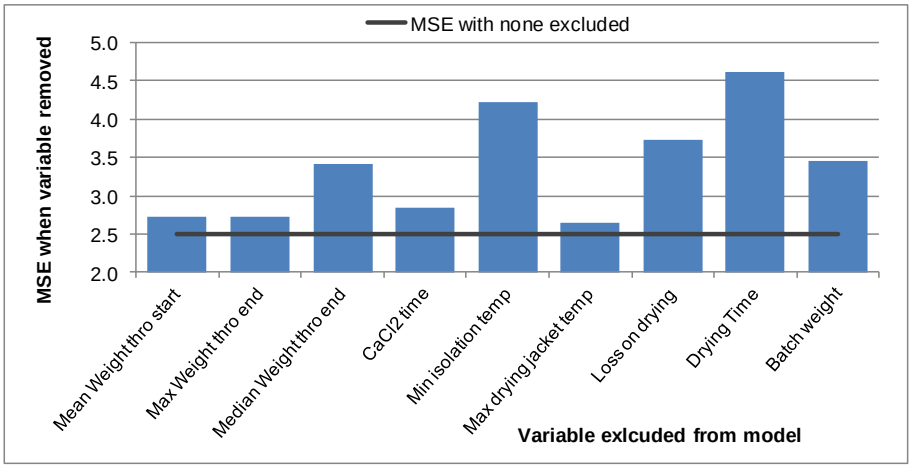


Figure 5-54: Mean squared errors with variables removed from dataset

The largest increases in error were seen when variables relating to the drying times and the precipitation temperature were removed, suggesting that these have the strongest relationship with the PSD and agreeing with the results from the PLS analysis (Section 5.4.2.3). When cross-validation was applied to the stacked NN model the R^2 for prediction was 61%, suggesting that the neural network model achieves a greater level of fit than the PLS model, which had a Q^2 of 36%. However the results from the neural network model do not explain how the variables are related to the particle size, other than to indicate which may be the most important variables. A greater scientific understanding of the process would be required to be certain of the relationships between the variables.

5.4.4 Discussion of Modelling Results

The poor model fit for the three PLS models suggests that there may not be a strong linear relationship between the process variables and the particle size distribution, for the range of each variable that was included in the dataset. The neural network model was able to identify a better fit but, with the exception of the drying time, individual variables do not show a relationship with PSD (Figure 5-46 to Figure 5-49).

While these results do not provide information in terms of explaining the batch to batch variation of the PSD, they do suggest that any variables that could have a significant effect on the PSD have been identified and are suitably controlled to a level that does not have a significant effect. A number of process variables, including the temperatures during precipitation and drying, are controlled to a level that has been found not to affect the quality of the product and hence have no impact on the particle size distribution. Similarly the screw feeder and mill speeds exhibit limited variation between batches, so the range of data is too small to identify the presence of a relationship with the particle size.

For a more detailed understanding of the relationship between the process variables and the particle size distribution, it may be necessary to increase the range of each variable by running the process outside of the normal operating range. However this approach may result in material being produced that cannot be released and is therefore not feasible.

A further limitation to the modelling is the sampling plan for milled material. Samples are taken from the start and end of each batch and blended together, which does not allow for changes in the PSD from the start to end of milling to be measured. This approach only allows differences in milling conditions to be compared batch-to-batch and not within a batch. Taking more samples throughout a batch and analysing them separately would allow an assessment of how changes to the milling conditions during a batch can affect the PSD.

5.5 Conclusions

A number of particle size distribution measurement techniques were initially discussed and of these laser diffraction and scanning electron microscopy were used to measure the PSD of an active pharmaceutical ingredient. The resulting data were analysed by principal component analysis, partial least squares regression and neural network modelling.

Laser diffraction (LD) was shown to produce detailed and repeatable measurements of the PSD for a number of batches. The ability of LD to produce results rapidly allowed for the comparison of material from a number of batches produced from different plants and processes. The difference in the material from the two plants was confirmed by examining the scanning electron microscopy images.

Principal component analysis (PCA) was applied to the data that was generated from the laser diffraction analysis. The PCA method was captured 99% of the variation in the data with three components and highlighted differences between the two processing plants and identified batches that exhibited unusual behaviour. By reducing the PSD to a set of principal components, batches can be compared on a scores scatter plot, as opposed to overlaying each distribution. However batches that that show a different PSD to the rest of the data set may not show differences in the formulation stage of the process and further study is required to understand the relationship between PSD and the characteristics of the product formulation.

Samples were analysed at two sites, using two different laser diffraction instruments. A comparison of the LD results showed an offset between the data generated from each site, however the same differences between the two processing plants were observed with both instruments. The results indicated that samples should only be compared when they have been analysed by the same instrument.

The PLS algorithm was unable to identify a strong linear relationship between the process data and the particle size distribution. The stacked neural network model identified that the drying time may have the strongest relationship with the PSD. The lack of model fit may be due to the range of each variable may be too small to show a relationship; for example some variables are controlled to a level of +/- 1°C. The most important variables in the process that may affect the particle size have previously been identified and are tightly controlled. As a result data from the normal operating range of the process may not show enough variation to identify any relationships with the PSD.

This study into particle size distribution has suggested that the key parameters of the drying and milling process are well controlled within their process specification limits. Process capability indices are used to capture how well a process runs within its specification limits and are investigated in Chapter Six, with a focus given to data that does not satisfy a normal distribution.

6 Process Capability Indices for Non-Normal Data

6.1 Introduction

A process with a measurable customer requirement will typically have a set of specification limits to which the data from the process or product must conform. How well the data lies within the specification limits shows how capable the process is of meeting the customer's target and can be quantified through the calculation of process capability indices (Spiring, 1995).

Pharmaceutical manufacturing processes have a number of specification limits that are registered with regulatory authorities. Data from the process must show that these limits are met to allow the finished product to be released to the market. Specification limits may be for processing conditions, such as reaction temperatures, timings and quantities of reactants or for quality control testing of the final product. For a batch process, if a measurement lies outside of a particular limit, then the batch will fail and be rejected.

In-process specification limits are used to confirm that a batch has been run under conditions that have been shown to produce a quality product and prevent impurities from forming. The limits are determined from knowledge gained during the development stage of the process and are registered with the relevant regulatory authorities. The monitoring of in-process measurements should confirm that the final product has been manufactured to the required quality, before being verified by quality control testing. This approach aligns with the US Food and Drug Administration's (FDA) principles for process validation, where quality assurance through process control is encouraged over end product testing, since quality can never be fully measured by testing (FDA, 2011).

The risk of a specification limit being breached can be quantified using process capability indices. The capability of a process describes how well the measured data sits within the specification limits and therefore shows the risk of a batch failure. Capability is a measure of the ability of the process to meet customer requirements (Spiring, 1995), with a highly capable process having a very low risk of failure. Process capability indices (PCIs) directly relate to the probability of a batch failing, so from the capability, the expected long term cost of batch failures can be calculated.

Most processes will have a number of specification limits for different measurements and stages of the process. A separate PCI is calculated for each registered range and the process is at most only as capable as the worst performing part of the process. By calculating a set of PCIs, the greatest risks to the process can be identified and prioritised for improvements to be made. Multivariate process capability techniques can also be used to handle multiple sets of specification limits (Section 6.1.3.3). By calculating the expected costs resulting from batch failures, the cost of improvement work can be justified.

Many of the standard process capability indices are based on the assumption that the data satisfies a normal distribution. However, data from industrial processes may not satisfy normality due to the manner in which the process is operated and controlled; hence the results from standard process capability indices may not reflect the actual capability of the process.

This chapter provides an introduction to the most commonly used process capability indices and how they should be interpreted and implemented. A number of non-parametric indices have been proposed in the literature, based either on the percentiles of the data or the proportion of data outside of the specification limits, which are discussed in Section 6.2. In Section 6.3 a simulation study of these metrics is conducted to understand their performance on data sampled from different statistical distributions with varying levels of skewness. Since the underlying distributions are known, the true capability of the population from which the data is sampled is also known. The results are then analysed to determine how close the calculated capabilities are to those of the underlying distribution, and to determine the level of variation caused by differences between samples taken from the same underlying distribution. In Section 6.4, these non-parametric indices are then applied to data from a manufacturing process operated by AstraZeneca.

6.1.1 Process Capability Indices for Normal Data

A number of process capability indices are regularly used by industry, which show both the current performance of the process (Section 6.1.1.1), and its potential capability if process improvements were introduced (Section 6.1.1.3).

6.1.1.1 P_p and P_{pk}

Two commonly used measures to quantify the process capability are P_p and P_{pk} (Kane, 1986, Kotz and Johnson, 1993). P_p measures how the width of the variation in the process data compared to the width of the specification limits (Figure 6-1):

$$P_p = \frac{USL - LSL}{6\sigma}$$ Equation 6-1

where USL is the upper specification limit, LSL is the lower specification limit, and σ is the process standard deviation.

To estimate the P_p value from a sample of data, $\hat{P}_p$ is calculated from the sample standard deviation, s:

$$\hat{P}_p = \frac{USL - LSL}{6s}$$ Equation 6-2

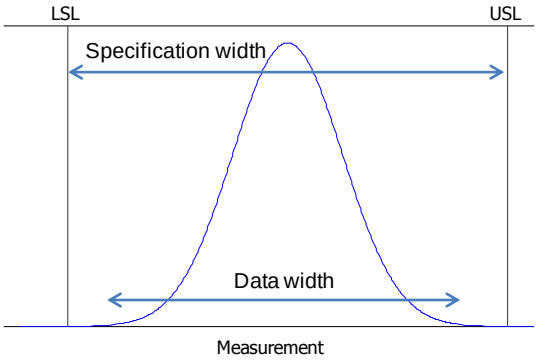


Figure 6-1: Example of a distribution and the measurements used to calculate P_p

Although the P_p measure does not indicate how much of the data sits within the specification limits, it shows the potential capability if the process measurement could be centred within the specification range without changing the level of variation. For a normal distribution, 99.73% of the data is expected to lie within the width of six standard deviations of the measurement. Consequently if six standard deviations is less than the specification width, the process has the potential to be capable, otherwise some data would be expected to lie outside of the limits.

When the process is not centred within the specification range, a capable process is required to have a minimum of three standard deviations between the mean and each limit. The P_{pk} value measures the shortest distance from the mean of the process to the specification limits, in units of three standard deviations:

$$P_{pk} = \min\left\{\frac{\mu - LSL}{3\sigma}, \frac{USL - \mu}{3\sigma}\right\}$$ Equation 6-3

where, USL is the upper specification limit, LSL is the lower specification limits, μ is the process mean and σ is the process standard deviation.

To estimate the P_{pk} value from a sample of data, $\hat{P}_{pk}$ is calculated from the sample mean, $\bar{x}$, and the sample standard deviation, s:

$$\hat{P}_{pk} = \min\left\{\frac{\bar{x} - LSL}{3s}, \frac{USL - \bar{x}}{3s}\right\}$$ Equation 6-4

The P_{pk} metric shows the actual capability of the process that is being run. If the process is centred then P_p is equal to P_{pk}, otherwise P_p is greater than P_{pk}. Hence P_{pk} is the actual capability and P_p is the potential that could be achieved if the process is centred.

6.1.1.2 Interpretation of P_{pk} Values

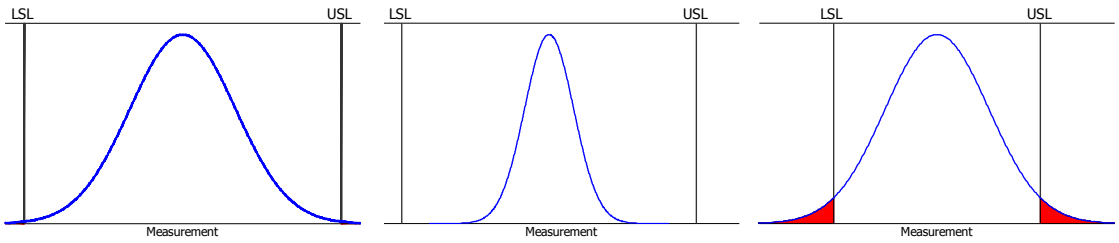
From the calculated P_{pk} value, the shape of the normal distribution can be used to calculate the expected failure rate of a process. Figure 6-2 shows examples of processes with various P_{pk} values, and the amount of data which fall outside of the limits.

Typically a process with a P_{pk} greater than one is considered to be a capable process, with at least 99.73% of the measurements within the specification limits (Figure 6-2a,b). The target for a highly capable process is 1.33 (Anjard *et al*, 1991). When a process has a P_{pk} of less than one, there is a risk of batches failing a specification limit. For example, a process with a P_{pk} of 0.67 (Figure 6-2c) would have an expected failure rate of 4.5%. The expected failure rate is calculated as:

$$Proportion\ out\ of\ specification = \left(1 - \Phi\left(P_{pk} * 3\right)\right) * 2$$

here Φ is the standard normal distribution function.

When comparing P_{pk} values, it is assumed that the data satisfies the shape of a normal distribution. If this assumption does not hold then the expected failure rate that is calculated from the P_{pk} value is not valid, and the value of P_{pk} may not accurately indicate the capability of the process (Section 6.1.4). It is also assumed that the data are randomly distributed about the mean and there are no step changes or drift over time in the process.



a) $P_{pk} = 1$ b) $P_{pk} = 2$ c) $P_{pk} = 0.67$

Figure 6-2: Typical processes with P_{pk} of 1, 2 and 0.67 respectively

6.1.1.3 C_p and C_{pk}

When using P_p and P_{pk} to measure the capability of a process, it is assumed that the process is in a state of statistical control (Section 2.2), that is the process mean is constant over time and the variation is random noise centred about the mean. However for some processes, a step or gradual change in the mean may occur, known as shift and drift respectively, or special cause variability. For example, a shift in the mean could be caused by a change of raw material added into the process, whilst a drift may be a result of an instrument calibration that changes over time. When the mean changes over time, the actual variation around the short term mean will be smaller than the overall variation of the whole data set (Figure 6-3).

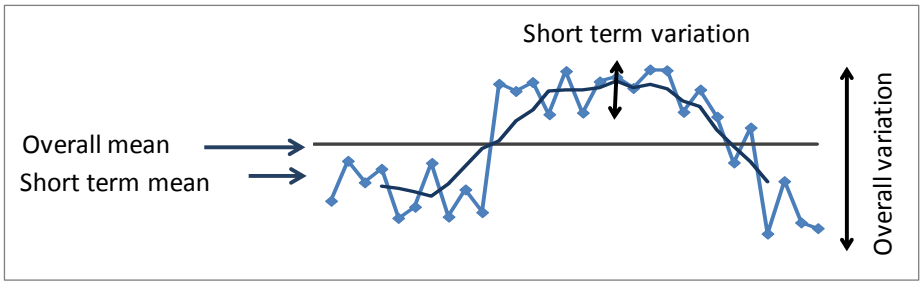


Figure 6-3: Example of a process where the mean shows shift and drift

An alternative capability measure is thus to calculate the short term variation of the process, excluding any shift or drift in the mean. The metrics C_p and C_{pk} use equations (Equation 6-2 and Equation 6-4), but the standard deviation, s, is replaced with the short term variation, which is measured from the absolute difference between consecutive data points. The values of C_p and C_{pk} will show the potential capability that could be achieved if the causes of shift and drift could be avoided, so the only variation is random noise about the overall mean. A large difference between C_{pk} and P_{pk} suggests that the mean is not constant and the capability of the process could be improved if the mean could be controlled and kept constant.

6.1.1.4 Comparison of P_p, P_{pk}, C_p, and C_{pk}

P_p, P_{pk}, C_p, and C_{pk} are four metrics which provide information about the current capability of the process and the level that could be achieved if improvements are made based on the information gained from comparing the four metrics (Figure 6-4). P_{pk} quantifies the actual capability that is currently being achieved by the process. P_p is the potential capability achievable if the process is centred between the specification limits. C_{pk} is the potential if the process mean could be controlled by removing special cause variability. C_p is the overall potential capability if the process is centred and the

mean is constant. The size of the difference between the four values shows where the largest gains in capability could be achieved.

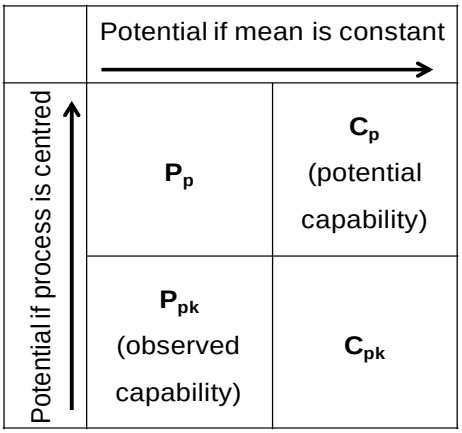


Figure 6-4: Information from differences between process capability metrics

6.1.2 Process Capability Analysis at Industrial Sponsor

At AstraZeneca, a review was undertaken to determine how to monitor the capability of a process with respect to in-process specification limits. Previously measurements from batches were checked individually against the specification limits, but little analysis was done to assess the overall capability of the process. Limited process capability analysis was conducted using P_{pk}, but without checking for the underlying assumptions of the method (Section 6.1.3.2).

The business required a system to quantify the capability of individual stages of a process, to highlight any risks to capability, indicate the need for improvements and to monitor the capability over time. A monthly capability review was initiated for in-process specification limits, where the P_{pk} values for each set of limits were displayed along with a 'red, amber, green' colour code to give a quick visual indication of the state of the process. A P_{pk} value less than one was coloured red, P_{pk} between one and 1.33 was coloured amber, and process with P_{pk} greater than 1.33 was labelled green. However it was noted that a number of the data sets did not conform to the assumption of a normal distribution, and hence process capabilities indices for non-normal data were investigated (Sections 6.1.4 to 6.3)

6.1.3 Implementation of Process Capability Indices

Process capability indices have been widely applied in a range of industrial sectors, including in the manufacturing and service industries (Spiring, 1995). Process capability can be applied to any situation with a measurable customer requirement, to quantify how will well the targets are being met.

6.1.3.1 Benefits

Process capability indices are dimensionless numbers and hence can be used to compare between measurements recorded in different units or magnitudes, allowing the capability to be compared across different stages of a process, or between different processes or locations (Anis, 2008). By comparing all stages of a process simultaneously, areas for improvement can be prioritised to focus on the lowest capability and targets can be set to achieve a minimum level of acceptable capability, for example a P_{pk} greater than 1.3. These comparisons are particularly appropriate for the manufacture of active pharmaceutical ingredients because there can be several stages and chemical processes, each with registered limits for the processing conditions. Process capability studies are also useful for setting performance targets, prioritising and implementing continuous improvement work and adopting a common language for process performance (Kane, 1986).

By comparing the values of C_{pk}, C_p, P_{pk} and P_p, the potential capability of a process can be found. The difference between the four metrics will indicate whether improvements should be targeted at special cause variation, common cause variation or by adjusting the mean of the process (Section 6.1.1.4). The use of structured improvement methodologies such as Six Sigma (Section 2.4.2) can assist in delivering an improvement to the capability (Yu, 2008). The benefits of the improvement work can be demonstrated by trending P_{pk} values over time, allowing changes in capability to be monitored (Kane, 1986).

The use of process capability also fits into the framework for Quality by Design (QbD, Section 2.3). In the implementation of QbD, the critical quality attributes (CQAs) and critical process parameters (CPPs) are identified, so the goal is to develop the process such that all the CQAs and CPPs all have high capability (Yu, 2007). Seibert *et al* (2008) suggested that the capability of a process parameter may be used to help determine whether it should be labelled as critical. For example when an important parameter is shown to have very high capability, it may not be necessary to treat the parameter as critical, since the risk of variation of the parameter impacting on the CQAs is very low. However it may still be necessary to periodically monitor these parameters to ensure that the capability remains high.

An example of applying capability metrics to the development of a cell tissue engineering process is reported in Liu *et al* (2010). P_p and P_{pk} were used to compare a manual and an automated method for the culture and expansion of human cells. Both approaches had very low P_{pk} values, less than 0.3. However the automated method

had a P_p value of 1.3, suggesting that high capability could be achieved with the automated method if the process could be centred within the specification limits, leading to an improved automated method with higher capability.

6.1.3.2 Challenges

Before undertaking a process capability study, the data must be checked to ensure that it satisfies the assumptions of normality, stability and independence. If the assumptions are not met then the calculated indices may not reflect the true capability of the process, leading to improvement work and resources being targeted at the wrong part of the process (Anis, 2008). When calculating process capability metrics, it is important to use data that represents all of the variation that is normally seen in the process, so that the results are a true reflection of the capability of the process (Deleryd, 1998).

The process must be run in a stable way so that the mean remains constant over time. If the process mean changes, the capability will only describe the current process and not indicate what is expected in the future (Deleryd, 1999, Palmer and Tsui, 1999). Additionally, measurements should be statistically independent, so there is no autocorrelation between consecutive data points (Porter and Oakland, 1991). The assumption of independence may be more difficult to achieve for a continuous process, since consecutive measurements may be expected to be similar. However for batch process when there is one measurement for each batch, each measurement may be expected to be independent of others.

Deleryd (1999) surveyed a number of companies who use process capability, including several from the manufacturing industry. It was found that the greatest benefits were gained from an increase in process knowledge and the ability to make fact based decisions for improvement work. However Deleryd (1999) found that the biggest drawback to process capability studies was the resources required both for training in how to run process capability studies and for carrying out improvement work.

6.1.3.3 Multivariate Process Capability Indices

When one process has a number of specification limits, it may be beneficial to know the overall risk that a failure will occur, for example to compare between different processes. The risk can be calculated using multivariate process capability metrics. Additionally, when the process data is multivariate in nature, the specification limits can define a region in multivariate space, rather than individual limits for each measurement.

The set of specification limits can be thought of as a multivariate tolerance zone, such that if the data points are all within the limits then they are within the tolerance zone. The zone may be rectangular for individual limits or ellipsoidal for multivariate limits. The process capability is calculated from the risk of a result occurring outside of the tolerance zone (Chen, 1994, Zahid and Sultana, 2008), where the capability index is the ratio of the acceptable risk over the actual risk. The capability can also be thought of as the ratio of the radius of the tolerance zone to the radius of the zone required to have the acceptable level of samples within the limits, which is a multivariate equivalent of the definition of P_p. Chen's (1994) approach makes the assumption that the data follows a multivariate normal distribution.

An alternative non-parametric approach is to calculate the probability of a failure occurring, based on the failure rate of a sample of measurements (Polansky, 2001). This method looks at the number of samples that have seen a failure, but not how many specification limits were breached for a particular sample, or how close the data points are to the limits.

6.1.4 Non-Normal Data

The standard PCIs, such as P_{pk}, are based on the assumption that the data are normally distributed. Therefore the data is assumed to be symmetrical about the mean and the range is represented by six standard deviations. However for an industrial process, several sources of variation can result in the data not satisfying normality (Porter and Oakland, 1991). For example, variation in batches of raw materials can impact on the product, ambient conditions such as temperature and humidity may affect the control of the plant, and different operators may run the plant in different ways, all of which can result in data with a shape that is different to the normal distribution. Many sources of data may not be expected to follow a normal distribution, such as particle size distribution, pH and chemical impurity content measurements (Anis, 2008).

In some cases, non-normality of the data may suggest that the process is not stable or in statistical control, for example if the process mean varies over time. However if the reasons for the non-normality can be explained and the process data is expected to remain within the current range, then it is appropriate to estimate the capability of the process.

Examples of non-normal industrial process data are shown in Figure 6-5. For these processes assuming a normal distribution and calculating P_{pk} may not accurately quantify the true capability. Figure 6-5a shows the minimum speed of the mill recorded

for 89 batches, the data is highly skewed because the mill generally runs at a constant speed, but may be reduced at the start and end of a run. Figure 6-5b shows the recorded temperature of a reaction; the distribution appears to be bimodal, which could be caused by different operating conditions such as the reaction being run during the day or night. Figure 6-5c shows the gas recirculation flow rate in the mill. It is desirable to run the flow rate close to the lower limit but the flow rate is controlled so that it does not fall below the lower limit. For each data set, the causes of the trends within the data are being investigated by the relevant technical teams.

There are several options for handling non-normal data to allow the process capability to be found, including the removal of outliers, transforming the data, fitting an alternative probability distribution or using a distribution free PCI.

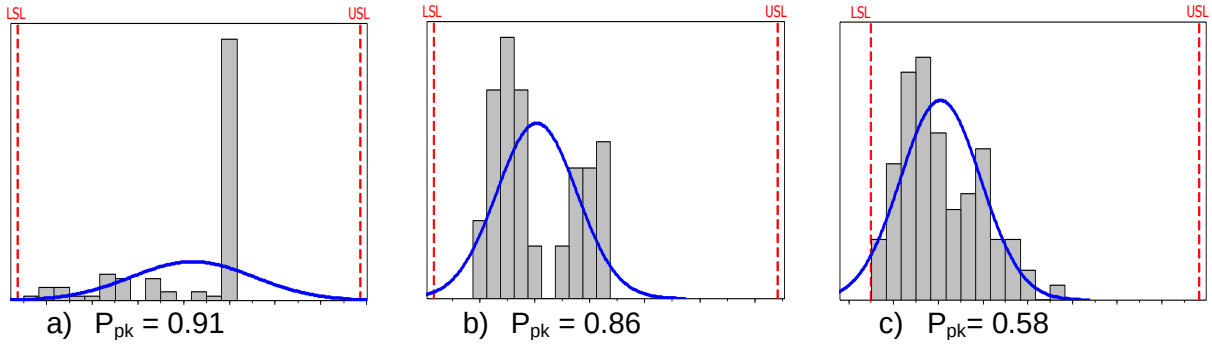


Figure 6-5: Examples of non-normal process data, a) mill speed, b) reaction temperature, c) gas flow rate

Removing extreme data points can improve the fit of the data to a normal distribution. However if the extreme points are part of the overall distribution of the data, then removing them from the data set will remove their influence on the calculated capability, making the process appear more capable than it is in practice. Outliers should only be removed from the dataset if a known error has occurred in collecting the data, such as a laboratory error occurring during quality control testing.

6.1.4.1 Data Transformations

Data can be transformed to increase its similarity to a normal distribution. The process capability can then be calculated on the transformed data. For example if the data is positively skewed then a log transformation can be applied to reduce the skew of the data. The Box-Cox transformation (Box and Cox, 1964) is commonly used to find the most suitable method to transform the data to a normal distribution. For this approach, a value of λ is calculated to transform each data point, y, to $y^{(\lambda)}$:

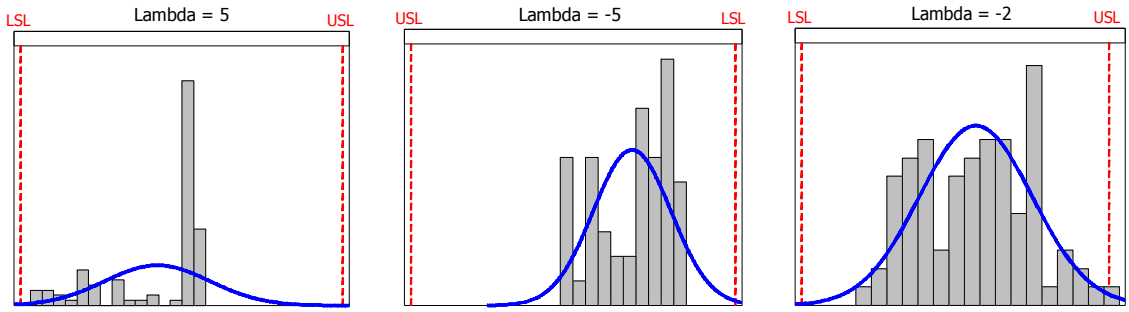
$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Equation 6-5

The value of λ is that which maximises the log likelihood function of λ, when it is assumed that the transformed data $y^{(\lambda)}$ follows a normal distribution. Therefore λ allows the data to be transformed to be as close as possible to a normal distribution.

When a transformation is applied to the data the specification limits must also be transformed. However transforming the data and the limits will lose the scale of the data, making the interpretation of the results more difficult. Additionally this method relies on finding a suitable transformation that results in data that follows a normal distribution.

Figure 6-6 shows the Box-Cox transformation of the data shown from Figure 6-5, with the P_{pk} values of the transformed data. Although the optimal values of λ are found, the distributions still do not appear to be close enough to a normal for the standard P_{pk} metric to be applied. The P_{pk} values for the transformed data are generally similar to those for the original data.



a) P_{pk} = 0.86 b) P_{pk} = 0.88 c) P_{pk} = 0.78

Figure 6-6: Box-Cox transformations of process data

6.1.4.2 Fitting an Alternative Distribution

When a normal distribution is not an appropriate fit to the data, it may be possible to find an alternative distribution which can produce a better fit. To determine which to use, several choices of distributions can be considered, for example the gamma or beta distributions. From the chosen distribution, the 0.135 and 99.865 percentiles are found to represent the range of the data and replace the six standard deviations in the process capability calculation. Additionally the mean is replaced by the median of the data. This approach is known as the Clements method (Tang and Than, 1999, Ahmad *et al*, 2008).

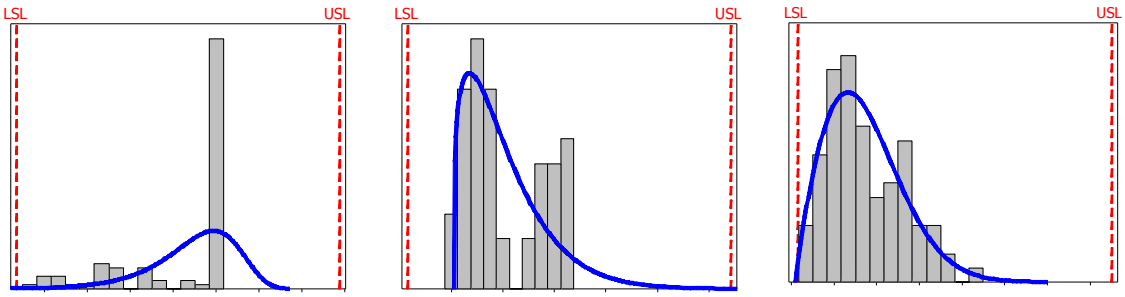
Rather than finding the optimal distribution for a data set, the percentiles of a data set can be estimated from the calculated skewness and kurtosis. Kotz and Johnson (1993) presented a table of values for the 0.135 and 99.865 percentiles of standardised data, for given skewness and kurtosis values. However Ahmad *et al* (2008) and Tang and Than (1999) both noted that estimates of skewness and kurtosis can be unreliable for small sample sizes.

Tang and Than (1999) applied the Box-Cox and Clements methods to data simulated from Weibull and lognormal distributions, to determine how well the capability calculated from these samples compared with the capability of the underlying distributions. Ahmad *et al* (2008) conducted a similar study, based on the Weibull and lognormal distributions.

Tang and Than (1999) found that with a lognormal (0, 0.5) distribution with a skewness coefficient of approximately 1.9, the Box-Cox method produced more accurate and less variable results than the Clements' method, for a sample size greater than 100. However with a lognormal (0, 1) distribution, with a skewness coefficient of around 5, Ahmad *et al* (2008) found that the capability was underestimated when the Box-Cox method was applied to the data. These results suggest that transforming the data may be reliable when the level of skewness is not too strong. The Clements' method was shown to overestimate the process capability when the data was highly skewed, but underestimate it when the skew was smaller.

For each set of process data shown in Figure 6-5, a Weibull distribution was found to produce the closest fit (Figure 6-7). The first set of process data is highly skewed (skewness=-1.2) but does not follow a smooth curve, so the shape of the data is difficult to represent with a statistical distribution. The second data set appears to be bimodal, which also cannot be represented by any standard statistical distribution, therefore the capability cannot be estimated from the Clements method.

A suitable fit could only be found for the third data set, the mill gas flow rate. For the gas flow rate, the estimated P_{pk} value of 0.99 appears to be appropriate since the data is close to the lower limit, but is all within the specification range.



a) P_{pk} = 0.86 b) P_{pk} = 1.16 c) P_{pk} = 0.99

Figure 6-7: Weibull distribution fitted to process data

Ideally a process capability index is required that does not rely on the data fitting a specific distribution or the application of a transformation, since a suitable fit cannot be guaranteed. Therefore distribution free process capability indices have been investigated

6.2 Distribution Free Capability Indices

A number of alternative process capability indices have been proposed in the literature which do not make any assumptions about the shape of the data. Indices have been found based either on the percentiles of the data, or the proportion of data that is outside of the specification limits.

6.2.1 Capability Indices Using Percentiles

In general, the basis for estimating the process capability using the P_p metric is to calculate the ratio of the width of the specification range to the width of the data. The P_{pk} measure extends this concept to looking at each side of the mean separately. The P_p and P_{pk} measures rely on the assumption that the data are centred about the mean and the width of six standard deviations represents 99.37% of the data. Consequently the objective of the P_p and P_{pk} measures is to assess how well 99.73% of the data fits within the specification limits.

For data that does not follow a normal distribution, equivalent measures are required for the centre and dispersion of the data. One option is to use the median rather than the mean, since the median is less influenced by skewness and the presence of outliers. The width of the data can be measured by percentiles, which will show the range of a specified proportion of the data.

Chen and Pearn (1997) proposed a set of process capability indices, denoted C_{Np} and C_{Npk}, which are based on the median, M, and the 0.135% and 99.985% percentiles of

the data ($P_{0.135}$ and $P_{99.865}$). These percentiles are equivalent to the mean ± 3 standard deviation for a normal distribution:

$$C_{Np} = \frac{USL - LSL}{P_{99.865} - P_{0.135}}$$ Equation 6-6

$$C_{Npk} = min\left\{\frac{USL - M}{1/2\,(P_{99.865} - P_{0.135})}, \frac{M - LSL}{1/2\,(P_{99.865} - P_{0.135})}\right\}$$ Equation 6-7

For data from a normal distribution, C_{Np} and C_{Npk} will be equal to P_p and P_{pk}. A method to estimate percentiles from a sample of data is given in Section 6.2.3.

The performance of C_{Np} and C_{Npk} for estimating the process capability was investigated by Wu *et al* (2007). A large number of data sets were simulated from a variety of distributions, including the Beta, Gamma and Student's t- distributions, with sample sizes ranging from ten to 3000. Percentiles and process capabilities were estimated from the samples and these values were compared to the underlying distributions to determine the relative bias of the estimates. The relative bias was defined as the ratio of the error to the true value of the capability of the underlying distribution.

Wu *et al* (2007) found that in general the median tended to be well estimated from samples, since the relative bias was close to zero, but C_{Np} and C_{Npk} were overestimated. This was because the width of the data in samples was not generally as wide as the width of the actual underlying distribution. For example, for a sample of size 100, the largest data point of a sample must be at least as large as the value of $P_{99.865}$ for the estimate to be correct. However the probability of this occurring is approximately 12.6% ($1-0.99865^{100}$), so for at least 87.4% of samples $P_{99.865}$ will be underestimated. As the sample size increases, the relative bias will reduce.

The errors from estimating $P_{0.135}$ and $P_{99.865}$ were compared for several distributions. For a sample size of 50, the smallest relative biases were attained for the Normal, Student's t-, Laplace and upper tail of the χ^2 distribution. The largest biases were recorded for the lower tails of the uniform, χ^2, Gamma and Weibull distributions. For these distributions, the true value of the percentile, 0.135, is close to zero so a small difference between the estimated and true value is perceived to be a large relative bias.

The accuracy of the estimators could be further analysed by assessing the variation due to sampling and quantifying how much the individual estimates vary between samples. For each distribution, Wu *et al* (2007) calculated the average capability estimate from a number of samples, but did not discuss the variability. However, high

variability between samples may cause individual samples to have low accuracy. Ideally a capability estimate is needed which will consistently give good results for a sample of data.

6.2.2 Asymmetric Data

When the shape of the data is highly skewed, the median does not lie in the centre of the range of the data, so the distribution will be more dispersed to one side of the median. In this case it is beneficial to assess each side separately.

Grau (2010) proposed modifying the percentile method to look at the dispersion on each side of the median separately, $C^{\#}_{pk}$:

$$C^{\#}_{pk} = min\left\{\frac{USL - M}{(P_{99.865} - M)}, \frac{M - LSL}{(M - P_{0.135})}\right\} \quad \text{Equation 6-8}$$

where M is the median. The difference between C_{Npk} and $C^{\#}_{pk}$ depends on how far the median is from the centre of the data range. For $C^{\#}_{pk}$ the variation in one tail of the data does not impact on the result in the other tail. However for C_{Npk} the overall variation in the data is measured for both tails of the data.

The use of C_{Npk} or $C^{\#}_{pk}$ depends on the interpretation of the shape of the data. Figure 6-8 shows two processes with positive and negative skew. Both processes have 0.135% of the data above the USL, so the USL is equal to $P_{99.865}$, equivalent to a P_{pk} value of one.

$C^{\#}_{pk}$ assigns both processes a capability value of one, since in both cases the distance from the median to the USL is equal to the distance from the median to $P_{99.865.}$ However C_{Npk} assigns a higher capability to the first process since the median is far from the upper limit. This process could be considered to be more capable because more of the data is far from the upper limit and a small shift in the median of the data will not cause a large proportion of the data to fall above the USL. However it could be considered that this process shows poor control, because there is a large positive skew towards the USL.

Conversely C_{Npk} assigns a lower capability to the second process because the median is close to the upper limit. Consequently as most of the data is close to the USL, a small shrift in the median could result in a large number of data points outside of the upper limit. It could also be argued that this process is well controlled at the upper end, since there is no long tail. The converse will also be true for the LSL. The overall

decision with regard to the process capability will depend on knowledge of the process being studied and how well it is controlled.

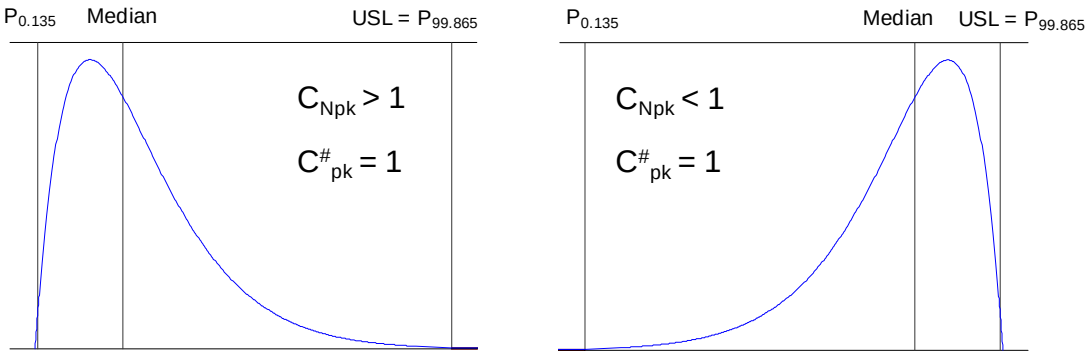


Figure 6-8: Examples of data from two skewed processes, each with 0.135% of the data about the USL

6.2.3 Method for estimating percentiles

When an exact percentile lies between two data points, the value can be estimated from the points on either side. Chang and Lu (1994) proposed a method that uses interpolation to calculate the percentile from the two data points.

The dataset of n points is sorted into ascending order: X_1, X_2,… X_n, where $X_i \leq X_j$ for all i<j. The percentile, P, is denoted as a percentage, so for the 99.865th percentile P is equal to 99.865%. To find the data point immediately below P, first calculate R:

$$R = \frac{(n-1)P + 100}{100}$$ **Equation 6-9**

Let [R] equal the largest integer less than or equal to R. Then the percentile P is located between the data points $X_{[R]}$ and $X_{[R]+1}$ (Figure 6-9). The value of R-[R] shows the proportion of the distance between $X_{[R]}$ and $X_{[R]+1}$ where P% is located. Assuming n is greater than one and P is less than 100, the percentile P is given by:

$$P = X_{[R]} + (R-[R])\left(X_{[R]+1} - X_{[R]}\right)$$ **Equation 6-10**

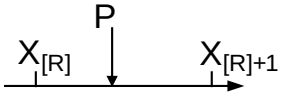


Figure 6-9: Location of percentile P between consecutive data points

6.2.4 Capability Index Using the Proportion of Data Out of Specification

An alternative description of the capability of a process is based on the likelihood of future observations lying outside of the specification limits. Chen and Ding (2001)

proposed a method to calculate the process capability based on the percentage of data outside of the specification limits and then calculating the equivalent P_{pk} value that would be observed for normally distributed data, $S_{pk}(\%)$:

$$S_{pk}(\%) = \frac{\Phi^{-1}(1 - {}^{p}/_{2})}{3}$$ Equation 6-11

where p is the proportion of the population that is outside of the specification limits, and Φ denotes the standard normal distribution function. Where data lies outside of both the upper and lower specification limits, p is the sum of the proportions outside of each limit.

To calculate S_{pk} from a sample of data, the proportion of data that is below each specification limit is found, denoted F(USL) and F(LSL), Figure 6-10. Then $\hat{S}_{pk}(\%)$ can be calculated as:

$$\hat{S}_{pk}(\%) = \frac{\Phi^{-1}\left(\frac{1 + F(USL) - F(LSL)}{2}\right)}{3}$$ Equation 6-12

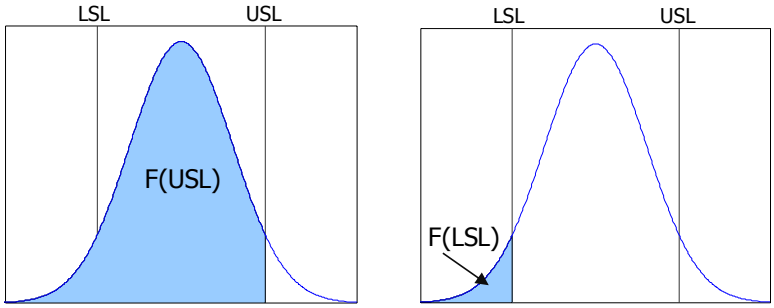


Figure 6-10: Definition of F(USL) and F(LSL)

The metric $S_{pk}(\%)$ can only be applied when some data lies outside of the specification limits. Otherwise F(USL) = 1 and F(LSL) = 0 and $\Phi^{-1}(1)$ does not exist. This method is thus only applicable to processes which exhibit poor capability, with observations lying outside of a specification limit. A process with data close to, but not outside of the limits would have a lower capability than a process where all the data lies well within the limits, but this difference would not be indicated by $S_{pk}(\%)$. Therefore $S_{pk}(\%)$ cannot differentiate between good and very good processes.

6.2.5 Percentile of Specification Limit

A potential novel improvement to the calculation of $S_{pk}(\%)$ would be to determine the percentiles of the data that correspond to each specification limit, as opposed to the proportion of data below each limit:

$$\hat{S}_{pk}\,(\text{percentile}) = \frac{\Phi^{-1}\left(\frac{1 + P(USL)/100 - P(LSL)/100}{2}\right)}{3}$$ Equation 6-13

where P(USL) and P(LSL) are the percentiles of the upper and lower specification limits respectively. That is, if the USL is the 95th percentile, then P(USL) is 95%.

A difference between $\hat{S}_{pk}\,(\%)$ and $\hat{S}_{pk}\,(\text{Percentile})$ will be observed when, for example, the first data point outside of the limit lies far away from the limit. To calculate $\hat{S}_{pk}\,(\%)$, the only information required is the percentage of data outside of the specification limits. However calculation of the percentile will depend on how close the data points on either side of a limit are to the limit itself and consequently more information is utilized, potentially improving the accuracy of the capability estimate.

For example Figure 6-11 shows two sets of data, both with one data point above the USL (83%), so both data sets would have the same value of S_{pk}(%), 0.46. It could be argued that data set A is more capable than set B, because the out of specification point in set A is just above the limit, whereas the out of specification point in set B is further outside the limit, suggesting that the capability is worse. The value of P(USL) for data set A is 90%, compared to 85% for set data set B, so the two data sets have S_{pk}(percentile) values of 0.55 and 0.48 respectively.

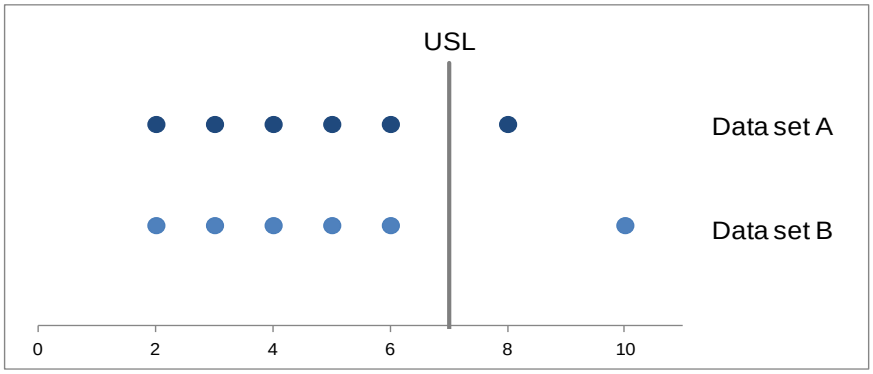


Figure 6-11: Examples of two data sets with different percentiles for the USL

The method shown in Section 6.2.3 to find the value of a percentile (Chang and Lu, 2004), can be adapted to find the percentile of a given value, for example a specification limit. For the USL, with the data sorted into numerical order, let r be the index of the highest data point below the USL, and $X_{(r)}$ the corresponding observation. The percentile of the USL is found as follows:

$$P(USL) = \frac{(r + \alpha)100 - 100}{n - 1} \qquad \text{where:} \qquad \alpha = \frac{USL - X_r}{X_{r+1} - X_r}$$ Equation 6-14

When simulating data (Section 6.3), if there is no data outside of the specification limits, the values of S_{pk}(%) and S_{pk}(percentile) will be set to one, since this value describes a process with a low risk of data failing a specification limit.

6.3 Assessment of Performance of Process Capability Indices for Simulated Data

The performance of the process capability indices discussed in Section 6.2 is now investigated with respect to their application to simulated non-normal data. A number of alternative statistical distributions were considered which reflect trends that have been observed in the industrial process data (Section 6.3.1). For each distribution, the upper specification limit was set such that the same proportion of data lay above the limit, hence it can be assumed that the individual sets of data have the same underlying capability. By drawing repeated samples from each distribution, the median and variability associated with each of the process capability metrics can be quantified.

A particular process capability index may exhibit good accuracy because the average from several repeat samples lies close to the true capability. However significant variation between individual values of the index could indicate that there is a large sampling variation, consequently an index calculated from an individual sample may not reflect the true underlying capability. For process data, sample sizes will typically be small, hence a process capability index must be able to exhibit good accuracy and precision for small samples, for example with less than 100 values.

The indices, $C^{\#}_{pk}$ and S_{pk}(%) have previously been studied to calculate the mean difference between the true and estimated capability for a number of distributions (Grau, 2010, Chen and Ding, 2001). Only Grau (2010) considered the variation resulting from drawing repeated samples from the same distribution and calculating $C^{\#}_{pk}$, however the results were not compared to other indices.

6.3.1 Alternative Distributions

A number of distributions were considered that reflect the non-normal trends that have been observed in industrial process data, such as those in Section 6.1.4. The Gamma (3, 10) shows a positively skewed distribution, where 3 is the shape parameter and 10 is the scale parameter (Figure 6-12). The Beta (3, 2) distribution has a more rounded shape, which can be seen when a process is controlled within certain limits, to prevent an upper tail from breaching a specification limit (Figure 6-13), both parameters define the shape of the distribution. Process data can also be bimodal (Figure 6-14). To generate this distribution, data from two normal distributions with the same variance

(one) but different means (ten and 14) were combined. Data from processes can show long tails on both sides of the main peak of the data (Figure 6-15), for example when poor control causes a large number of more extreme results. To generate this distribution, data from two normal distributions with different variances (0.5 and two) but the same mean (ten) were combined. The normal distribution is also included, to assess how the alternative PCIs compare to the values of P_{pk} (Figure 6-16).

For each distribution, the 95th and 99.986th percentiles are shown (Figure 6-12 to Figure 6-16), these denote the specification limits used in this simulation study. In this study, only the upper specification limits of the distributions were considered, since all of the non-normal trends to be studied are captured in the upper tails of the distributions.

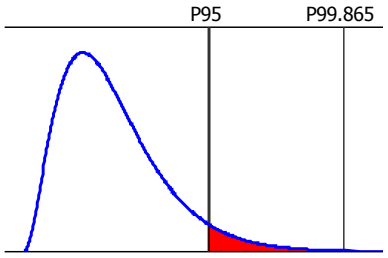


Figure 6-12: Gamma(3, 10)

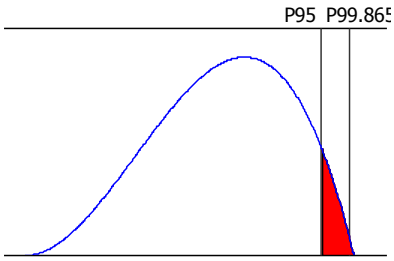


Figure 6-13: Beta (3, 2)

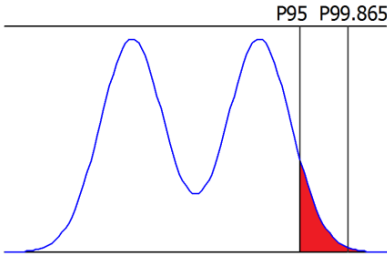


Figure 6-14: Bimodal distribution

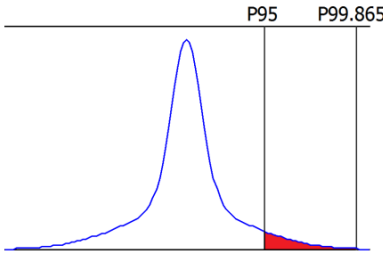


Figure 6-15: Peaked distribution

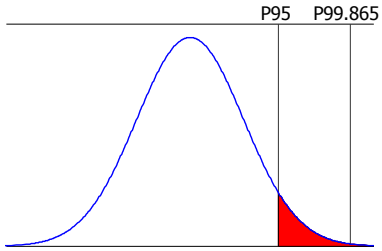


Figure 6-16: Normal distribution

6.3.2 Methodology for Simulation Study

Simulations were run for each of the distributions described in Section 6.3.1. Two sets of upper specification limits were considered. Firstly the limits were set so that 5% of the distribution was above the USL, therefore it is assumed that each distribution has an underlying capability of 0.55. Secondly the limits were set so that 99.865% of the distribution was below the upper limit, which is equivalent to a process with a P_{pk} value of one. The USLs for each distribution are shown in Figure 6-17.

For the Gamma, Beta and normal distributions, samples were generated of 100 observations. The peaked and bimodal distributions were generated from two samples

of normal distributions with different parameters, each of size 50. The two samples were combined to produce one dataset of size 100.

For each sample generated, the following PCIs were calculated: P_{pk}, C_{Npk}, $C^{\#}_{pk}$, S_{pk}(%) and S_{pk}(percentile). Sample generation was repeated 10,000 times for each distribution to observe the variability associated with each estimator. The median and interquartile range (IQT) were calculated for each set of samples. Figure 6-17 shows an overview of the simulation study. Simulations were run using Matlab 7.7.0

Distribution	Gamma(3,10)	Peaked: Combine N(10,0.5), N(10,2)	Beta(3,2)	Bimodal: Combine N(10,1), N(14,1)	Normal (10,1)
USL1	63.0	12.6	0.902	15.3	11.6
USL2	108.7	15.6	0.985	16.8	13.0

Each distribution

Start with USL1				
Sample 100 data points				
Calculate mean, standard deviation Calculate P_{pk} (Equation 7-2)	Calculate median, $P_{99.965}$ (Equation 7-10) Calculate $P_{0.135}$ Calculate C_{Npk} (Equation 7-7)	Calculate $C^{\#}_{pk}$ (Equation 7-8)	Calculate proportion below USL: F(USL) Set F(USL)=0 Calculate S_{pk}(%) (Equation 7-12)	Calc percentile of USL: P(USL), (Equation 7-13) Set P(LSL)=0 Calculate S_{pk}(prcntl) (Equation 7-13)
Repeat sampling 10,000 times				
Calculate median P_{pk} of samples Calc IQR of samples	Calculate median C_{Npk} of samples Calc IQR of samples	Calculate median $C^{\#}_{pk}$ of samples Calc IQR of samples	Calculate median S_{pk}(%) of samples Calc IQR of samples	Calculate median S_{pk}(prcntl) of samples Calc IQR of samples
Repeat for USL2				

Figure 6-17: Methodology for simulation study

6.3.3 Simulation Results

Summary statistics for each distribution are shown in Table 6-1. For comparison, the distributions were scaled so that each had a mean of ten and a P_{95} of 11.6, consistent with a Normal (10, 1) distribution (Figure 6-18 to Figure 6-22). Each data point, x, in the distribution was scaled using the population mean and 95th percentile:

$$\frac{x-\mu}{P_{95}-\mu} * 1.6 + 10$$ Equation 6-15

Scaling was applied to a allow comparison of the statistics that are used to calculate the various PCIs in the simulation study.

Figure 6-23 and Figure 6-24 shows the median capability estimates from the repeated samples, where the true capabilities of the simulated processes are 0.55 and one respectively. With the exception of C_{Npk}, the median results are similar for the Gamma and peaked distributions, because they both exhibit a long tail on the direction of the upper limit and the process capability indices are concerned with the tail ends of the distribution. Similarly, the results are also comparable for the Beta and bimodal distributions, which both have short tails. Therefore, the peaked and bimodal distributions are excluded from the analysis of P_{pk} and S_{pk}.

	Gamma	Peaked	Beta	Bimodal	Normal
Mean	10	10	10	10	10
Median	9.84	10	10.08	10	10
SD	0.86	0.96	1.09	1.10	1.00
Mean+3 SD	12.6	12.9	13.3	13.3	13.0
P 95	11.6	11.6	11.6	11.6	11.6
P 99.865	13.9	13.6	12.1	12.4	13.0

Table 6-1: Summary statistics for simulated distributions

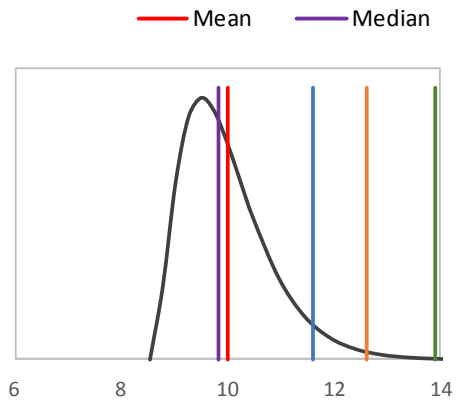


Figure 6-18: Scaled Gamma distribution

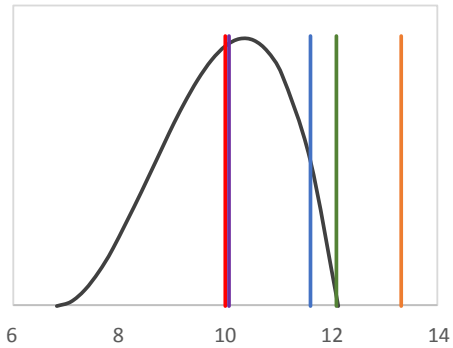


Figure 6-19: Scaled Beta distribution

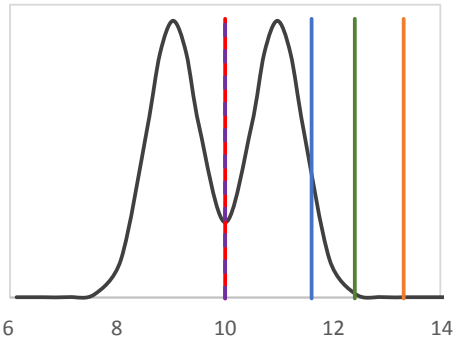


Figure 6-20: Scaled bimodal distribution

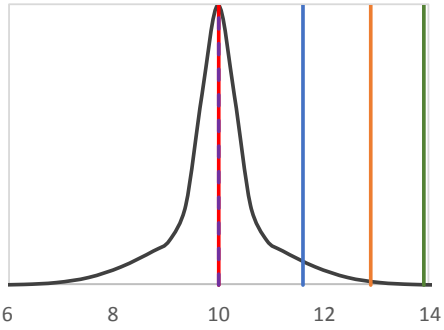


Figure 6-21: Scaled peaked distribution

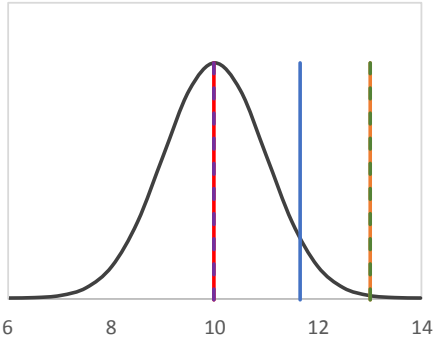


Figure 6-22: Scaled normal distribution

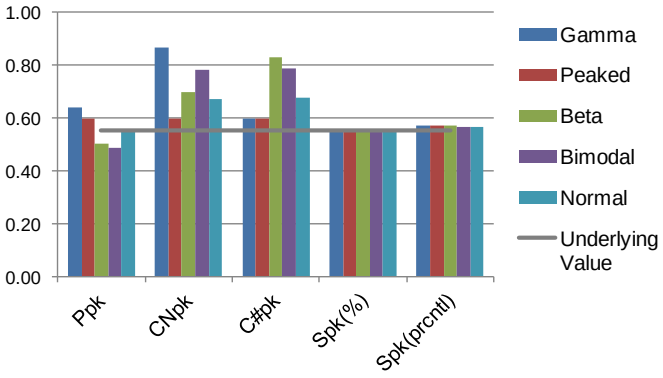


Figure 6-23: Median capability results, underlying capability is P_{pk}=0.55

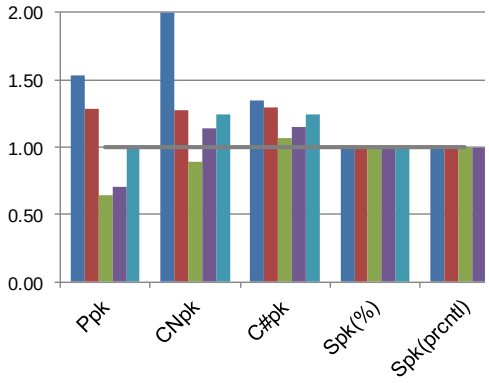


Figure 6-24: Median capability results, underlying capability is P_{pk}=1.0

6.3.3.1 Median results for P_{pk}

Figure 6-25 and Figure 6-26 show the distributions of the capability estimates from the repeated samples, for the two capability levels. For the long tailed Gamma distribution, the capability tends to be overestimated since the distance from the mean to $P_{99.865}$ is greater than three standard deviations of the distribution and hence three standard deviations does not capture the variability within the data (Table 6-1). In contrast, for the Beta distribution, three standard deviations overestimates the variability within the

data and hence P_{pk} underestimates the capability. As expected for P_{pk}, the average capability estimate for the normal distribution aligns with the true capability of the data.

Comparing the samples with lower and higher underlying capabilities, the specification limits change very little for the Beta distribution (11.6 to 12.1), so the calculated P_{pk} values are similar for both capability levels and therefore are generally underestimated for the higher capability. However for the Gamma distribution, the specification limits are much wider for the higher capability, 11.6 compared with 13.9, so this process appears particularly capable.

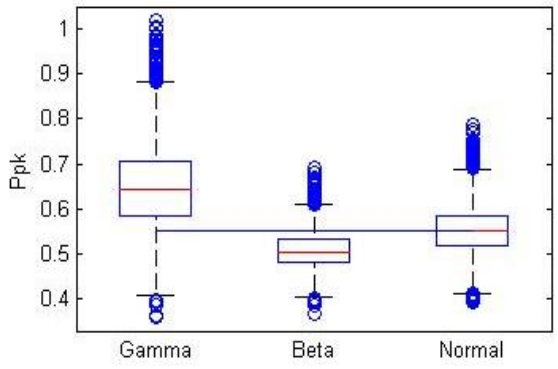


Figure 6-25: Simulation results for P_{pk}, underlying capability is P_{pk}=0.55

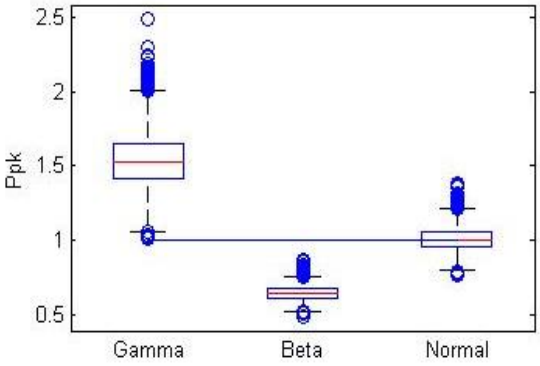


Figure 6-26: Simulations results for P_{pk}, underlying capability is P_{pk}=1.0

6.3.3.2 PCIs based on Percentiles

In general the process capability is overestimated for the metrics C_{Npk} and $C^{\#}_{pk}$, for both capability levels (Figure 6-27 and Figure 6-30). When samples are generated there will be cases where the data will not lie in the ends of the tails of the underlying distribution, so the percentiles of the sample will not reflect the full range of the distribution and consequently the process will appear more capable than it is in practice.

The Beta distribution has a particularly higher than expected $C^{\#}_{pk}$ for the lower underlying capability (0.55). The upper tail of this distribution is very short, so the estimate for $P_{99.865}$ is close to the USL (P_{95}), resulting in a capability estimate close to one. Since the upper tail is very narrow, the 5% of the distribution that is above the USL is very close to that limit, so the capability appears to be good ($C^{\#}_{pk}$=0.83). Conversely for the Gamma distribution, some of the out of specification results will be much larger than the USL, resulting in some samples with lower than expected capability and a lower median than the Beta distribution at the 0.55 capability level.

As expected, the results for C_{Npk} and $C^{\#}_{pk}$ are similar for the symmetric distributions; peaked, bimodal and normal. However more differences are observed for the Gamma and Beta distributions. To use C_{Npk} it is assumed that the median lies in the centre of

the data range, so there is equal variation above and below the median. For the Gamma distribution, there is more variation above the median than below and this is not captured by C_{Npk}, so the estimated capability is high. Conversely for the Beta distribution, the variation above the median is small, so the capability is underestimated. With the exception of the beta distribution at the lower capability level, the median $C^{\#}_{pk}$ value is as close as or closer than the median C_{Npk} value to the underlying capability. Hence in general the accuracy is higher for $C^{\#}_{pk}$ than C_{Npk}, and $C^{\#}_{pk}$ is the preferred method for estimating capability based on percentiles.

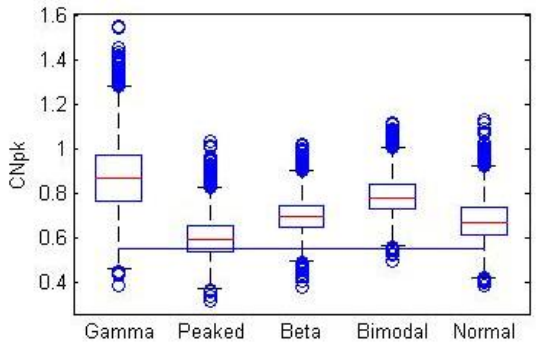


Figure 6-27: Simulation results for C_{Npk}, underlying capability is 0.55

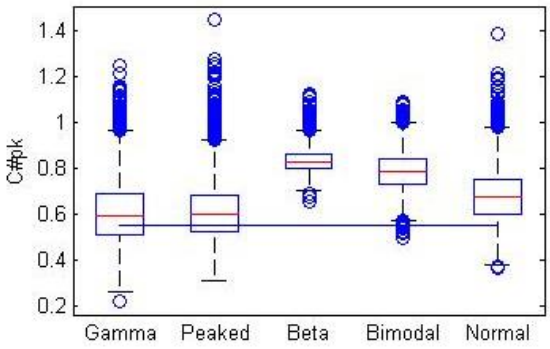


Figure 6-28: Simulation results for $C^{\#}_{pk}$, underlying capability is 0.55

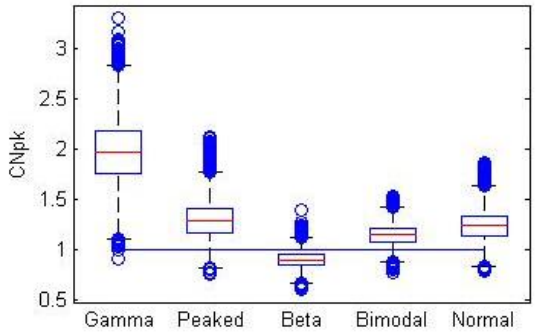


Figure 6-29: Simulation results for C_{Npk},, underlying capability is 1.0

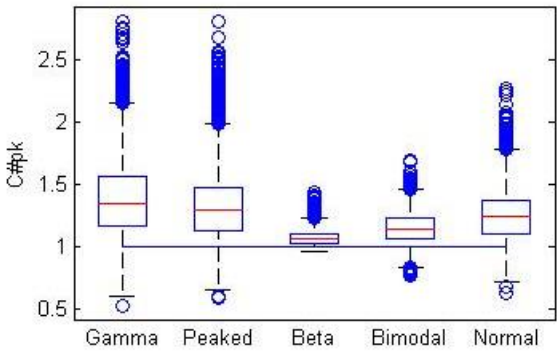


Figure 6-30: Simulation results for $C^{\#}_{pk}$, underlying capability is 1.0

6.3.3.3 PCIs based on proportion of data out of specification

For the two S_{pk} metrics, the median capability estimates are consistent across the different distributions and are close to the expected capabilities, 0.55 and 1.0 (Figure 6-31 to Figure 6-34). Also minimal differences are observed between the median values for the two methods, S_{pk}(%) and S_{pk}(percentile), suggesting that utilising the percentile of the limit as opposed to the percentage of failures does not affect the accuracy of the method.

The difference between S_{pk}(%) and S_{pk}(percentile) is found in the individual values of the capability estimates. The values of S_{pk}(%) change in steps as the number of data points outside of the specification limits change, so for a given samples size there is a

finite set of values that S_{pk}(%) can take. This trend was particularly apparent when the underlying capability was set to one, with only four unique values of S_{pk}(%) observed (Figure 6-32). In contrast, calculating the percentile between data points allows for a greater range of values, resulting in greater variability in the capability results (Figure 6-34).

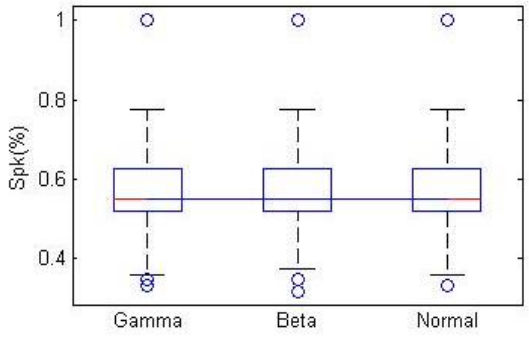


Figure 6-31: Simulation results for S_{pk}(%),underlying capability is P_{pk}=0.55

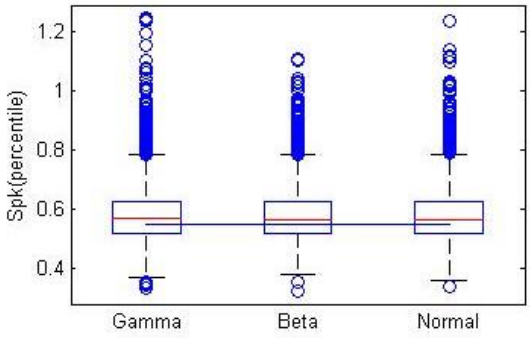


Figure 6-32: Simulation results for S_{pk}(percentile), underlying capability is P_{pk}=0.55

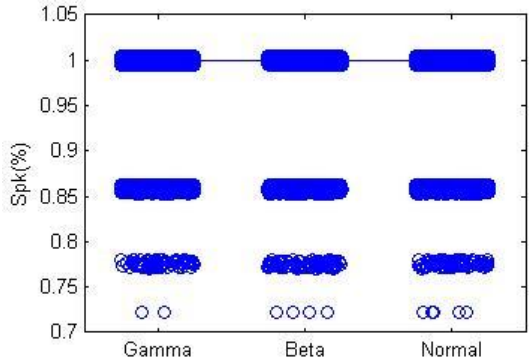


Figure 6-33: Simulation results for Spk(%), underlying capability is P_{pk}=1.0

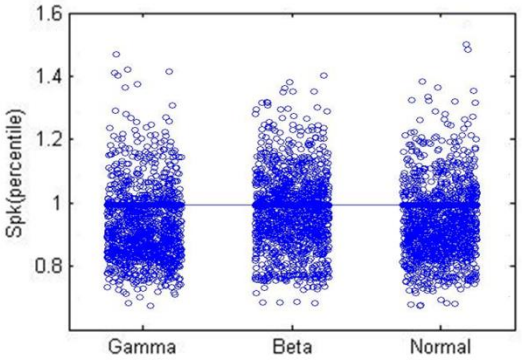


Figure 6-34: Simulation results for S_{pk}(percentile), underlying capability is P_{pk}=1.0

6.3.3.4 Variation of Estimates

The variation of the capability estimates can be quantified by the interquartile range of the repeated samples. In general, the P_{pk} estimates have the lowest sampling variability (Figure 6-35 and Figure 6-36). The mean and sample standard deviation calculations use all of the data in the sample and hence the results are more consistent for P_{pk} than for $C^{\#}_{pk}$ and S_{pk}, which consider fewer data points lying in the tails of the distribution.

When the distribution has a long tail, $C^{\#}_{pk}$ generally exhibits particularly large sampling variability, since the calculation depends on the largest data points and these results will vary between samples. However when applied to the Beta distribution, $C^{\#}_{pk}$ exhibits low variation because the upper tail is short and hence greater consistency is observed

between samples. Compared to $C^{\#}_{pk}$, C_{Npk} appears to exhibit larger variation for the Gamma and Beta distributions, but smaller variation for the symmetric distributions. Since the variation is largest for the Gamma distribution, $C^{\#}_{pk}$ is preferred, to avoid high sampling variation.

For the lower capability, the results from the S_{pk} metrics are generally more variable than the P_{pk} results, again because only the largest values of the sample are considered in the calculation. The variation for S_{pk} is consistent across the three distributions, because it is the number of out of specification results that is considered, rather than the distance from the specification limit. When the underlying capability was set to one, the interquartile range was zero for S_{pk} because the majority of the samples had no out of specification results and hence of the capability estimates were set to one.

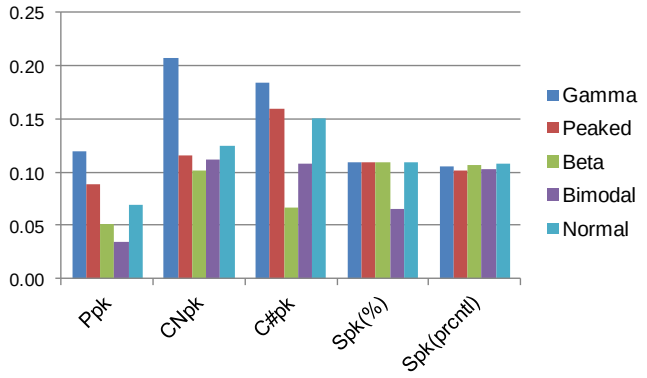


Figure 6-35: Interquartile range results, when underlying capability is P_{pk}=0.55

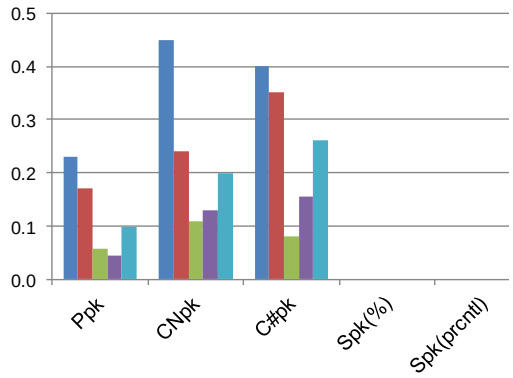


Figure 6-36: Interquartile range results, when underlying capability is P_{pk}=1.0

6.3.4 Sample Size

Data from industrial processes may be limited in terms of sample size, particularly for new processes or when a process is being investigated after a significant change has been implemented. The effect of sample size was investigated with the upper specification limits set so that 5% of the distribution is above the limit. The sample sizes considered varied between 30 and 100, in increments of 10, and a sample size of 1000 was also considered. The peaked and bimodal distributions were again excluded because they exhibit the same trends as the Gamma and Beta distributions respectively. Additionally C_{Npk} is excluded since $C^{\#}_{pk}$ is the preferred metric that uses percentiles of the data, as discussed in Section 6.3.3.2.

For P_{pk}, the median value changes minimally when the sample size is reduced (Figure 6-37), since the expected values of the mean and standard deviation are invariant of sample size. The $C^{\#}_{pk}$ estimates increase as the sample size is reduced (Figure 6-38). For a smaller sample, the data is less likely to fall in the tails of the distribution and

hence the sample appears to be narrower than the underlying distribution and the calculated $C^{\#}_{pk}$ is larger than expected.

The median values of S_{pk}(%) are consistent as the sample size is reduced (Figure 6-39), since the median number of data points above the USL does not change with sample size. However the value of S_{pk}(percentile) increases as the sample size is reduced (Figure 6-40). Since the percentile of the USL is larger than the percentage of data below the USL, S_{pk}(percentile) will be greater than or equal to S_{pk}(%). When the sample size is reduced, there will be a greater distance between the two data points on either side of the USL, so S_{pk}(percentile) can take larger values and hence the median is greater. For example, for a sample of 100 with five points (5%) above the USL, P(USL) will be between 95% and 96%. However for a sample of 40 with two points (5%) above the USL, P(USL) will be between 95% and 97.5% and hence the median S_{pk}(percentile) from a number of samples will be greater with a smaller sample size.

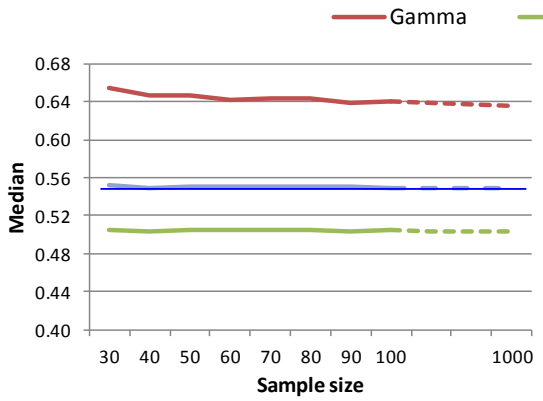


Figure 6-37: Median P_{pk} vs. sample size

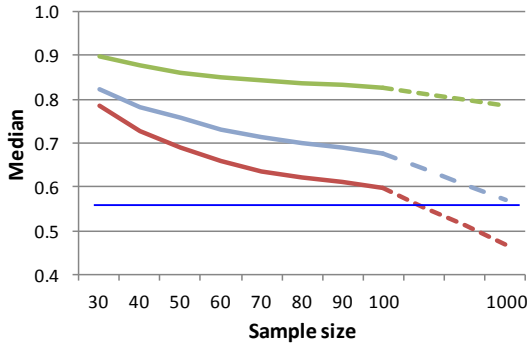


Figure 6-38: Median $C^{\#}_{pk}$ vs. sample size

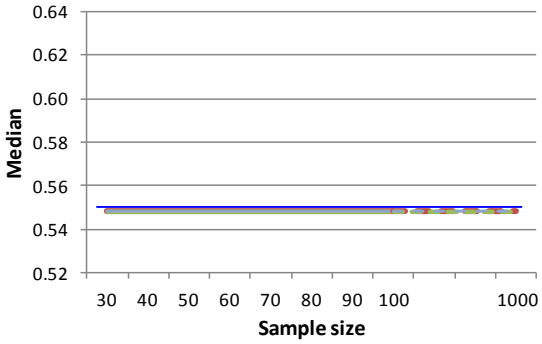


Figure 6-39: Median S_{pk}(%) vs. sample size

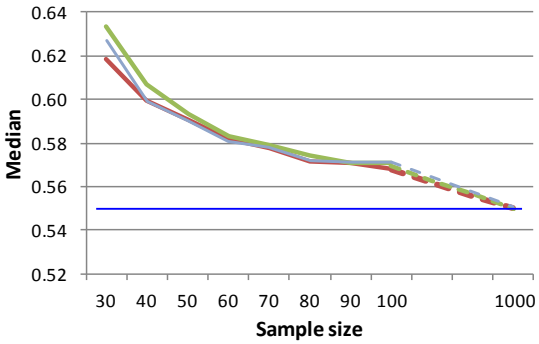


Figure 6-40: Median S_{pk}(percentile) vs. sample size

The variability was quantified by the inter quartile range of the individual process capability measurements from the repeated samples (Figure 6-41 to Figure 6-44). For S_{pk}(%) and S_{pk}(percentile), the interquartile range results were very similar for the three distributions and hence the lines are overlaid (Figure 6-43 and Figure 6-44).

As expected, the variability in the capability indices increases as the sample size reduces, because there is greater variation between the data in each sample. The metric $C^{\#}_{pk}$ has a particularly high sampling error for the Gamma distribution when the sample size is below 50, and so may not provide a reliable capability estimate for smaller sample sizes (Figure 6-42). This trend was also noted by Grau (2010), where the distribution of repeated estimates was seen to become wider when the sample size was reduced from 250 to 50.

Each distribution shows a reduction in variation when the sample size is increased from 100 to 1000, indicating that a limited sample size will contribute to variability in the resulting capability estimate. Therefore where possible a larger sample size is recommended, especially when the underlying distribution is known to have a long tail.

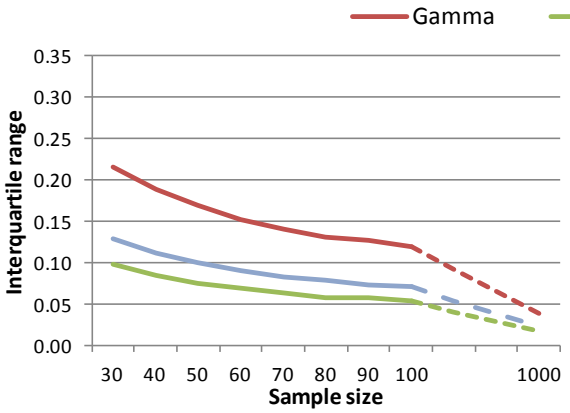


Figure 6-41: Sampling error of P_{pk} vs. sample size

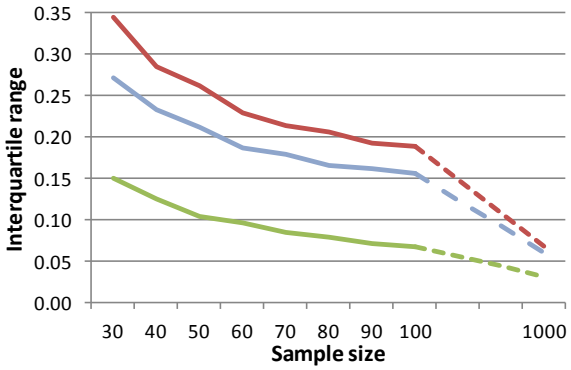


Figure 6-42: Sampling error of $C^{\#}_{pk}$ vs. sample size

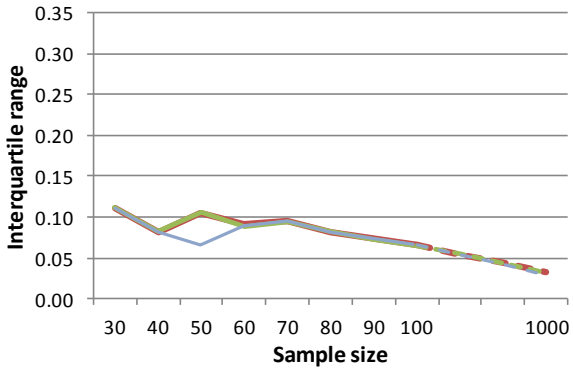


Figure 6-43: Sampling error of S_{pk}(%) vs. sample size

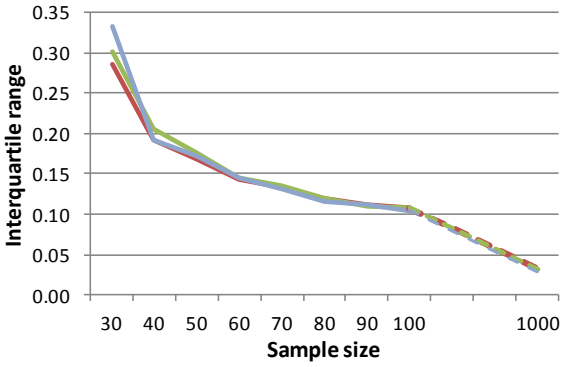


Figure 6-44: Sampling error of S_{pk}(percentile) vs. sample size

6.3.5 Discussion of Simulation Study

The results from the P_{pk} metric are centred about the capability of the underlying distribution and show low sampling variation compared to the alternative metrics, especially for smaller sample sizes of 50 and lower. Therefore P_{pk} may be a reliable metric for data that is not too dissimilar to a normal distribution.

Of the metrics based on the percentiles of the sample, the capability tends to be overestimated because the samples do not reflect the full width of the distribution, particularly for smaller sample sizes. In general P_{pk} was found to be more accurate when the underlying capability is 0.55 and $C^{\#}_{pk}$ was more accurate when the underlying capability was set to one. C_{Npk} has been found to be effective for distributions that are symmetric or have low skewness, but $C^{\#}_{pk}$ additionally provides a more reliable metric for data skewed towards a specification limit. However for a highly skewed Gamma distribution, sampling error can cause high variation in the $C^{\#}_{pk}$ results that are calculated.

The metrics based on the amount of data outside of the limits show consistent results for all of the distributions, with good accuracy when there is some data out of specification. The sampling variation is generally lower than for $C^{\#}_{pk}$, so for processes with some data outside of the limits, S_{pk} may provide a reliable estimate of the capability. However when there are no failures in the data, S_{pk} cannot be used to determine if the capability is greater than one. The S_{pk}(percentile) metric was found to provide more detail than S_{pk}(%), because more information about the location of the specification limit is used in the calculation. However the sampling variation for S_{pk}(percentile) increases as the sample size reduces.

6.4 Application to Process Data

The performance of the distribution free capability indices is now assessed on data from an industrial process. Four in-process variables from a pharmaceutical manufacturing process were selected that all exhibited some form of non-normal behaviour. The data were scaled to have a minimum of zero and a maximum of ten (Table 6-2). The data for each variable must lie within a range that is registered with the regulatory authorities, to allow a batch of product to be released. The selected variables all have P_{pk} values below one, which suggests that they have poor capability; however none of the data lies outside of the specification limits. The sample sizes vary depending on the amount of data available for each variable.

The results are summarised in Table 6-3. Since none of the data lies outside of the specification limits, the values of S_{pk}(%) and S_{pk}(percentile) cannot be calculated and hence they are set to one.

Variable	Variable one	Variable two	Variable three	Variable four
N	89	89	49	92
Mean	3.74	7.83	4.30	5.87
Median	3.24	9.73	2.92	5.69
Standard deviation	2.17	3.08	3.26	1.89
6*SD	13.0	18.5	19.5	11.3
Range	10	10	10	10
Skewness	0.57	-1.23	0.49	-0.23

Table 6-2: Summary statistics of process data

	Variable one	Variable two	Variable three	Variable four
P_{pk}	0.58	0.91	0.86	0.79
$C^{\#}_{pk}$	1.02	1.10	2.40	1.06
S_{pk}	1.00	1.00	1.00	1.00

Table 6-3: Process capability estimates for plant data

6.4.1 Variable One

Figure 6-45 shows a histogram of the data for variable one, the upper and lower specification limits and the percentiles used to calculate $C^{\#}_{pk}$. There is a large amount of data close to the LSL as the optimal operating conditions lie in this region; however the process is controlled such that the measurement should not fall below the lower limit. Both P_{pk} (0.58) and $C^{\#}_{pk}$ (1.02) generally identify the capability as being low (Table 6-3), because the data lies close to the lower limit. More specifically, by fitting a normal curve to the data the distribution extends to below the LSL, hence the P_{pk} value is well below one. The data is a similar shape to a beta distribution, for which P_{pk} tends to overestimate the capability, and $C^{\#}_{pk}$ tends to underestimate (Section 6.3.3.2), hence the true capability of this process could be between 0.6 and one.

6.4.2 Variable Two

Variable two (Figure 6-46) displays a negatively skewed distribution, with the majority of the data falling close to the median. The $C^{\#}_{pk}$ estimate is 1.1, which may be a good indication of the actual capability if the current dataset shows the whole range of values that would be expected from future observations. However if there is a risk of future observations lying below the LSL then the true capability of the process may be lower and the P_{pk} value of 0.91 may be more representative of the true capability. Of the distributions considered in the simulation study, the data is most similar to a gamma distribution, for which P_{pk} and $C^{\#}_{pk}$ are generally over estimated. However since the shape of the data is unusual, more understanding of the sources of variation and control for this variable may be required to determine the actual capability.

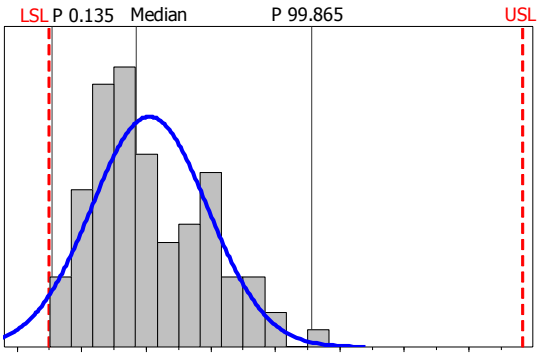


Figure 6-45: Histogram of variable one

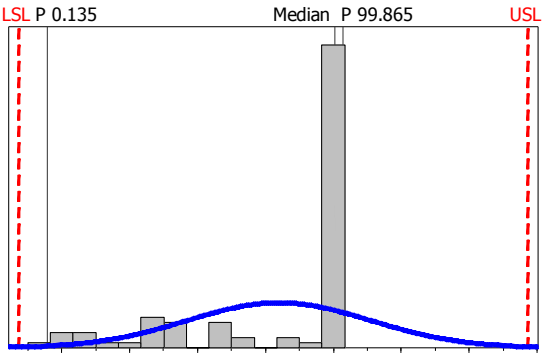


Figure 6-46: Histogram variable two

6.4.3 Variable Three

The distribution of variable three (Figure 6-47) appears to exhibit a bimodal distribution. For this data set, six standard deviations is wider than the range of the data (Table 6-2) and the normal curve extends beyond the data to the LSL, resulting in a low P_{pk} value (0.86). All of the data for this variable lies well within the limits, which suggests that there is good capability. The $C^{\#}_{pk}$ result is very high (2.4) since the lower limit of the data is well above the LSL and is close to the median, hence $C^{\#}_{pk}$ may provide a better estimate of the true capability. Generally P_{pk} underestimates and $C^{\#}_{pk}$ overestimates the capability for a bimodal distribution, so the true capability is expected to be between 0.86 and 2.4. The sample size for variable three is small, with 49 batches, so the results must be treated with caution since the sample may not completely reflect the underlying distribution.

6.4.4 Variable Four

From a sample of 92 batches, data has been observed close to both specification limits for variable four, but no values lie outside of the limits (Figure 6-48). The P_{pk} value (0.79) suggests that the capability is poor, which may be a reasonable estimate because a future observation occurring just outside of the current range of the data would fail a specification limit. The $C^{\#}_{pk}$ result is slightly higher (1.06) because the $P_{0.135}$ and $P_{99.865}$ percentiles are within the limits. The shape of the data approximately resembles a normal distribution and both capability metrics overestimate the capability for this distribution, hence both capability measures agree that the process capability should be improved.

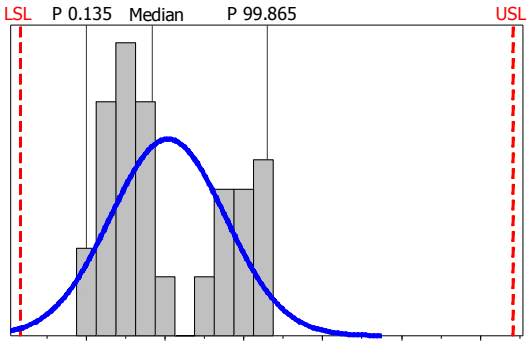


Figure 6-47: Histogram variable three

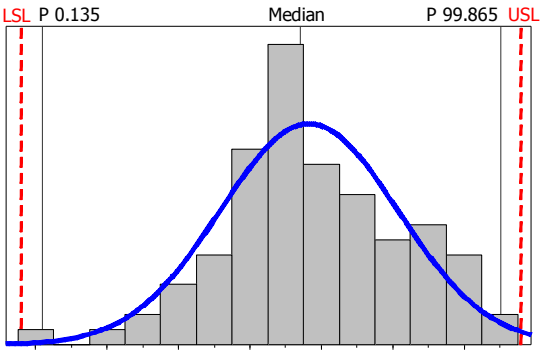


Figure 6-48: Histogram of variable four

6.4.5 Discussion of Process Data

Some examples of data from industrial processes have shown non-normal trends, such as the process being controlled close to a specification limit or a small number of outlying results extending towards a limit. In these cases, calculating the percentiles of the data may provide a better measure of the data width than the standard deviation.

The variables selected in this study all have P_{pk} values below one, which suggests that capability is currently poor; however none of the data lies outside of the specification limits. In all cases, the value of $C^{\#}_{pk}$ appears to provide a likely estimate of the capability, but visualisation of the data and knowledge of the process control are required to gain an understanding of the risk of failing a specification limit.

6.5 Conclusions

Process capability indices have been found to be useful for summarising the risk of a process failing a registered specification limit that is required for a batch to be released to the market. Using process capability, many sets of specification limits can be compared to identify the greatest risks to the process and to justify the cost of improvement work. The capability can be tracked over time to highlight improvements and to ensure that they remain effective.

The P_{pk} index is known to be appropriate for normally distributed data. However as a result of various sources of variation, industrial process data does not always exhibit normality and the P_{pk} metric may not provide an accurate representation of the underlying capability. While visualising the data and using process knowledge will provide a detailed understanding of the capability of a process, it is important to use a single metric to summarise the capability, allowing a number of variables to be compared efficiently and areas for improvement to be identified.

Several alternative process capability indices have been proposed in the literature for data that does not satisfy a normal distribution. Simulated and industrial process data have shown that the various capability indices will generate a range of results for the same dataset (Figure 6-23 to Figure 6-33). While none of the proposed indices show consistently good accuracy and low sampling variation when applied to samples from a variety of simulated distributions, each of the metrics appeared to outperform P_{pk} in certain situations.

The S_{pk}(percentile) and S_{pk}(%) metrics produced consistent and accurate results in the simulation study with 5% of the data outside of the specification limits (Figure 6-31 and Figure 6-32). Similar results were seen across all of the statistical distributions, suggesting that these metrics could be applied to a variety of data sets where failures have occurred.

For data with a higher underlying capability, $C^{\#}_{pk}$ has been found to give the most accurate results for non-normal data (Figure 6-24), which is also reflected in the results from the process data. However the sampling variation is very high for data with long tails (Figure 6-42) so this metric should be used with caution for small sample sizes, particularly with less than 50 data points. The standard P_{pk} metric has been shown to have low sampling variation, particularly for small sample sizes, but is likely to overestimate the capability of data with long tails in the distribution and underestimate the capability for data with very short tails.

This study has shown that a universal capability metric that is applicable to all data may not be possible, but by comparing several metrics and visualising the data, the underlying capability of a process may be estimated. The results of this study were shared and discussed with the Global Operations Statisticians Forum at AstraZeneca, who are using the methodologies and results to develop a best practice for process capability analysis.

Process capability metrics have been found to be useful for tracking the capability over time (Section 6.1.3.1), for example to compare the capability month by month. However when a small sample size is available in each time period, the calculated capability estimate may be unreliable (Section 6.3.4) so it may be necessary to combine data from several time periods to create a large enough sample size. In Chapter 7, the use of Bayesian methods is investigated to calculate the capability using data from one time period but also using older data as prior information.

7 Bayesian Methods in Pharmaceutical Process Development

7.1 Introduction

In classical statistics, such as linear regression analysis, parameters are assumed to be fixed quantities that can be estimated from random samples of data taken from the population. An alternative approach is the use of Bayesian statistics, here it is assumed that each parameter is a random variable with an associated probability distribution (Lunn, 2012).

The term Bayesian statistics stems from the application of Bayes Theorem, which is presented in Section 7.1.1. Section 7.1.2 provides a comparison with classical Statistics. Bayesian methods allow prior knowledge to be quantified and incorporated into an analysis (Section 7.1.3), allowing all of the available information to be utilised. A Bayesian analysis often requires the use of simulation based computational methods, which are presented in Sections 7.1.4 and 7.1.5. An overview of application is presented in Section 7.1.6.

A research theme for this thesis is to investigate potential uses of Bayesian statistics in pharmaceutical manufacture. There are a limited number of examples in the literature of applying Bayesian statistics to manufacturing processes, compared to methods such as multivariate analysis (Chapter 3). However there are some examples of how the Bayesian approaches to formalising prior information and estimating distributions for parameters have been found to be beneficial for gaining information about a process (Section 7.2 to 7.4).

In Sections 7.2 to 7.4 a number of applications of Bayesian statistics to pharmaceutical process development are presented, along with examples of case studies. In Section 7.2, Bayesian applications to experimental design are presented, including a methodology for Bayesian D-Optimal designs and a method to maximise the expected accuracy of predictions from the resulting models. These models can be used to find the optimal operating conditions for a process and Bayesian methods can be applied to identify the design space in which the process has the highest certainty of meeting the required quality conditions (Section 7.3). In Section 7.4, a Bayesian approach to calculating the process capability is presented that measures the certainty that a process is capable. Section 7.5 concludes the examples of applications.

In Section 7.6 a novel process capability methodology is proposed for the situation where the process capability is analysed over time. Bayesian techniques are proposed

to combine data from the current month of data with prior knowledge from previous months, to sequentially update the capability estimate.

7.1.1 Methodology

The origin of Bayesian statistics dates back to a manuscript on the subject of probability, written by Thomas Bayes in 1763. He proposed Bayes rule, a process for combining information from data with prior information to obtain posterior information (LeBlond, 2009). Bayes theorem (Equation 7-2) was stated by Laplace in 1774 (Colosimo and del Castillo, 2007) and builds on from conditional probability (Equation 7-1). For events A and B:

$$P(A|B) = \frac{p(A \text{ and } B)}{p(B)} \qquad \text{Equation 7-1}$$

$$P(A|B) = \frac{p(B|A)\, p(A)}{p(B)} \qquad \text{Equation 7-2}$$

Equation 7-2 is a function of event A and the denominator on the right hand side, p(B), is constant with respect to A, hence the relationship can be redefined as a proportion:

$$p(A|B) \propto p(B|A)\, p(A) \qquad \text{Equation 7-3}$$

Therefore p(B) is a constant of proportionality that will ensure that the conditional probabilities of all possible values of A sum to one.

An underlying theme of Bayesian statistics is that of making inferences about a parameter by combining two sources of information: the likelihood, or new information from data, and prior information that is already known. Both sources are captured as probability density functions of the parameters to be estimated. The prior and likelihood are then combined to form a 'posterior' distribution of the parameter of interest.

Equation 7-3 can be applied to a statistical inference problem to estimate the distribution of a parameter, θ, where new data has been collected and some prior knowledge exists about θ. The prior information is captured as a probability density function, $p(\theta)$, and the new information from the data is termed the likelihood and denoted $p(x|\theta)$. Combining the likelihood and prior gives the posterior distribution for θ given the data:

$$p(\theta|x) \propto p(x|\theta)p(\theta) \qquad \text{Equation 7-4}$$

Posterior $\propto$ Likelihood × Prior

θ can be an individual parameter or a set of parameters, for example the mean and standard deviation of a normal distribution.

7.1.2 Comparison with Classical Statistics

There are a number of differences between Bayesian and classical statistics in the way in which the data is handled and interpreted. These differences are summarised in Table 7-1. In classical statistics parameters are treated as fixed values that are estimated from random samples of data. The level of uncertainty is captured through a confidence interval about the estimated value of the parameter, which will contain the true value for a given percentage of samples. For example, if data is repeatedly sampled and a 90% confidence interval calculated on each sample, 90% of the calculated intervals will contain the true parameter value.

However in Bayesian analysis the data is treated as fixed and is used to estimate the distribution of parameters. The distribution shows the most likely values that the parameter will take and hence the width of the distribution indicates the uncertainly associated with the parameter estimates. A Bayesian interval shows the most likely range that a parameter will be found in, whereas a confidence interval shows the range in which future estimates are most likely to be found, and hence a Bayesian interval may be considered to be more intuitive (LeBlond, 2008).

Since Bayesian methods assume that parameters are variables rather than fixed quantities, these methods may be more relevant to process data. Industrial processes are not fixed and can exhibit variation over time, hence Bayesian methods can be used to capture this variation.

A further difference is the formal use of prior information in Bayesian methodologies. Information that is already known about a subject can be included informally in classical statistics. For example in an experimental design, existing knowledge is used to decide on which factors to include and what levels to run the experiment at, or in a process modelling context prior information is used to decide which terms to include in a model. In a Bayesian context, prior information is quantified can be used to design an experiment that is expected to result in the most information being gained, or to influence the values of model parameters when there is limited data available.

	Classical Statistics	Bayesian Statistics
Parameters	Fixed values to be estimated	Random variables with a probability distribution
Data	Random, due to sampling	Fixed values
Parameter Intervals	Confidence intervals: Contain the true value for a specified percentage of samples	Most likely range that the parameter will take
Prior information	No formal use	Quantified and influences the posterior distribution

Table 7-1: Comparison of the philosophies of classical and Bayesian statistics

7.1.3 Prior information

When a statistical analysis is run to analyse a set of data, information about the subject may already exist that can be incorporated into the analysis. Prior information may originate from previous studies on the same subject or from scientific knowledge of what is expected to be true (Congdon, 2003). In a process development context, prior information may be obtained from previous development work, mechanistic models of the process or information in the literature about similar processes. Prior knowledge is considered to be subjective, since experts can disagree about the information that is known about a parameter being estimated (LeBlond, 2008).

Within a Bayesian analysis, prior information is summarised as a statistical distribution for a particular parameter. In the case when data is available that specifically relates to the parameter, it can be used to calculate a prior distribution, otherwise knowledge about the parameter must be used to define a statistical distribution. For example a uniform distribution can be used to capture the range over which the parameter is expected to lie, or a normal distribution can be used to define the mean and variance of the expected value of the parameter.

If there is no appropriate knowledge available, a wide prior can be selected so that it does not influence the results of the analysis. A prior for which all values are equality likely is termed a non-informative prior (Colosimo and del Castillo, 2007).

The width of the prior distribution will influence the width of the posterior distribution. If there is a lot of prior knowledge about the location of a parameter, the prior will be narrow and hence the posterior will narrower. However if there is little certainty in the prior information, the prior distribution will be wider and the subsequent posterior will be wider.

7.1.4 Markov Chain Monte Carlo

Implementation of a Bayesian analysis requires the calculation of the posterior distribution from the prior and the likelihood distributions. If the prior and likelihood are from the same family of distributions, for example both are normal distributions, the posterior can be found analytically by multiplying the functions of the prior and likelihood distributions. However in many cases the prior and likelihood do not allow the posterior to found analytically and so sampling methods are required to obtain the posterior distribution (Lunn, 2012).

Calculation of the posterior distribution requires the constant of proportionality in Equation 7-4 to be obtained, which is equivalent to evaluating the integral $\int p(\theta)p(x|\theta)d\theta$. However evaluating this integral analytically can be difficult or even impossible and numerical integration methods can be computationally expensive. In Bayesian statistics, Markov chains are commonly used to generate samples from the posterior distribution and Monte Carlo integration is implemented to compute summaries of the posterior, such as the mean or percentiles. These two steps are collectively known as Markov Chain Monte Carlo (MCMC, Brooks, 1998). The posterior distribution is simulated by constructing a Markov chain that will converge to the required posterior distribution, termed the stationary distribution. Therefore, when run for a long time, the Markov chain will converge and provide samples from the posterior distribution.

The Markov chain will produce dependant samples of θ, denoted θ_1, θ_2, θ_3,…, θ_T, when the chain is run up to the time T. A Markov chain is defined such that the distribution of θ at time t+1 (θ_{t+1}) depends only on the location of θ_t and not on any of the previous time steps. Additionally, the chain will theoretically converge to the same stationary distribution from any initial value of θ_0; the only difference would be the time taken for the chain to converge.

The initial values of θ that are generated before the chain has converged will not form part of the posterior distribution and these samples should be discarded. Convergence can be checked by viewing a trend plot of the samples (Figure 7-1) and determining when the chain remains in the same location. The samples before the chain has converged are known as the burn in samples.

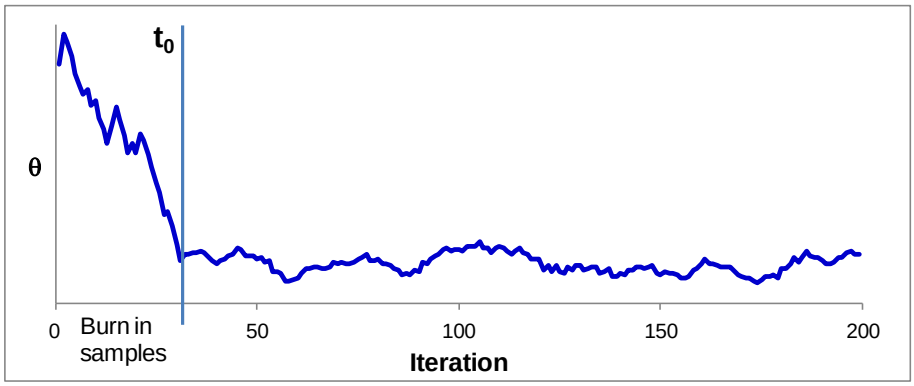


Figure 7-1: Samples of θ generated from a Markov Chain

By removing the first T_0 samples that represent the burn in samples, the remaining $T-T_0$ represent the desired posterior distribution. These samples can then be used to calculate summaries of the posterior distribution, such as the mean or standard deviation, using Monte Carlo integration. For a function of the data, g(θ), the expected value of g(θ) is calculated as:

$$E[g(\theta)] = \int g(\theta)p(\theta|x)\,d\theta \approx \frac{1}{T-T_0}\sum_{T_{0+1}}^{T} g(\theta_t) \qquad \text{Equation 7-5}$$

The use of MCMC methods can require a large amount of computational power and consequently the use of Bayesian statistics has developed rapidly over the past 30 years, due to the increase in computer power that is available (Brooks, 1998, Ntzoufras, 2009). Examples of software include WinBUGS (Lunn *et al*, 2000), which has been developed to run Bayesian analyses using simulation methods, and MATLAB, which includes commands to run Bayesian simulations.

7.1.5 MCMC Algorithms

For the application of MCMC methods in Bayesian statistics, a Markov chain must be defined such that the stationary distribution is the required posterior. A number of algorithms are available to construct such a Markov chain, including the Metropolis-Hastings (Hastings, 1970) and Gibbs Sampling (Geman and Geman, 1984). An overview of these two algorithms is given in Sections 7.1.5.1 and 7.1.5.2. The Bayesian analysis applied in this chapter is implemented in Matlab 7.7.0, which uses slice sampling (Section 7.1.5.3).

7.1.5.1 Metropolis-Hastings Algorithm

The Metropolis-Hasting algorithm is explained in Brooks (1998). A Markov chain, $\theta_{0,}$ $\theta_{1,\ldots}$ is constructed such that the samples that are generated will converge to the required posterior distribution.

The aim is to sample from an unknown posterior density, $f(\boldsymbol{\theta}|\mathbf{x})$, when a function $c(\boldsymbol{\theta}|\mathbf{x})$ is known that is proportional to $f(\boldsymbol{\theta}|\mathbf{x})$. An arbitrary function $q(\boldsymbol{\theta}_{t-1}|\ \boldsymbol{\theta}_t)$ is defined that denotes the probably of the Markov chain moving from $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$ at time t. The algorithm proceeds as follows:

1. Start with $\boldsymbol{\theta}_0$ such that $f(\boldsymbol{\theta}_0|\mathbf{x}) > 0$
2. At time t, sample $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}_{t-1}|\ \boldsymbol{\theta}_t)$
3. Accept $\boldsymbol{\theta}_t = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*)$, otherwise $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$

   Where $\alpha = \min\left\{1, \frac{f(\boldsymbol{\theta}^*|\mathbf{x})q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_{t-1}|\mathbf{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})}\right\}$

4. Repeat steps 2 and 3 until the chain converges

7.1.5.2 Gibbs Sampling

The Gibbs algorithm is explained in Gelfand and Smith (1990). Suppose the aim is to estimate the distribution $f(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]$, and each of the conditional distributions:

$f(\theta_i \mid \theta_j, \text{ for } i \neq j),\ i=1,\ldots, k$

are known.

The algorithm proceeds as follows:

1. Start with arbitrary starting values of $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$
2. At time t, sample: $\theta_1(t)$ from $f(\theta_1^{(t-1)} \mid \theta_2^{(t-1)}, \ldots, \theta_k^{(t-1)})$

   $\theta_i^{(1)}$ from $f(\theta_i^{(t-1)} \mid \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_k^{(t-1)})$
3. Repeat step 2 until the chain converges. Then the samples of θ_i are from the distribution $f(\theta_i)$

7.1.5.3 Slice Sampling

Slice sampling was developed by Neal (2003) and is used to find the density distribution for a parameter by sampling from the region under its density function. The method works by constructing a Markov chain that alternates between sampling from the vertical direction and sampling from a horizontal slice in the current vertical position (Figure 7-2, Figure 7-3).

The algorithm is used to find the probability density function of a single or set of parameters, p(θ), using a known function, f(θ), that is proportional to the required density function. In a Bayesian sense, f(θ) could be the product of the prior and likelihood densities that are proportional to the posterior distribution p(θ|x). A Markov chain is created to generate samples of θ by sampling from the region under the f(θ). Then the samples of θ are used to compute summaries of p(θ) by Monte Carlo integration.

To implement slice sampling, a new variable, y, is created such that the joint density p(θ,y) is uniform over the region U={(θ,y): 0<y<f(θ)}. The joint density is defined as:

$$p(\boldsymbol{\theta}, y) = \begin{cases} 1/Z & if\ 0 < y < f(\boldsymbol{\theta}) \\ 0 & otherwise \end{cases}, \qquad \text{Equation 7-6}$$

where Z is equal to the value of $\int f(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and hence Z is the constant of proportionality required to scale f(θ) to a probability density function. The marginal distribution for θ is given by:

$$p(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} p(\boldsymbol{\theta}, y)\, dy = \int_{0}^{f(\boldsymbol{\theta})} 1/Z\, dy = f(\boldsymbol{\theta})/Z \qquad \text{Equation 7-7}$$

Then it follows from the definition of Z that p(θ) is the probability density function that is proportional to f(θ). If f(θ) is the product of the prior and likelihood distributions, then p(θ) is the corresponding posterior distribution.

For the case where θ is a single parameter, the algorithm for slice sampling proceeds as follows. Firstly an initial value of θ_0 is defined and then samples of (θ_i, y_i) are generated by alternately updating the values for θ and y:

1. Given θ_i, sample y_i uniformly between zero and $f(\theta_i)$, Figure 7-2, i.e. sample from $p(y|\theta_i) \sim$ Uniform $(0, f(\theta_i))$.

2. Given y_i, sample θ_{i+1} from the range of θ over which y_i slices through the curve for $f(\theta)$, Figure 7-3, i.e. for the horizontal slice $S=\{\theta:y_i<f(\theta)\}$, sample θ_{i+1} uniformly over the slice S.

Steps 1 and 2 are repeated to generate a chain of (θ_i, y_i). The samples of θ_i are from the distribution of the parameter θ. Samples of y_i are not required in the subsequent analysis. When $\boldsymbol{\theta}$ is multivariate, each parameter is sampled in turn to generate the chain $(\theta_{1,i}, \theta_{2,i}\ldots \theta_{n,i}, y_{,i})$.

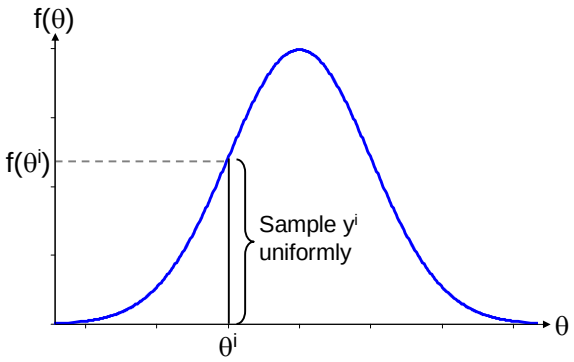


Figure 7-2: Slice sampling step 1

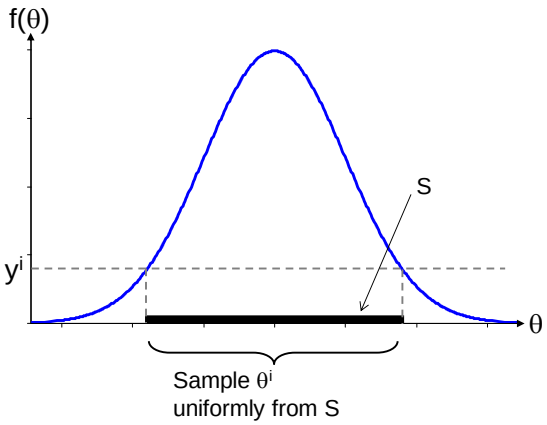


Figure 7-3: Slice sampling step 2

7.1.6 Overview of Applications

An advantage of using Bayesian methods in process modelling applications is that the width of the posterior distribution captures the uncertainty both in estimating the model parameters and the variation inherent within the process. Therefore when a model is used to predict a critical quality attribute of a process, the posterior distribution will show the full range of values that the CQA may be expected to take. If there is high uncertainty in estimating the parameters, then the posterior distribution will be wide, indicating that more information may be needed to predict the true range of the CQA (Section 7.3). Additionally Bayesian methods can be applied to a mechanistic model of a process, when they are some parameters to be estimated. By sampling from the posterior distributions of the parameters and applying the mechanistic model with the sampled values, a posterior distribution for the model output is generated, which reflects the certainty of the model parameter estimates (Section 7.3.2.2 and Section 7.3.2.3).

Bayesian methods have been applied to many areas of statistical inference, including linear regression analysis, hierarchical modelling and time series analysis (Lee, 2012, Congdon, 2003). Hierarchical models are applicable for modelling data from different groups by allowing specified parameters to vary for each group. For example in the

social sciences, hierarchical models have been used to model data from surveys, by allowing the comparison of parameters from different countries or regional areas (Jackman, 2009). Additionally, in epidemiology, a binomial Bayesian analysis has been applied to model the spread of disease in different demographic groups (Geweke, 1989).

Bayesian methods have also been applied in the analysis of clinical trials (Berry, 2006), allowing data to be analysed as soon as it is collected during the trial. The results can influence the design of the remainder of the trial, for example to favour better performing therapies or to target patient groups that appear to respond better to the treatment. In statistical process control (SPC), Bayesian methods have been developed to improve upon the standard Stewhart chart (Section 2.2). For example Colosimo and del Castillo (2007) developed a Bayesian SPC methodology that quantifies the posterior probability that a problem has occurred and also accounts for known drift in a process that is not attributed to the problem. In general, Bayesian methods allow for a more flexible model structure to be developed than classical methods, so the most appropriate model can be found to fit the data that is available.

In the subsequent sections (7.2 to 7.4), examples are presented how the theory of Bayesian statistics and Markov Chain Monte Carlo has been applied in the development of pharmaceutical manufacturing processes.

7.2 Bayesian methods in experimental design

In pharmaceutical process development, experimental design is used to gain information about a process, such as assessing how the variation of the inputs to the process can affect the outputs. For example, an experiment may be planned to gain knowledge of which are the most important factors and then to optimise the operating space or to make predictions from the process data. When a new process is developed using the Quality by Design framework, data from experimental studies is used to develop a design space in which the process will be validated and run (Section 2.3.2). Good experimental design is needed to maximise the information that is gained so that a robust process can be developed (Lunney et al, 2008, van de Ven *et al*, 2011).

Bayesian methods allow previously known information to be included into the design and analysis of an experimental program. When designing an experiment, prior knowledge usually exists about the product or process being studied (Chaloner and Verdinelli, 1995). For example, prior knowledge may exist from data from previous experimental work or mechanistic models of the process being studied (Lunney et al, 2008). The use of Bayesian methods allows the prior information to be quantified and

used formally to develop an experimental design (Dubé et al, 1996). Prior knowledge can be used to determine a number of features of the experimental design, including which factors should be included, the levels at which each factors will be tested and to provide an indication of the levels of quadratic and interaction terms that should be estimated.

The number of experimental runs can be a limiting factor in experimental design. Bayesian designs provide more flexibility than standard fractional factorial designs, allowing the number of runs to be reduced while maximising the information that can be gained (Nabifar *et al*, 2011). Using Bayesian analysis, an experimental design can be run sequentially, with the prior information updated as each set of results is collected. Then the design for the subsequent runs is updated each time more results become available (van de Ven *et al*, 2011). When experiments are being run to gain information about a process, this approach allows the design space of the experiment to move so that the focus shifts towards those factors that appear to be the most important. In a process optimisation context, a sequential experiment can be moved towards the optimal region of the process, to gain more information about the space in which the process will be run.

7.2.1 Bayesian modification of D-optimal designs

In the context of process development, the objective of an experimental program may be to estimate how a set of factor variables, or inputs to a process, will impact on the final product or outcome of a process, such as the yield. To create an optimal design, one of the goals is to maximise the information that is expected to be gained from the resulting data. Since the outcome of an experiment cannot be known at the design stage, the information that will be gained can be estimated from the prior information and the experimental design (Chaloner and Verdinelli, 1995).

A method to quantify the information that is gained from an experiment is to consider the variance of the model parameters. The optimal design will be selected to minimise the parameter variation and hence to maximise the certainty with which the parameters are estimated. Model parameter variance is measured in terms of the determinant of the matrix $[\mathbf{X}^T\mathbf{X}]^{-1}$, where $\mathbf{X}$ is the design matrix (Eriksson *et al*, 2008). The design that minimises $\det[\mathbf{X}^T\mathbf{X}]^{-1}$ is termed a D-optimal design.

When Bayesian methods are used to analyse the resulting data, the prior information is combined with the information gained from experiments (likelihood). Therefore the objective becomes to minimise the expected posterior variance of the model

parameters, known as a Bayesian D-optimal design (Dubé *et al*, 1996, DuMouchel and Jones, 1994).

When prior information is obtained from a mechanistic model or from existing data, the prior can be specified as a multivariate normal distribution of the vector of parameters to be estimated, $\boldsymbol{\theta}$ (Dubé *et al*, 1996):

$\boldsymbol{\theta} \sim \text{mvN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known.

A linear regression model is the fitted to the data:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim \text{mvN}(\mathbf{0}, \mathbf{I}\sigma^2), \quad \text{Equation 7-8}$$

where σ^2 is the variance of the response and is assumed to be known. The posterior variance of $\boldsymbol{\theta}|\mathbf{y}$ is $[\boldsymbol{\Sigma}^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{X}]^{-1}$. Minimising the posterior variance is equivalent to selecting the Bayesian D-optimal to maximise the determinant of $[\mathbf{I} + \sigma^{-2}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T]$ (Nabifar *et al*, 2010).

An alternative method for capturing the prior information was proposed by DuMouchel and Jones (2004) that is appropriate when there is more certainty about some model parameters than others. Suppose there are p primary terms that will be included in the model and q potential terms that are less likely be important, including high order terms such as quadratic and interaction terms. As before the prior information is captured in the form $\boldsymbol{\theta} \sim \text{mvN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The primary terms will have a wide prior distribution because the model coefficients are not expected to be zero. However the potential terms have a multivariate normal distribution of the form $\text{N}(\mathbf{0}, \tau^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and τ is set to one if the data has been scaled to unit variance.

To determine the posterior distribution, let $\mathbf{K}$ be a (p+q; p+q) matrix with ones for the last q elements on the main diagonal and zero otherwise:

$$K = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & & 0 & \vdots \\ \vdots & & 0 & & & \vdots \\ \vdots & & & 1 & 0 & \vdots \\ \vdots & 0 & & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{bmatrix} \begin{matrix} \} p \\ \\ \\ \} q \end{matrix}$$

$$\underbrace{\qquad}_{p} \quad \underbrace{\qquad}_{q}$$

The posterior covariance matrix is given by $[\mathbf{X}^T\mathbf{X} + \mathbf{K}/\tau^2]^{-1}$ and hence the Bayesian D-Optimal design is selected to maximise $\det[\mathbf{X}^T\mathbf{X} + \mathbf{K}/\tau^2]$. This method is appropriate if

the number of experiments to be run is between p and p+q and hence it is not possible to estimate each factor (Jones *et al*, 2008).

7.2.2 Maximising the prediction accuracy

Van de Ven *et al* (2011) suggested that the Bayesian D-optimal approach can result in more experiments being run than are necessary because the focus is on estimating unknown parameters, as opposed to identifying the optimal design space of a process. As an alternative approach, van de Ven *et al* (2011) proposed that the aim should be to minimise the variation of the predictions, rather than the variation of the model parameter estimates. Adopting a sequential design approach, the design can be focused on the prediction accuracy for the region in which the process is most likely to be run. This approach, named the Iw criterion, avoids running unnecessary experiments to improve predictions for regions in which that process will not be run.

For a particular point in the design space, **x**, the prediction variation is measured by:

$$\mathbf{x}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{x}^{\mathrm{T}},$$

where **X** is the design matrix.

Over an experimental region, **R**, the average prediction variance can be found by dividing **R** into a grid, G(**R**), and finding the prediction variance at each point in G(**R**). The average variance across the experimental region is known as the I-criterion:

$$\mathrm{I}(\mathbf{X}) = \sum_{\mathbf{x}\epsilon\mathrm{G}(\mathbf{R})} \mathbf{x}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{x}^{\mathbf{T}} \qquad \textbf{Equation 7-9}$$

When prior knowledge exists about the region within R in which the process is most likely to be run, I(**X**) can be extended to a weighted sum, where the weights, w(**x**), are selected to be larger in the region of interest:

$$\mathrm{Iw}(\mathbf{X}) = \sum_{\mathbf{x}\epsilon\mathrm{G}(\mathbf{R})} \mathrm{w}(\mathbf{x})\mathbf{x}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}^{\mathbf{T}} \qquad \textbf{Equation 7-10}$$

For a sequential design, the weights can be updated as more information is generated. By selecting a design to minimise Iw(**X**), the experiment will be designed to provide the most accurate predictions in the region in which the process is expected to be run, and hence the properties of the final registered design space can be predicted will a high level of accuracy.

7.2.3 Implementation of Bayesian experimental Design

The main advantage of using a Bayesian approach to experimental design is that prior information can be used formally rather than subjectively, allowing the experimental design to be optimised to gain the most information from the fewest number of experimental runs. Bayesian D-optimal designs are thought to be more powerful than traditional D-optimal designs for detecting non-linear effects such as quadratic and interaction terms (Lunney *et al*, 2008).

When there is limited prior information about a process, it can be useful to run experiments in sequential blocks, so that information from one block can be used to develop the design for the next block. From the first block of experiments, the prior and likelihood are combined to form a posterior distribution for the parameters being estimated. This posterior can then be carried over and used as the prior for the next block (Dubé et al, 1996). This approach allows the focus of the design to evolve to concentrate on the parameters and input variables that are found to be the most important in the initial blocks of experiments.

Dubé et al (1996) applied a sequential Bayesian D-optimal design to a multi-component polymerisation process. The aim was to gain an understanding of how seven process variables affected the production rate and product quality. Initial prior information was gained from the running of a mechanistic model of the process and combining the output with knowledge from process experts to estimate the size of each factor to be modelled. The prior variance was estimated based on data from previous experimental work. The experiment was run in three blocks consisting of four, four and ten experimental runs respectively. After each block of experiments was run, the posterior distribution was calculated and used as the new prior distribution for the next block. Experiments were run until the posterior variability of the model parameters was considered to be suitably small and the response parameters could be estimated with the required level of certainty, allowing for a detailed study of the polymerisation kinetics of the process.

A similar example was described in Vivaldo-Lima *et al* (2006) to determine the important process variables that influence the particle size in a polymerisation process. Six process variables were studied to determine their effect on two outputs: mean particle size and variation of particle size. A mechanistic model was used to generate a prior distribution, in the form of a multivariate normal model of the variables. The aim of the experimental work was to improve upon existing knowledge of the process, so a limited design of two sets of four runs was implemented, with the prior information for the second set updated with the posterior information from the results of the first set of

runs. The results highlighted the importance of the mixing process in controlling the particle size and allowed existing mechanistic models to be updated.

7.3 Identifying a design space

Bayesian methodologies have been found to be useful when applied to experimental designs that are used to identify the design space of a process. Following the collection of data, Bayesian methods can be utilised to analyse the data and identify a robust design space, in which the process can be run with an acceptably low risk of failing the relevant quality requirements. Typically, the data from experimental results is used to create a response surface model of the process that will allow the prediction of the expected outputs for a given set of inputs. Ranges of the input variables can then be determined for which the critical quality attributes of the process are predicted to fall within their specification limits. This region is defined as the design space of the process and is registered with the regulatory authorities. Ideally the process will be run within a subset of the design space: the normal operating range.

The response surface model will predict the mean output for any point in the design space, but it does not take into account the variation inherent in the process and the uncertainty associated with estimating the model parameters. This approach can lead to a design space being identified in which the mean predicted response is favourable, but high variation results in a low probability of achieving the quality conditions for every batch. A more robust approach is to define the design space based on the probability that a favourable outcome will be achieved (Peterson, 2004).

Through the use of Bayesian modelling, a posterior predictive distribution (PPD) can be calculated for the response variables; the width of the PPD reflects the variation in the process and the uncertainty associated with estimating the model parameters (Section 7.3.1.1 and 7.3.1.2). Monte Carlo methods can be used to simulate the response for a particular set of input settings, to calculate the probability, or reliability, that the process outputs will all meet their desired quality conditions (Section 7.3.1.3). By determining the reliability at multiple points within the potential operating space, a reliability surface is created for the process, thus allowing for the identification of a design space that is robust to the variation in the process (Section 7.3.1.4).

In some cases, wide variation in the PPD can result in the estimated reliability being too low to identify a suitable design space. Pre-posterior analysis can be applied to determine whether the variation is inherent in the process or is in the estimates of the parameters and thus the reliability could be improved by running further experiments

(Section 7.3.1.5). A number of applications of the use of Bayesian models within the quality by design framework are discussed in Section 7.3.2

7.3.1 Methodology for the posterior predictive approach

Peterson (2004) presented a methodology for implementing Bayesian modelling to identify the optimal design space to validate and run a process. A model of the process is created that combines prior information and experimental data to generate a posterior predictive distribution for the process outputs. The resulting model is then used to estimate the reliability that a set of quality conditions will be met, for a given set of inputs.

7.3.1.1 Bayesian reliability

Let $\mathbf{Y}=(Y_1,..., Y_p)^T$ be a vector of response variables that must satisfy specified quality criteria. A set of input variables, such as temperatures or raw material characteristics, is denoted $\mathbf{x}=(x_1,..., x_k)^T$. The matrix **A** is the acceptance region, which is a set of quality conditions that must be satisfied for each of the responses, with probability Q. For example, **A** can be a (px2) matrix of upper and lower specification limits. The design space of the process is defined as:

$\{\mathbf{x}: P(\mathbf{Y} \epsilon \mathbf{A}|\mathbf{x}, \text{data}) \geq Q\}$, for a predefined value Q.

The design space is the set of input factor settings that is expected to produce response variables that are within the acceptance region, with a suitably high probability. The value of Q is selected to define the required reliability of the process. The required reliability is not a set value and should be defined using a risk based approach, with consideration given to the phase of product development, complexity of the process, required performance and intended use of the product (Stockdale and Cheng, 2009). Additionally, the financial cost of a failure may be taken into consideration (Mockus *et al*, 2011a).

The probability of the quality conditions being met for a given set of input variables is denoted $p(\mathbf{x})$:

$$p(\mathbf{x})=P(\mathbf{Y} \epsilon \mathbf{A}|\mathbf{x})$$

Equation 7-11

where **x** is a vector of settings for each input and $p(\mathbf{x})$ is termed the Bayesian reliability. To indicate whether or not the quality conditions have been met, the discrete desirability function $I(\mathbf{Y} \epsilon \mathbf{A})$ can be used, where $I(\mathbf{Y} \epsilon \mathbf{A})=1$ if the conditions are met, or zero otherwise. Samples are taken from the posterior predictive distribution of $\mathbf{Y}|\mathbf{x}$ to

calculate I(**Y**ϵ**A**|**x**) for each sample. Then the average of I(**Y**ϵ**A**|**x**) is calculated to give p(**x**).

7.3.1.2 Regression model

To estimate the value of p(**x**), a regression model is built to identify a relationship between the input variables, **x**, and the response variables **y** (Peterson, 2004):

$$\mathbf{y} = \mathbf{B}\mathbf{z}(\mathbf{x}) + \mathbf{e}$$

Equation 7-12

where **B** is a (pxq) matrix of regression coefficients, for the q model terms and p response variables. **z(x)** is a vector valued function of **x** that is used to create the terms in the model. For example if a quadratic model is being used, **z**=(1, x_1, ..., x_k, x_1^2, ..., x_k^2). **e** is the error vector that follows a multivariate normal distribution with zero mean and covariance matrix ∑. Data obtained from experimental work is used to estimate **B** and ∑.

For the calculation of the posterior predictive density, f(**y**|**x**,data), a prior distribution must be specified for the model parameters **B** and ∑. If suitable information exits, an informative prior can be defined, for example from results of previous experimental work or a mechanistic model of the process (Peterson, 2008). Alternatively a non-informative joint prior can be specified so that **B** and ∑ are proportional to $|\sum|^{(p+1)/2}$.

Data from the prior distribution is combined with data from experimental work to determine the posterior distribution, f(**B,**∑|**x**). Since an analytical solution may be difficult to obtain, Monte Carlo methods are used to simulate the posterior distribution. These samples are then used applied to Equation 7-12 to obtain a posterior predictive distribution for the process outputs: f(**y**|**x**).

7.3.1.3 Simulating the Bayesian reliability

Using Monte Carlo methods, the posterior distribution of the responses can be simulated for a given set of inputs, **x**. By generating N samples of the response vector, $\mathbf{y_1}$, $\mathbf{y_2}$, …, $\mathbf{y_N}$, the Bayesian reliability is estimated as:

$$p(\mathbf{x}) \sim \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{y_i} \in \mathbf{A}|\mathbf{x})$$

Equation 7-13

where **A** is the acceptance region and I(**Y**ϵ**A**) is the discrete desirability function described in Section 7.3.1.1. The value of p(**x**) gives the probability, or Bayesian reliability, that a selected point in the design space will produce a response that will satisfy the quality conditions.

7.3.1.4 Identifying the optimal region

Calculation of the optimal design space in which to operate the process is achieved by simulating the value of p(**x**), the Bayesian reliability, across multiple points of the potential operating space. Comparison of the p(**x**) values will determine which combinations of input settings are expected to produce a response that satisfies the quality conditions, with a suitably high reliability (Stockdale and Cheng, 2009). The optimal region is the region in which p(**x**) is greater than the required reliability.

When there are up to three controllable factors, values of p(**x**) can be determined for a grid across the experimental region and the reliability visualised to identify the highest reliability. However, when there are a large number of factors, significant time and computational power will be required to cover the experimental region. An alternative method is to create a regression model for p(**x**) based on the process inputs, to represent the reliability space (Peterson *et al*, 2009). A number of points can be simulated to create a factorial design and then a regression model is fitted to estimate p(**x**) across the experimental region. Finally an optimisation procedure is applied to identify the region that has the greatest reliability.

7.3.1.5 Pre-posterior analysis

The variation in the posterior predictive distribution (PPD) has two potential sources: variation that is inherent within the process and uncertainty in the model parameters that are estimated (Peterson *et al*, 2009). When aiming to identify a suitable design space, the mean predicted response may be suitable but the reliability may be too low due to large variation in the posterior predictive distribution of the response, i.e. for a particular response some of the PPD lies outside of a specification limit.

In this case it is useful to determine whether the high variation is due to the process or uncertainty in the model parameters. The results will indicate whether future work should be focused on understanding and reducing the process variation, or collecting more data to improve the estimates of the model parameters. If the parameters are highly uncertain, then the calculated PPD for the responses will be much wider than the actual variation that is seen in the process. Increasing the certainty of parameter estimates will reduce the width of the PPD and potentially increase the calculated reliability.

Pre-posterior analysis is used to determine whether the variation in the posterior predictive distribution could be reduced by collecting more data and hence improve the estimate of the reliability that a process will meet the required quality conditions

(Peterson, 2004). The impact of collecting more data can be assessed by simulating new data that will result in the same values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ being estimated (Gilmour and Mead, 1995). The new and existing data is then combined and the posterior predictive distribution of the response and the Bayesian reliability of meeting the quality conditions is determined. The results are then compared to the original analysis to determine whether the reliability has improved with the addition of simulated data.

7.3.2 Examples of applications of Bayesian modelling in process development

7.3.2.1 Early phase synthetic chemistry

Peterson (2008) presented an application of the Bayesian reliability method in early phase synthetic chemistry. A process under development for an API product was investigated to determine the operating conditions that would reduce the level of impurities that formed during the manufacturing process. Four input factors were studied: temperature, pressure, catalyst loading and reaction time. A 2^4 factorial design with eight axial points and six centre points was used to assess their effect on four response variables relating to the impurity levels following a reaction. Each of the response variables had an upper or lower specification limit.

A response surface model was developed following the application of a logit transformation to the response data to improve the distribution of the residuals. The response surface was modelled as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Equation 7-14

where $\mathbf{Y}$ = (logit(Y_1), logit(Y_2), logit(Y_3), logit(Y_4))^T, $\mathbf{X}$ = diag($\mathbf{m_1}$, $\mathbf{m_2}$, $\mathbf{m_3}$, $\mathbf{m_4}$) and each $\mathbf{m_i}$ contained the model terms for the response Y_i. The model terms were a combination of linear, quadratic and first order interactions of the input variables. The vector $\boldsymbol{\beta}$ contained the coefficients for the model terms. The residuals follow a multivariate normal distribution, i.e. $\mathbf{e}$ ~ N($\mathbf{0}$, $\boldsymbol{\Sigma}$).

A non-informative prior of the form:

$$f(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(r+1)/2}$$

Equation 7-15

was used, where r is the number of responses, i.e. four. The prior information was combined with the experimental results using Monte Carlo simulation to determine the posterior distribution of ($\boldsymbol{\beta}$,$\boldsymbol{\Sigma}$|$\mathbf{x}$).

For a given set of values for the input variables, the posterior predictive response of Equation 7-14 was simulated by sampling from the posterior distribution of ($\boldsymbol{\beta}$,$\boldsymbol{\Sigma}$|$\mathbf{x}$). For each response, the posterior predictive distribution defined the probability of satisfying the specification limits at that particular point in the operating space. Calculating the product of the results from each response gave the resulting Bayesian reliability for the whole process.

By simulating the response across a grid of the experimental region, the Bayesian reliability could be assessed across the potential operating space. Contour plots were used for a visual assessment of how the reliability varied when the input settings were changed (Figure 7-4). The plots show the Bayesian reliability as the four input variables, temperature, pressure, catalyst loading and reaction time are varied. The results suggested that the optimal operating conditions were found when the temperature and pressure were low, the reaction time was high and the catalyst loading was in the centre of the experimental range.

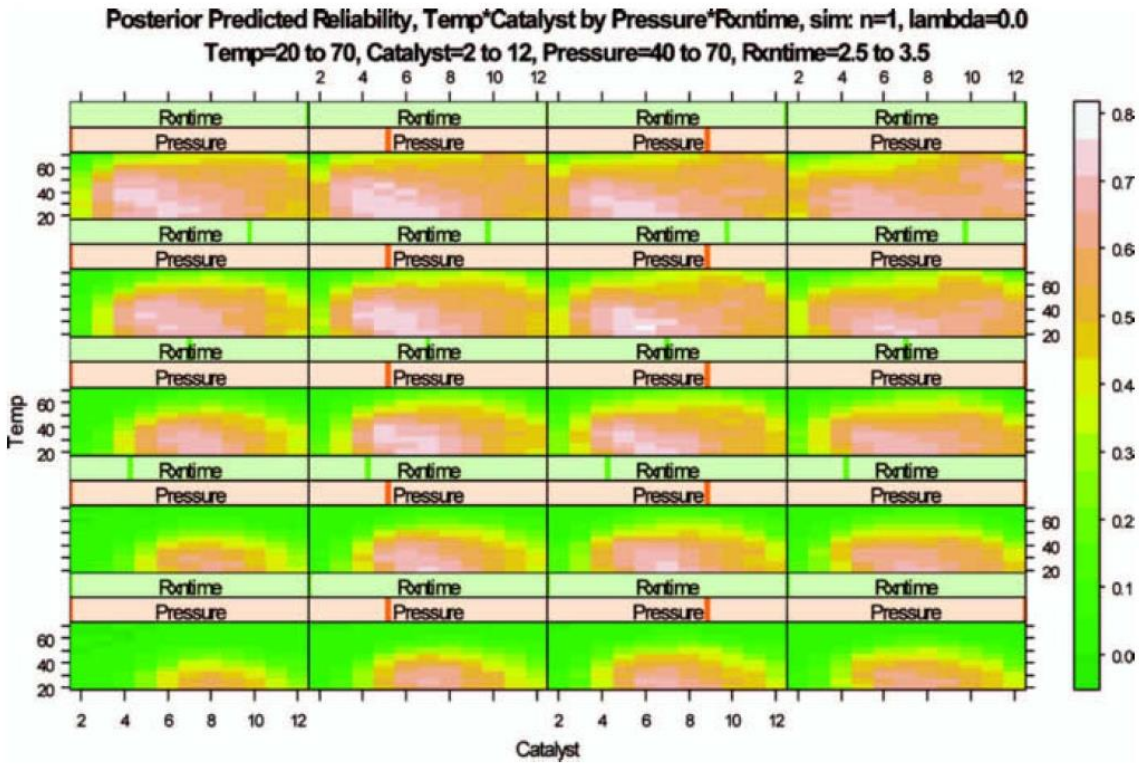


Figure 7-4: Contour plot of Bayesian reliability (Peterson, 2008, Figure 2)

The maximum reliability was found to be 71%, suggesting that the process may not be capable of consistently meeting the specification limits for the impurity levels. Stockdale and Cheng (2009) extended the work of Peterson (2008) and applied pre-posterior analysis to this example. It was found that by adding an additional four replicates to the original 30 experimental runs, the variation in the model parameters could be reduced

such that the maximum reliability could be as large as 83%. Additionally, it was suggested that if the variation in the process could be reduced by 30%, the residual error ($\boldsymbol{\Sigma}$) would be reduced and the reliability could be as high as 98%. The results of the study suggested that a suitable design space for a capable process could be achieved if additional experimental work was undertaken and the variation in the process reduced.

7.3.2.2 Quality by Design in a tablet manufacturing process

When a mechanistic model is used to represent part of a process, a number of parameters within the model may be estimated from experimental data. By applying Bayesian modelling, the uncertainty in estimating the model parameters can be input into a mechanistic model, to produce a posterior distribution for the model response.

Mockus *et al* (2011a) applied a Bayesian approach when implementing Quality by Design to a tablet manufacturing process, with the aim of controlling tablet hardness and the formation of an impurity over the product shelf life. A number of process variables were investigated from the wet granulation, drying, blending and tabletting stages of the process, to determine their effect on the CQAs: tablet hardness and degradation that results in the impurity forming over time. Process variables included compression force, median particle size, dryness and bulk density.

A Bayesian linear model was used to predict the tablet hardness:

$$\log(\text{Tablet hardness}) = \mathbf{X\beta}$$

Equation 7-16

where **X** contains the process variables: compression force, median particle size, bulk density and the particle size*bulk density interaction, and **β** is a vector of the model parameters. Posterior distributions for the model parameters were estimated using Monte Carlo methods in WinBUGS (Lunn *et al*, 2000).

A mechanistic model was used to predict the degradation over time:

$$L_t = L_0 + k_1 V_0 (1 - e^{-k_2 t})$$

Equation 7-17

where L_t is the level of impurity at time t, L_0 is the initial impurity level and V_0 is the rate of degradation. Posterior distributions for the parameters k_1 and k_2 were estimated from experimental data. For each batch, the values of $\log(L_0)$ and $\log(V_0)$ were estimated using linear models of the process variables:

$$\log(L_0) = b_0 + b_1\,\text{dryness} + b_2\,\text{particle size} + b_3\,\text{bulk density}$$ Equation 7-18

$$\log(V_0) = c_0 + c_1\,\text{compression force} + c_2\,\text{particle size} + c_3\,\text{bulk density}$$ Equation 7-19

Again, the posterior distributions for the parameters b_0 to b_3 and c_0 to c_3 were estimated from experimental data.

For an assessment of the design space of the process, the tablet hardness and degradation were estimated for various values of the process variables. Samples were generated from the posterior distributions of the model parameters, first to predict $\log(L_0)$ and $\log(V_0)$ in Equation 7-18 and Equation 7-19, and then the results were carried through to Equation 7-16 to determine the posterior predictive distribution of degradation. Finally the posterior distributions for degradation and tablet hardness were used to calculate the Bayesian reliability of meeting the required specification limits. Contour plots were then used to visualise how the reliability varied across the potential operating space. From Figure 7-5 it is suggested that the lowest risk of failure could be found when the compression force and dryness are high, the bulk density low and the particle size in the middle of the range.

7.3.2.3 Estimation of the distribution of the drying phase duration

As part of a Quality by Design study for a lyophilisation process, Mockus *et al* (2011b) used Bayesian methods to estimate the parameters in a mechanistic model, with the aim of estimating the distribution of the drying phase duration. The case study relates to the development of the manufacture of a parenteral product that requires a low moisture content. During the lyophilisation process, the product is frozen, ice is sublimed to form a cake and the product cake undergoes desorption. The primary drying phase occurs during sublimation and comprises the majority of the lyophilisation time, in the order of days.

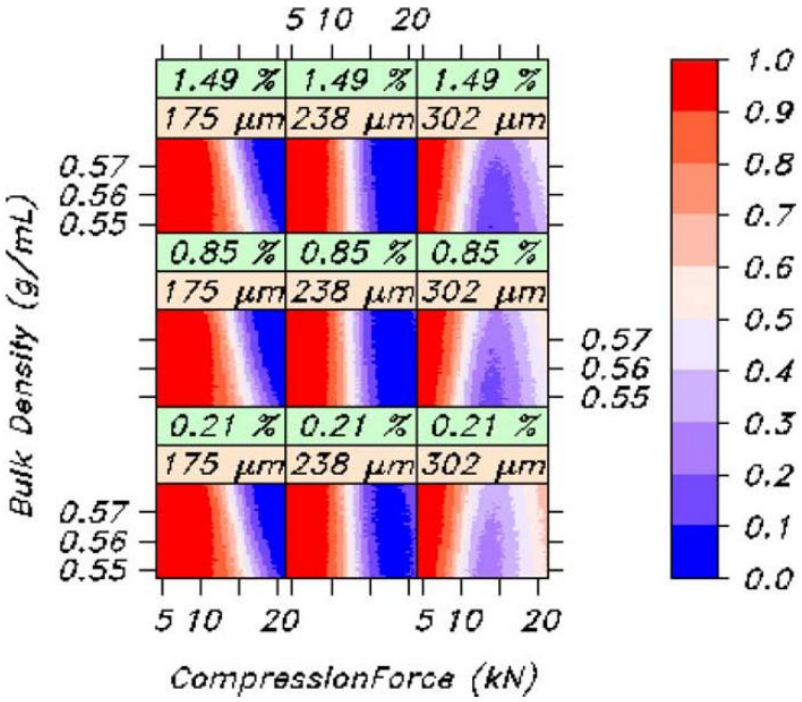


Figure 7-5: Contour plot of risk of failure, from Mockus *et al* (2011a), Figure 12

The drying time can be predicted from a mechanistic model which requires the average resistance as an input. The average resistance (RP) can estimated from the nucleation temperature, T_n:

$$RP = \alpha + \beta T_n + \varepsilon \qquad \text{for } \varepsilon \sim N(0, \sigma_{RP}^2)$$

Equation 7-20

A prior distribution was defined for T_n, where $T_n \sim N(\mu_T, \sigma_T^2)$. The values of μ_T and σ_T^2 were estimated from data taken from 11 batches. Non-informative prior distributions were used for α, β and σ_{RP}^2. By combining the data of the average resistance and nucleation temperature with the prior distributions, posterior distributions were found for the model parameters in Equation 7-20. Sampling from these posterior distributions produced a posterior predictive distribution for the average resistance. Then by taking samples from the PPD of the average resistance as inputs into the mechanistic model, a posterior predictive distribution was found for the drying time.

This approach allowed the distribution of the drying time to be estimated from a heat and mass transfer model, which was expected to be scalable. Therefore it will be possible to estimate the distribution of the drying time as the process is scaled up. This method also allowed the inputs into the mechanistic models to be varied, to determine their importance on the variation in drying times.

7.4 Process Capability

Process capability indices (PCI) have been found to be a useful tool to represent how well data from a process falls within a set of specification limits and to highlight the risk of a failure (Chapter 6). When a sample of data is taken from a process, the calculated PCI provides a point estimate of the capability, but does not give any indication of the certainty of the capability estimate. To have some certainty that a process is capable, it is necessary to look at the range of values in which the true capability could lie. In particular, the lower limit of the true capability will indicate how low the capability could actually be in a process. A process may not be considered to be capable unless the whole range of values in which the true capability could lie is above the minimum value that is required for a process to be capable.

The process capability indices presented in Chapter 6 are based on frequentist methods, so it is assumed that the mean and standard deviation of the process are fixed quantities. However processes are not expected to remain constant over time and so a Bayesian approach may be more appropriate. In a Bayesian context, it is assumed that the mean and standard deviation are stochastic variables, each with an associated probability distribution (Cheng and Spiring 1989).

Confidence intervals on the P_{pk} metric can be found but are difficult to compute because the interval will depend on the distributions of both the mean and variance (Kotz and Johnson, 1993). However when Bayesian methods are applied a posterior distribution is found for the PCI, from which the percentiles can be derived. The lower percentile of the posterior distribution will give the lowest value that the capability is expected to take and if this value is suitably large, there is high certainty that the process is capable. This approach links to the methodology used to identify a design space (Section 7.3), since the focus is given to assessing the reliability that the process is capable, rather than finding the average capability estimate.

7.4.1 Methodology

In a Bayesian approach, the posterior predictive distribution of the process capability index (PCI) can be determined (Cheng and Spiring, 1989). The minimum required capability is denoted by w, for example w=1.33. For the process to be capable in a Bayesian context, it is required that the PCI is greater than w, with some probability p, for example p=0.95:

$P(PCI>w \mid data)>p$

i.e. for p=0.95, 95% of the posterior distribution is required to be greater than w, then with 95% certainty the process is capable. The width of the posterior distribution will depend on the sample size n, so when the sample size is larger, the posterior distribution will be narrower and hence there will be greater certainty that the process is capable.

7.4.1.1 Method for P_p

Cheng and Spiring (1989) presented a Bayesian approach to determining the process capability when the process is assumed to be centred and the capability is measured by P_p. To calculate P_p, only the standard deviation, σ, needs to be estimated. For a sample of data, $\mathbf{x}=(x_1, x_2, \ldots x_n)$, that is assumed to come from a normal distribution, the likelihood of the data is:

$$f(\mathbf{x}|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} * \exp\left\{-\sum(x_i-\mu)^2/2\sigma^2\right\} \qquad \text{Equation 7-21}$$

A non-informative prior distribution was used for μ and σ^2:

$$f(\mu,\sigma^2) = \frac{1}{\sigma}\text{, for } -\infty < \mu < \infty, 0 < \sigma < \infty \qquad \text{Equation 7-22}$$

The prior is chosen to maximise the difference between the information provided by the prior and posterior distributions, therefore minimising the prior information (Shaiu *et al*, 1999). Combining the prior and likelihood distributions gives the posterior distribution. The marginal posterior distribution for σ given the data is:

$$f(\sigma|X) = 2\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1}\left[\frac{(n-1)S^2}{2}\right]^{\frac{n-1}{2}}\sigma^{-n}\,exp\left\{-\frac{(n-1)S^2}{2\sigma^2}\right\} \qquad \text{Equation 7-23}$$

A process is considered to be capable if:

$$P_p > w \Rightarrow \frac{USL-LSL}{6\sigma} > w \Rightarrow \frac{USL-LSL}{6w} > \sigma$$

So the posterior probability that a process is capable is found by:

$$p = P(P_p > w|X) = P\left(\sigma < \frac{USL-LSL}{6w}\middle|X\right) = \int_0^{(USL-LSL)/6w} f(\sigma|X)d\sigma \qquad \text{Equation 7-24}$$

Since the width of the posterior distribution depends on n, for given values of n, w and p, the minimum required value of P_p can be found, such that p% of the posterior distribution is greater than w, which is denoted C*(p). Then if a calculated value of P_p is greater than C*(p), the process can be considered to be capable in a Bayesian sense.

Examples of C*(p) values presented in Cheng and Spriring (1989) are shown in Table 7-2. For example, for a sample size of 50, the calculated P_p value must be greater than 1.2 for 95% of the posterior distribution to be greater than one.

n	C*(0.95)
10	1.65
20	1.37
30	1.28
40	1.23
50	1.20

Table 7-2: Values of C*(0.95), w=1

7.4.1.2 Method for P_{pk}

Pearn and Wu (2005) extended the above method to be applicable to P_{pk}, by finding the joint posterior distribution for μ and σ^2.

To find $P(P_{pk}>w \mid x)$, first note that:

$$\min\{USL-\mu, \mu-LSL\} = d - |\mu-m|$$

Equation 7-25

where d=1/2(USL-LSL) is half of the specification width, and m=1/2(USL+LSL), is the mid-point of the specification range. Then the posterior probability that the process is capable is given by:

$$\begin{aligned} p = P(P_{pk} > w|x) &= P\left(\frac{d - |\mu - m|}{3\sigma} > w \middle| x\right) \\ &= P(|\mu - m| < d - 3\sigma w|x) \\ &= \int_0^{\infty} \int_{m-d+3\sigma w}^{m+d-3\sigma w} f(\mu, \sigma|x) d\mu \, d\sigma \end{aligned}$$

Equation 7-26

For given values of n, p, w and $\delta = \frac{|\bar{x}-m|}{s}$, the minimum required value of P_{pk} can be found, denoted C*(p), examples of which are shown in Table 7-3. As a result it is straightforward to determine whether a process is capable in a Bayesian sense, by comparing P_{pk} to C*(p).

7.4.1.3 Illustrative example

An illustration of how the above method works is now presented. Two processes are compared with P_{pk} values greater than 1.33. In each case, n=50, p=0.95, δ=0 and w=1.33 and hence from Table 7-3, C*(0.95) = 1.67, i.e. the calculated P_{pk} value must be greater than 1.67 for the process to be considered capable.

	δ				
n	**0**	**0.5**	**1**	**1.5**	**2**
10	2.47	2.33	2.27	2.25	2.23
20	1.98	1.88	1.85	1.85	1.85
30	1.81	1.74	1.72	1.72	1.72
40	1.72	1.67	1.65	1.65	1.65
50	1.67	1.62	1.61	1.61	1.61

Table 7-3: Values of C*(0.95) for w=1.33, taken from Pearn and Wu (2005)

Figure 7-6 shows the two posterior distributions for P_{pk}, where the calculated P_{pk} values are 1.7 and 1.5 respectively. In Figure 7-6a, less than 5% of the posterior distribution is below 1.33, so there is greater than 95% certainty that the process is capable. In addition, the calculated value of P_{pk} is greater than C*(p) and hence the process is capable in a Bayesian context. Conversely in Figure 7-6b, 21% of the posterior distribution is below 1.33 and the value of P_{pk} is less than C*(p), hence there is not enough certainty that the P_{pk} value is greater than 1.33. Therefore, even though the P_{pk} value is 1.5, the process is not considered capable in a Bayesian context.

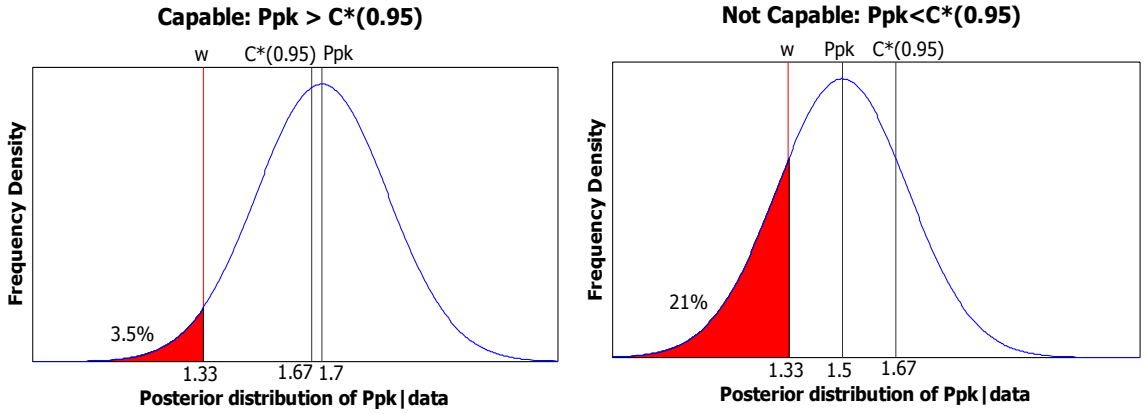


Figure 7-6 (a, b): Examples of posterior distributions for P_{pk}|data

7.4.2 Application to process data

A number of process variables have been selected to assess the results of the Bayesian approach to process capability (Figure 7-7 to Figure 7-11). These particular variables were selected because their distributions are all similar to the normal distribution. For each variable, the value of P_{pk} was compared to values of C*(95), for w=1, 1.3, 1.5 and 2, taken from Pearn and Wu (2005) and Table 7-4. In each case a sample size of 50 was used.

Variables A and B have very high P_{pk} values, greater than two, and there is high certainty that the distribution of P_{pk} is greater than 1.5 (Table 7-4), so these variables are considered to be highly capable. Variables C and D also have high P_{pk} values,

greater than 1.5. However there is less than 95% certainly that P_{pk} is greater than 1.33, suggesting the capability is high but may not have reached the target of 1.33. The P_{pk} value for variable E is 1.07, suggesting the process is only just capable. Additionally, there is less than 95% certainty that P_{pk} is greater than one, suggesting that the capability needs to be improved.

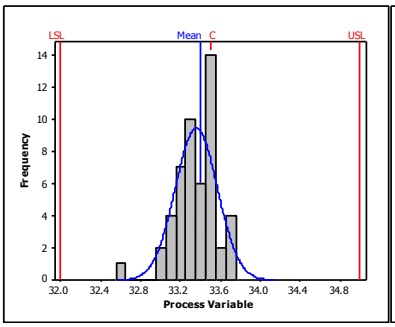


Figure 7-7: Variable A

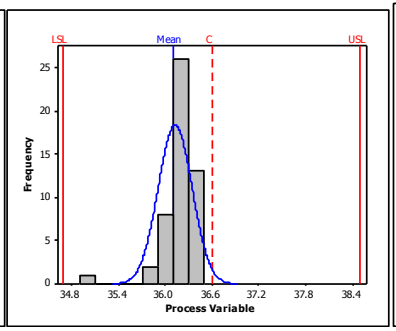


Figure 7-8: Variable B

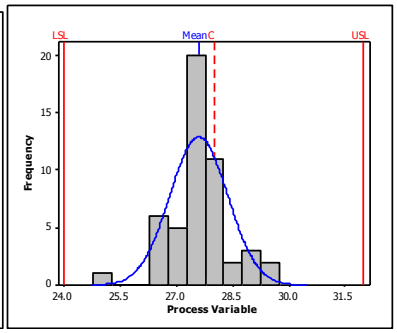


Figure 7-9: Variable C

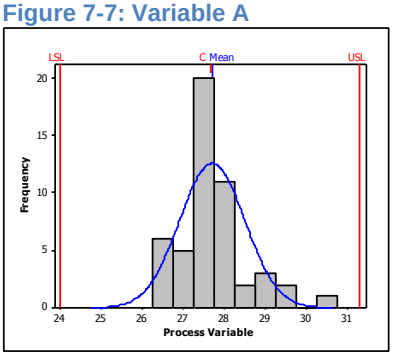


Figure 7-10: Variable D

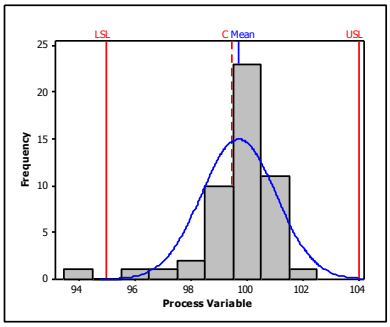


Figure 7-11: Variable E

Variable	P_{pk}	δ	w			
			1	1.33	1.5	2
A	2.14	0.68	✓ 1.22	✓ 1.62	✓ 1.82	× 2.41
B	2.19	2.19	✓ 1.22	✓ 1.62	✓ 1.82	× 2.41
C	1.54	0.54	✓ 1.22	× 1.62	× 1.82	× 2.42
D	1.52	0.05	✓ 1.29	× 1.68	× 1.86	× 2.49
E	1.07	0.16	× 1.27	× 1.64	× 1.84	× 2.46

Table 7-4: C*(0.95) values for process variables, compared to P_{pk} values

The examples above are all processes for which the P_{pk} value is greater than one. The Bayesian approach is used to determine whether there is a high certainty that the process is capable. A high level of certainty suggests that enough data has been collected to have confidence in the conclusions of the capability study. When the P_{pk} value is high but less than C*(p), a larger sample size is required to be certain that the process is capable.

When a process capability study is used to compare variables with lower capabilities, to identify the priorities for improvement, the Bayesian approach may not add any extra information other than to confirm that the capability is low. Therefore this Bayesian method is most appropriate for processes with a P_{pk} value greater than one.

7.5 Conclusion from the reviewed applications

The use of Bayesian methods in statistical analysis allows for all of the available information to be utilised to give the greatest possible certainty in the results. In addition, the output from a Bayesian analysis enables the certainly of the results to be quantified, so that rather than finding the most likely result to occur, the likelihood of a favourable result can be found. Bayesian methods are more computationally expensive and complicated to implement compared to classical methods. However the results provide a greater certainty of achieving the goals of the analysis, or highlight areas in which there is high uncertainly and hence more information is required.

An issue in Bayesian statistics is the use of prior information, which is required to be quantified to be used formally in a Bayesian analysis. In a process development context, it is likely that some prior knowledge will exist as a process is scaled up from the laboratory to full scale manufacturing. The methods presented for process validation and process capability (Sections 7.3 and 7.4) both made use of a non-informative prior distribution, so no prior knowledge was captured. However these methods could be adapted to incorporate relevant prior information. For example, when the process capability is reported at regular intervals, the most recent data can be used to calculate the capability index. A novel method would be to use information from older data to form a prior distribution. This method is developed in the following section.

7.6 Bayesian approach to sequential P_{pk} calculations

Process capability indices produce an overview of the capability of the process, for the time period over which the data is collected. A useful application of PCIs is to track the capability over time, allowing for the identification of changes in the behaviour of a process that could present a risk of failing a specification limit. For a particular process at AstraZeneca, the process capability of a number of in-process variables is presented every month to provide a summary of the state of the process. Typically the P_{pk} values are calculated from the set of batches manufactured in the past month. There could be up to 20 batches, but at times there are significantly fewer depending on demand, planned shut downs and process delays.

When the sample size is small, the calculated P_{pk} values may be unreliable because there could be high variation due to sampling. A confidence interval around the P_{pk} value indicates the range of values that the true P_{pk} could take, but the interval will be wide when the sample size is small. For example, with a sample size of 20, a calculated P_{pk} value of 1.33 would have a 95% confidence interval width of +/-0.42, suggesting that the true capability could be as low as 0.91. Therefore it is difficult to determine whether the process is truly capable.

An alternative method is required for calculating a P_{pk} value when the sample size is small, so that the capability estimates can be updated every month with confidence in the P_{pk} value that is presented. A capability estimate is required for small sequential data sets, which allows the business to detect genuine changes to the process, without responding to variability in the capability metrics caused by small sample sizes. Such a metric would allow AstraZeneca to prioritise technical resources to investigate potential risks to the process specification limits.

Bayesian methods have been found to be useful for finding a posterior distribution for the P_{pk} value (Section 7.4). These methods could be extended to sequentially update the capability every month, by combining the new data from the current month with older data from previous months. Using a Bayesian structure, the distribution of the P_{pk} from the previous month would comprise the prior information and the new data collected would be the likelihood. Then when the sample size is small, more information would be taken from the previous month to allow a reliable estimate of the current process capability to be calculated.

In this section, the methodology is described for a novel Bayesian solution to the sequential calculation of P_{pk} values. This method is applied to simulated data to test how the P_{pk} results can detect changes in the mean or variability of the data, and then is applied to process data to determine how the method could be implemented at AstraZeneca.

7.6.1 Methodology for Bayesian sequential P_{pk}

A novel methodology is proposed for sequentially calculating the capability, using Bayesian techniques. For the proposed method, it is assumed that the process data will satisfy a normal distribution and hence the P_{pk} metric will accurately describe the process capability. There is scope to extend the method to be applied to non-normal data in future work.

The structure of the methodology is presented in Figure 7-12. P_{pk} is calculated from the mean and standard deviation of the data, so a posterior distribution for P_{pk} can be found from the posterior distributions of the mean and the standard deviation. Initially, a prior distribution must be specified for each parameter, either based on process knowledge or from previous data that has been collected. When there is limited prior knowledge, a very wide prior can be specified. Then data collected from the batches manufactured during the previous month will form the likelihood, which is combined with the priors to calculate the posterior distributions of the mean and standard deviation. By sampling from the two posterior distributions, values of P_{pk} can be calculated to provide a posterior distribution for P_{pk}.

The posterior distributions of the mean and standard deviation from the previous month will then form the prior distributions for the next month. The priors are then combined with new data to update the posteriors and calculate the new posterior for P_{pk}. Markov Chain Monte Carlo (MCMC) methods are used to combine the prior and likelihood information to produce the posterior distribution.

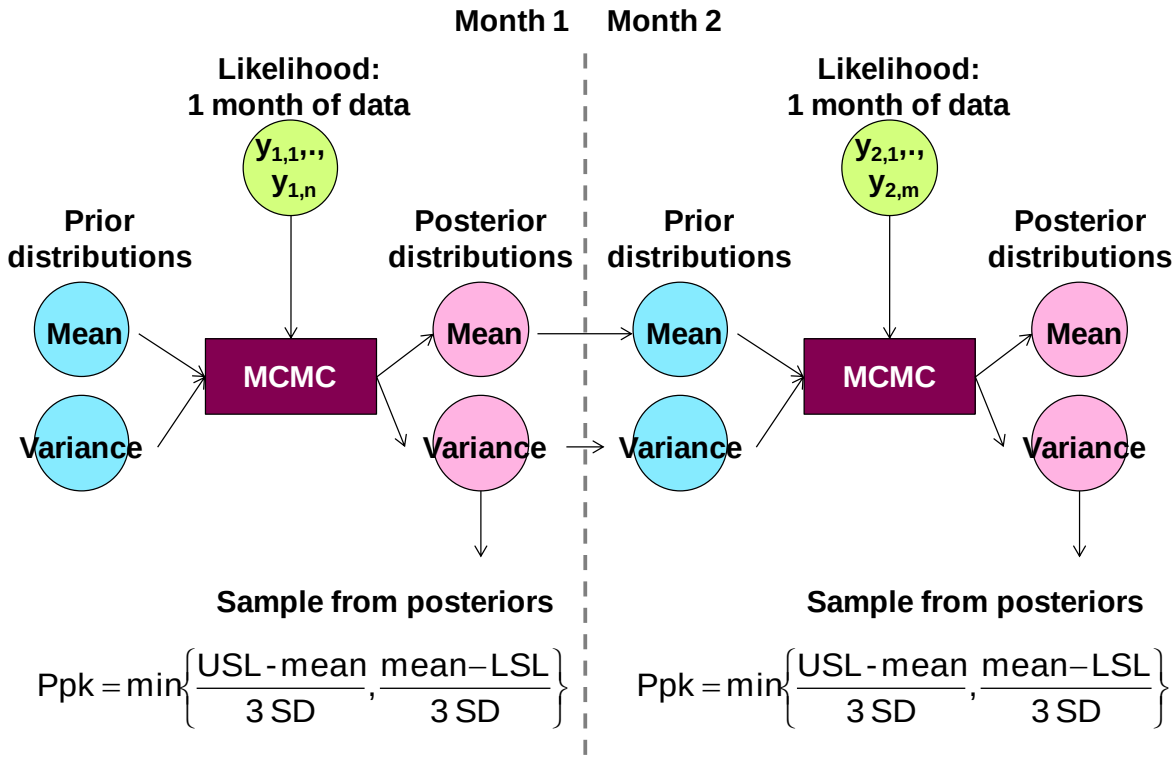


Figure 7-12: Structure of Sequential P_{pk} methodology

The model structure is as follows:

Prior Distribution

A normal prior is used for the sample mean, μ:

$\mu \sim N(m, v)$

The initial hyper-parameters, m and v, must be specified based on prior knowledge of the process mean. After the first sequential update, m and v will be taken from the mean and variance of the previous posterior distribution for μ.

The standard deviation, σ, is restricted to be a positive number, so a Gamma prior is used:

$\sigma \sim \text{Gamma}(a, b)$

where a is the shape parameter and b is the scale parameter. The initial hyper-parameters must be specified by the user and could be based on the level of variability that is observed in the process data. After the first sequential update, a and b are estimated from the mean and variance of the posterior for σ:

$$a = \text{mean}(\sigma)^2/\text{variance}(\sigma)$$ Equation 7-27

$$b = \text{variance}(\sigma)/\text{mean}(\sigma)$$ Equation 7-28

Likelihood

For a sample of n data points, y_1, y_2,... , y_n:

$y_i \sim N(\mu, \sigma^2)$, for i=1,..., n

Posterior

The posterior distributions are found by combining the information from the priors and the likelihood. At each iteration of the MCMC algorithm, samples are generated of the posteriors for μ and σ; these paired samples are then used to calculate samples of the posterior distribution of P_{pk}.

7.6.2 Application to simulated data

The ability of the Bayesian sequential P_{pk} method to represent the process capability is investigated using simulated data. The underlying distribution of the simulated data is known and hence the accuracy of the calculated P_{pk} values can be quantified. Initially the method is applied to data that is stable over time and then to data that exhibits changes in the mean or standard deviation.

7.6.2.1 Methodology for simulated data

The simulated data comprised ten samples of size 20, to represent ten monthly updates. Initially a stable dataset was created, for which each sample was generated from a normal distribution with a mean of ten and variance of one. The upper specification limit was set to 14, so the P_{pk} of the underlying distribution is 1.33. The analysis was run in Matlab 7.7.0 and uses slice sampling (Section 7.1.5.3) in the MCMC algorithm.

The prior distributions were specified as:

$\mu \sim N(10, 1)$

$\sigma \sim Gamma(2, 2)$

Since this is a simulated example and no prior information exists, the initial prior distributions were selected to have minimal impact on the subsequent posterior distributions. Both prior distributions include the underlying values of the parameters, but were wide so that they would have minimal impact on the posterior distributions. Choices of the prior distributions for the hyper-parameters are investigated in Sections 7.6.2.3 and 7.6.2.4.

The sampling and calculations proceeded as follows:

Step 1:

a. Set prior distributions as: $\mu \sim N(10,1)$, $\sigma \sim Gamma(2,2)$
b. Sample 20 data points from N(10,1)
c. Find the mean and SD of the sample to calculate P_{pk} ("data value" on graph)
d. Combine the sampled data with the prior distributions using MCMC to generate paired posterior samples of μ and σ
e. From each pair, calculate a sample of the posterior distribution for P_{pk}.
f. Find medians of the posterior distributions for μ, σ and P_{pk} (plot on graph)
g. Find the mean and variance of posterior samples of μ, to give $\mu \sim N(m_1, v_1)$
h. Find the mean and variance of posterior samples of σ, use Equation 7-27 and Equation 7-28 to find a_1 and b_1.

Step 2 to end:

a. Set priors distributions for the next step as $\mu \sim N(m_1, v_1)$ and $\sigma \sim Gamma(a_1, b_1)$ Repeat b. to f. from step 1
g. Find the mean and variance of posterior samples of μ, to give $\mu \sim N(m_2, v_2)$

h. Find the mean and variance of posterior samples of σ to find σ~Gamma(a_2, b_2)

Repeat for ten samples. The initial ten samples that were generated were used in the subsequent analysis.

7.6.2.2 Results for stable data

Ten updates of the posterior distributions were generated from the ten samples. The medians of the posterior distributions were monitored and compared to the calculated values from the individual samples of data and the values of the underlying distributions (Figure 7-13). The trend of the median P_{pk} over the 10 updates is smoother than the P_{pk} values calculated individually from each data set. This result suggests that by combining the new data with the prior taken from the previous update, the variation from sample to sample is filtered out. When a process is stable over time, but the individual samples exhibit variation, calculating the Bayesian sequential P_{pk} will provide a value that is more representative of the true capability than the P_{pk} values calculated from individual samples.

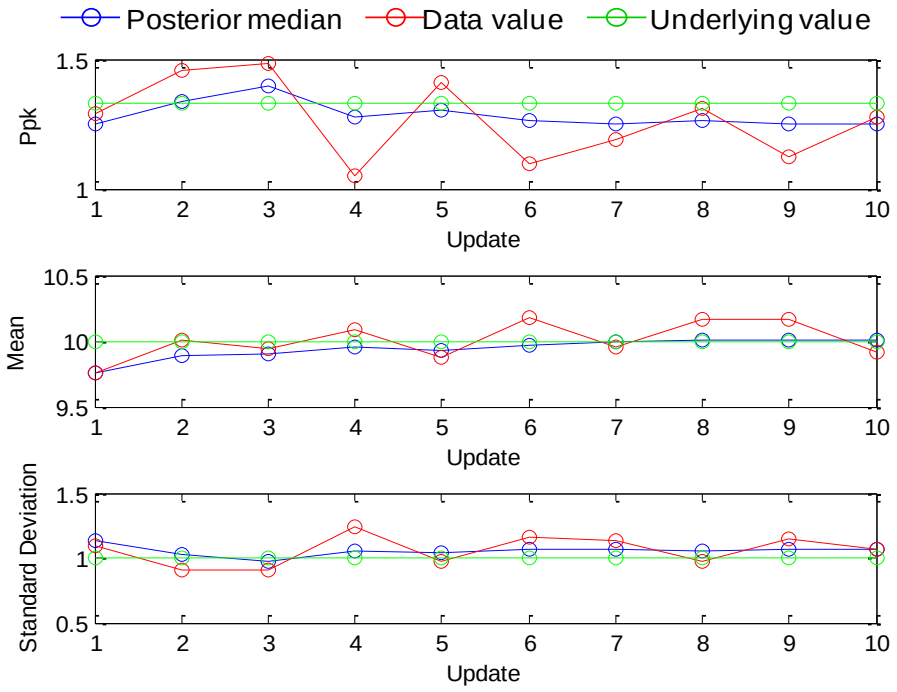


Figure 7-13: Results from ten updates of stable data

As the number of updates increases, more information is gained and hence the width of the posterior distribution for P_{pk} reduces (Figure 7-14). Although the underlying P_{pk} is 1.33, the target for a good process, the posterior distributions all extend to below 1.33 and hence it cannot be shown that the process is capable in a Bayesian sense (Section 7.4). However the width of 95% of the posterior distribution is approximately 0.5,

whereas a 95% confidence interval calculated from one month of data would have a width of approximately 0.84, suggesting that the P_{pk} value can be estimated with a higher precision with the Bayesian method.

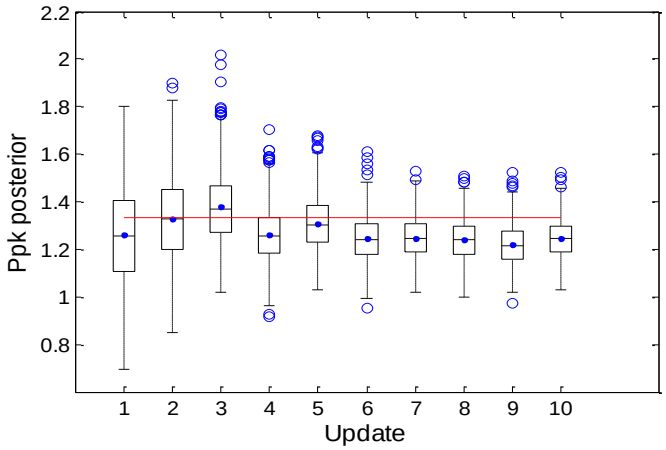


Figure 7-14: Posterior distribution P_{pk} for ten updates of stable data, red line is the underlying capability

7.6.2.3 Prior for the mean

The values of the hyper parameters that are specified in the prior distributions will affect the posterior distributions of the initial updates. Eventually the information from the data collected every month will have a stronger influence on the subsequent posterior distributions and the effect of the initial priors will not be seen. Ideally the prior distributions will reflect the past behaviour of the process. However when there is limited information available, a prior distribution is required to be specified that does not have a significant impact on the resulting posterior. The effect of the initial prior distributions was investigated by varying the values of the hyper-parameters and observing the effect on the posterior distributions, firstly for the prior for μ and then for σ (Section 7.6.2.4).

The methodology from Section 7.6.2.1 was implemented with just one step. This method was repeated with the prior mean for μ set to 10, 12, 14, 16, 18 and 20, and the prior standard deviation set to one, generating posterior distributions of μ (Figure 7-15) and σ (Figure 7-16).

When the prior mean was between ten and 16, the median of the posterior for μ is seen to increase slightly as the prior mean increases (Figure 7-15). This result indicates that the prior for μ has a small effect on the posterior and the posterior is more similar to the distribution of the data, which is sampled from Normal (10,1) distribution. However, when the prior mean is larger than 16, the posterior median trends upwards, reflecting

the distribution of the prior rather than the data. This effect suggests that when the mean for the prior is within the range of the likelihood, the posterior will be similar to the data. When the mean of the prior is outside of the range of the data, the prior has a stronger effect and pulls the posterior towards it.

Additionally, when the prior mean is greater than 16, the posterior median for σ becomes larger (Figure 7-16). This result suggests that some of the difference between the means of the prior and the data increases the overall variation in the model, which is captured as a larger σ.

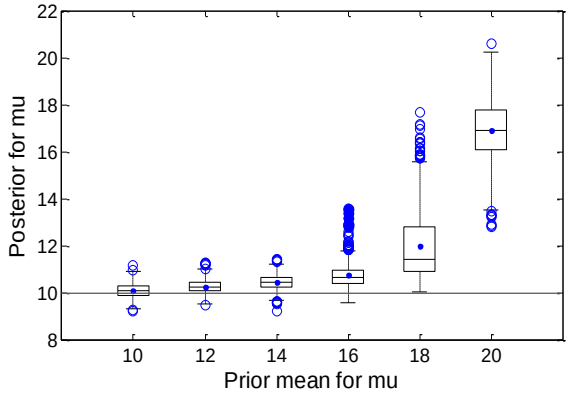


Figure 7-15: Posterior distributions for μ, varying the prior mean, with SD=1

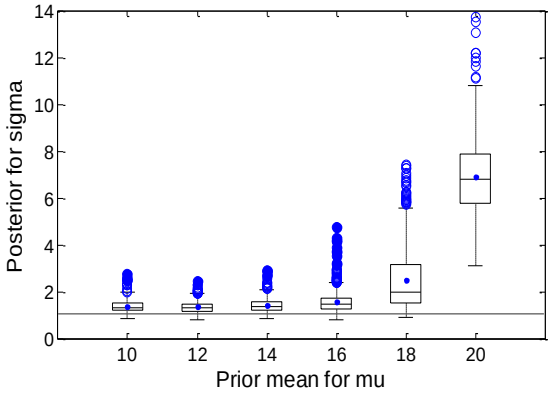


Figure 7-16: Posterior distributions for σ, varying the prior mean, with prior SD=1

The above analysis was subsequently re-run with the standard deviation for μ increased to two, to increase the range of the prior distribution. Figure 7-17 indicates that increasing the prior mean up to 20 is not seen to have an effect on the resulting posterior for μ. This result suggests that the wider prior allows the posterior to take more information from the data and hence all of the posterior medians are close to the mean of the data. Similarly the posterior distribution for σ remains constant when the prior mean is increased (Figure 7-18).

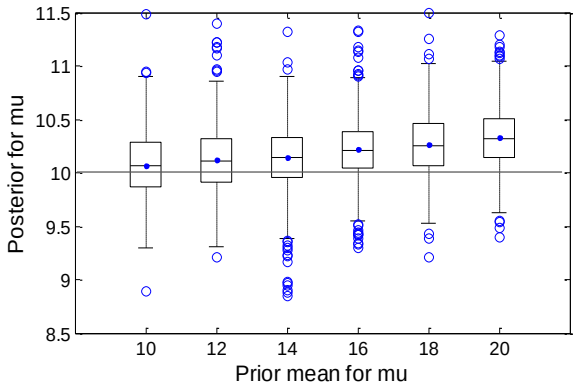


Figure 7-17: Posterior distributions for μ, varying the prior mean, with prior SD=2

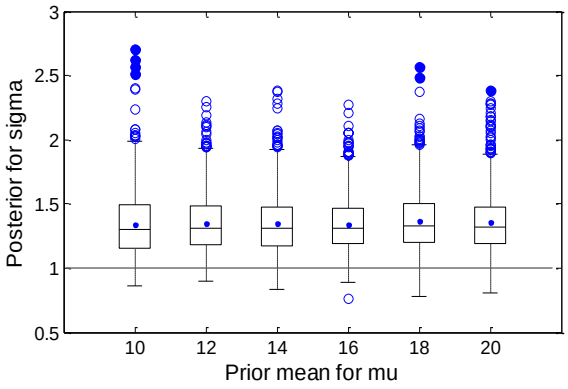


Figure 7-18: Posterior distributions for σ, varying the prior mean, with prior SD=2

The results indicate that if the prior mean is uncertain, then a wide prior distribution for μ is necessary to capture the information from the data. If there is information available of the likely values of the data, then the range of the prior distribution should include these values. When no prior information exists, then a very wide prior should be used. In a process context, knowledge of the measurement being studied would indicate a potential range of values that a measurement could take.

The use of a wider prior could result in wider posterior distributions and therefore less certainty in the P_{pk} estimate. The interquartile ranges (IQRs) of the posteriors for μ and σ were compared when the width of the prior was doubled (Table 7-5) and it was found that there was minimal change in the IQRs. Therefore it is more important to use a wide prior that is expected to include the range of the data than to use a narrower prior to reduce the posterior width.

Prior SD for μ	IQR of posterior for μ	IQR of posterior for σ
1	0.41	0.327
2	0.43	0.337

Table 7-5: Interquartile range for posterior distributions when prior mean set to ten

7.6.2.4 Prior for standard deviation

The prior distribution for the standard deviation is specified as a Gamma(a, b) distribution, where the shape (a) and scale (b) hyper parameters together define the location and width of the distribution. The values of a and b can be calculated by specifying the prior mean and variance for σ and applying Equation 7-27 and Equation 7-28.

A further study was undertaken to understand the effect of changing the hyper parameters for σ. With variance(σ) in the prior for σ set to one and the mean for σ was varied between 0.5 and 5 (Figure 7-19), and one update run to generate posterior distributions for μ (Figure 7-21) and σ (Figure 7-22).

When the prior mean of σ was set to five, i.e. much larger than the standard deviation of the data, the posterior for σ is wider and takes larger values. Similarly the posterior for μ is wider when the mean for σ is larger. As before, it appears that when the range of the prior for σ includes the data value (one), the resulting posteriors for μ and σ reflect the information in the data. However when the mean for σ is increased to five, the prior distribution does not include the value one (Figure 7-19) and the resulting posterior takes more information from the prior. Additionally, when the prior mean for σ

was increased to five, the posterior for μ becomes wider, suggesting that increased uncertainty in the model is reflected in wider posterior distributions.

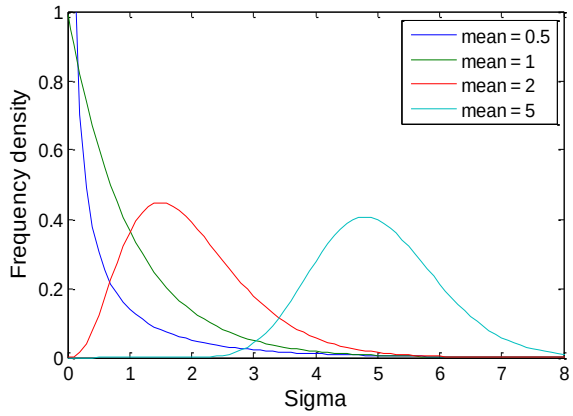


Figure 7-19: Prior distributions of σ, with variance(σ)=1

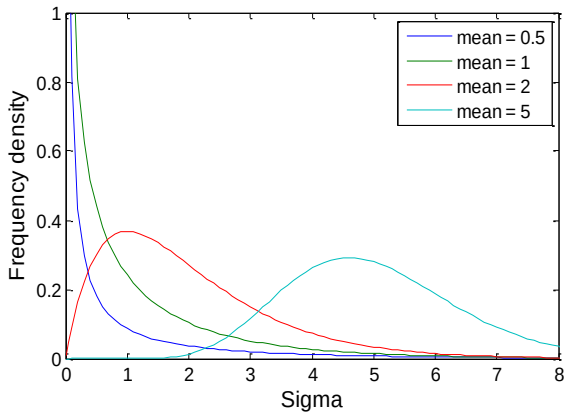


Figure 7-20: Prior distributions of σ, with variance(σ)=2

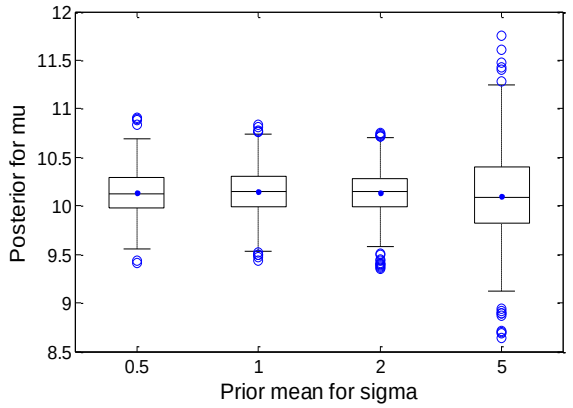


Figure 7-21: Posterior distributions for μ, varying the prior mean of σ, variance(σ)=1

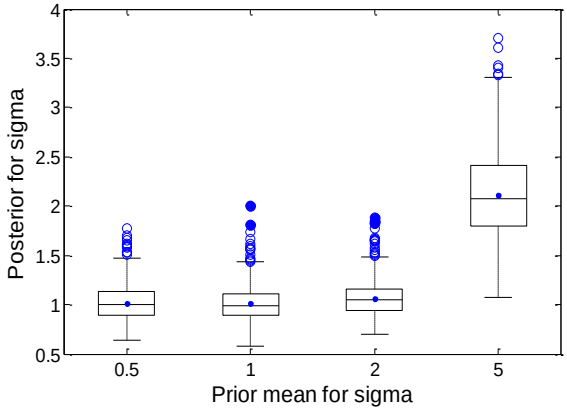


Figure 7-22: Posterior distributions for σ, varying the prior mean of σ, variance(σ)=1

When the variance of the prior for σ is increased to two, the ranges of the prior distributions are wider (Figure 7-20) and hence changing the mean has less of an effect on the resulting posterior distributions (Figure 7-23 and Figure 7-24). As before, the results indicate that the prior has a smaller effect on the posterior when the range of the prior includes the value of the standard deviation that is calculated from the data.

7.6.2.5 Change in process mean

The Bayesian sequential P_{pk} has been shown to filter out sample to sample variation in the data. However it is also required to detect changes in the mean and variability of the process, which indicate a change in the process capability.

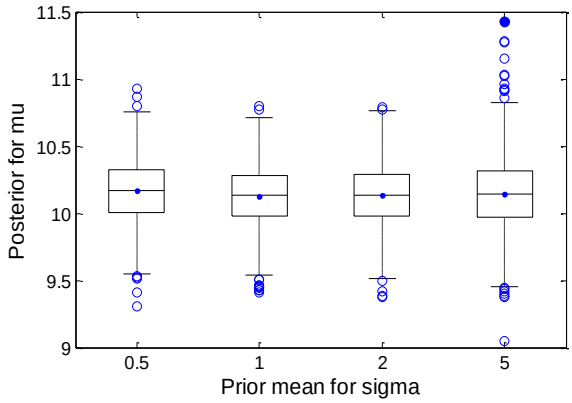


Figure 7-23: Posterior distributions for μ, varying the prior mean of σ, variance(σ)=2

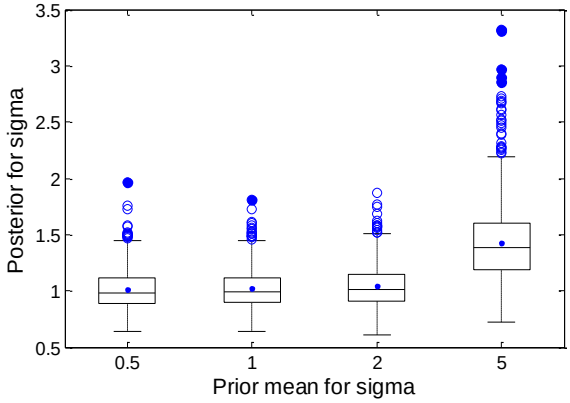


Figure 7-24: Posterior distributions for σ, varying the prior mean of σ, variance(σ)=2

The dataset created at the beginning of Section 7.6.2 was modified to show changes in the mean and standard deviation of the underlying distribution and the sequential P_{pk} method applied. Firstly the dataset was changed to represent a drift in the mean from ten to eleven over five updates and then a shift of the same magnitude over one update (Table 7-6). Furthermore, each dataset was scaled to unit variance so that variation in the standard deviation of the data did not influence the resulting P_{pk} values. Finally the dataset was adjusted to represent changes in the standard deviation (Section 7.6.2.6). In each case, the initial prior distributions used were Normal (10,1) for the mean and Gamma (0.5,2) for the standard deviation, which is equivalent to a mean of one and variance of two.

Following the median of the posterior P_{pk} over time shows that the sequential method is slow to reflect the changes in the mean. When the mean drifts from ten to eleven, the actual P_{pk} changes from 1.33 to one, but the posterior median after the final update is 1.16 (Figure 7-25). When the mean shifts, the posterior P_{pk} catches up more quickly, reaching a median of 1.04 by the final update (Figure 7-26).

	Drift of Mean				Shift of Mean		
Step	Underlying value				Underlying Value		
	μ	σ	P_{pk}		μ	σ	P_{pk}
1-5	10	1	1.33		10	1	1.33
6	10.2	1	1.27		11	1	1.00
7	10.4	1	1.20		11	1	1.00
8	10.6	1	1.13		11	1	1.00
9	10.8	1	1.07		11	1	1.00
10	11	1	1.00		11	1	1.00

Table 7-6: Underlying values of samples where the mean changes

Additionally, the posterior standard deviation increases when the underlying mean changes. The results show that the posterior P_{pk} is unable to reflect actual changes in

the process mean from sample to sample, because too much dependence is given to the prior distribution. A potential solution is developed in Section 7.6.2.7.

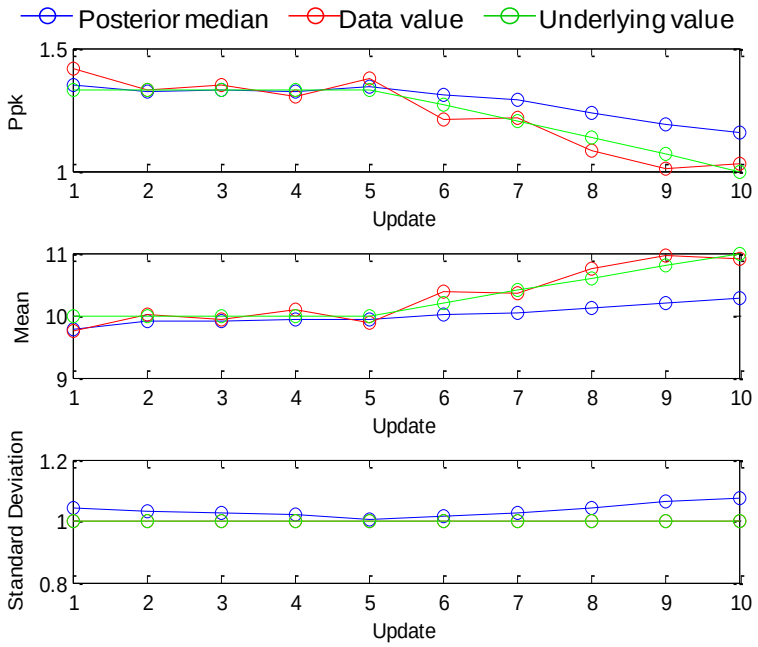


Figure 7-25: Results from ten updates, showing a drift in the mean from the 6th update

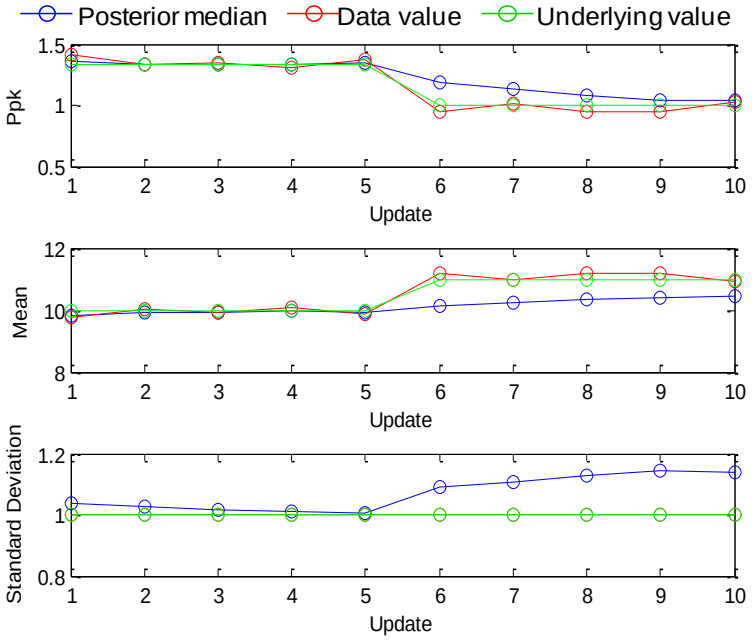


Figure 7-26: Results from ten updates, showing a shift in the mean at the 6th update

7.6.2.6 Change in standard deviation

The above analysis was repeated to assess how the posterior P_{pk} could follow changes to the standard deviation of the data. The stable dataset was changed to show a drift in standard deviation from one to 1.33 over five samples and a shift from one to 1.33 (Table 7-7).

	Drift of standard deviation						Shift of Standard Deviation				
	Underlying value			Gamma Parameters			Underlying value			Gamma Parameters	
Step	μ	σ	P_{pk}	a	b		μ	σ	P_{pk}	a	b
1-5	10	1.00	1.33	100	0.100		10	1.00	1.33	100	0.100
6	10	1.07	1.25	87.9	0.114		10	1.33	1.00	56.3	0.178
7	10	1.13	1.18	77.9	0.128		10	1.33	1.00	56.3	0.178
8	10	1.20	1.11	69.4	0.144		10	1.33	1.00	56.3	0.178
9	10	1.27	1.05	62.3	0.160		10	1.33	1.00	56.3	0.178
10	10	1.33	1.00	56.3	0.178		10	1.33	1.00	56.3	0.178

Table 7-7: Underlying values of the samples where the standard deviation changes

Similar to the results from changing the mean, the posterior standard deviation is slow to represent changes to the standard deviation of the data (Figure 7-27 and Figure 7-28). When the underlying P_{pk} reduces from 1.33 to one, the median posterior P_{pk} only reduces to 1.17 and 1.15 for drift and shift respectively, after five updates. The results suggest that too much weighting is given to the prior distributions and hence the posterior cannot pick up changes in the new data.

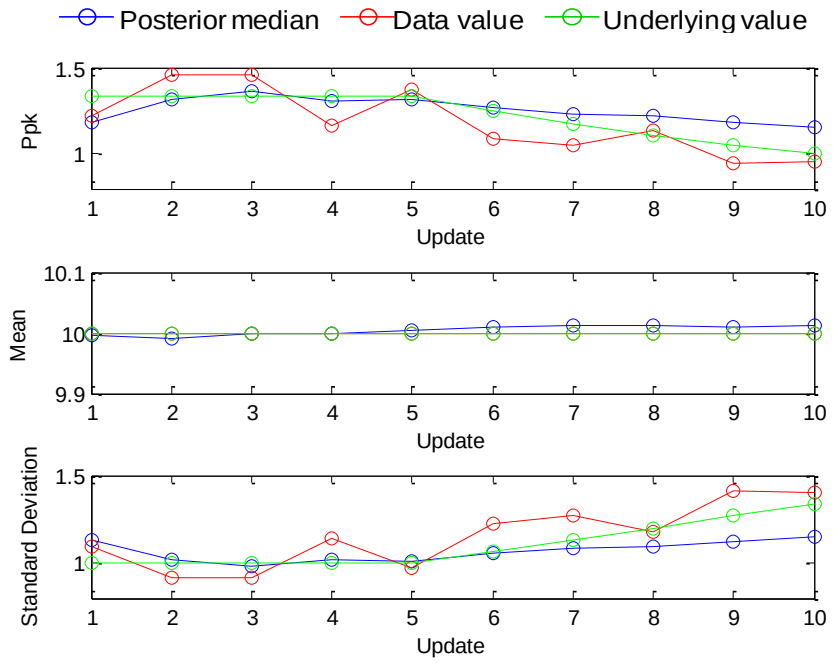


Figure 7-27: Results from ten updates, showing a drift in standard deviation from the 6th update

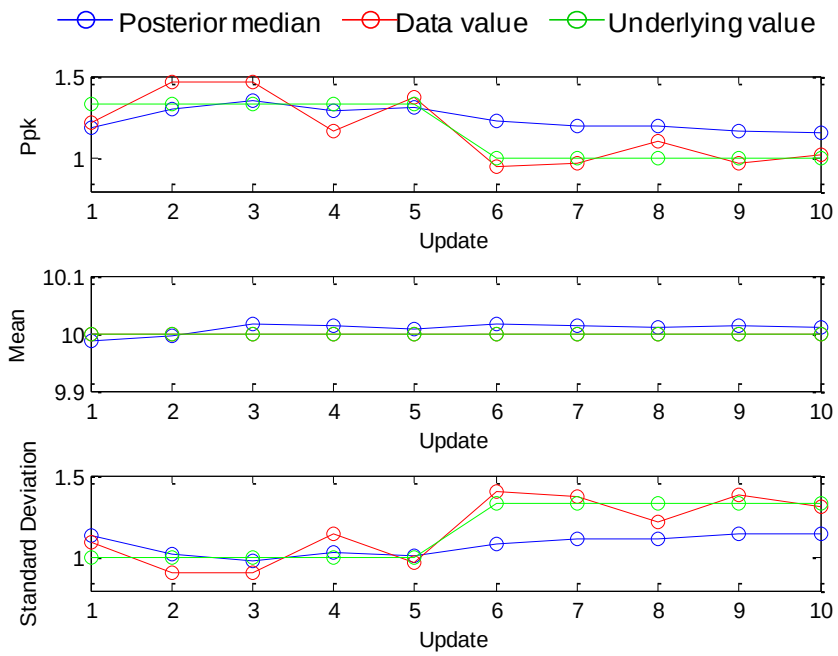


Figure 7-28: Results from ten updates, showing a shift in standard deviation at the 6th update

7.6.2.7 Reducing the impact of the prior distribution

For an effective Bayesian sequential P_{pk} method, a balance is required between filtering out the noise in the data and responding to actual changes in the process. This balance could be achieved by adjusting the influence of the prior distribution on the subsequent posterior. Reducing the influence of the prior distribution will increase the amount of information that is attained from the new set of data, allowing for the detection of process changes but also allowing more sample variation to be captured. This adjustment could be achieved by increasing the widths of the prior distributions and hence passing less information from the previous posteriors to the next update.

7.6.2.7.1 Prior for μ

Firstly the prior for the process mean (μ) is assessed. From the previous posterior distribution, the width of the prior is increased by scaling the variance by factor, k_1, so that the prior for μ becomes:

$$\mu \sim N(m, v*k_1^2)$$

This scale factor was applied to the five simulated datasets: stable data, drift and shift of the mean, and drift and shift of the standard deviation. The value of k_1 was varied between one and 2.5. For each dataset the mean squared error (MSE) was measured from the difference between the median of the posterior P_{pk} and the calculated P_{pk} of the underlying distributions (Figure 7-29). The changes to the mean and variation were

implemented from the 6th update, so the MSEs were measured from the 6th to 10th updates.

When the value of k_1 is increased, the MSE increases for the stable dataset, because the prior has less influence and hence the posterior is more similar to the data and does not filter out as much sample to sample variation (Figure 7-29). However the error reduces for the datasets that show a change in the mean, because the posterior P_{pk} is more able to follow these changes. Little effect is seen on the datasets exhibiting a change in the standard deviation because the underlying mean remains constant; although the error increases slightly for the dataset exhibiting a shift in standard deviation, due to less noise being filtered out.

7.6.2.7.2 Prior for σ

The sequential P_{pk} is also required to follow changes in the underlying variability in the data, captured by σ. The prior for σ was widened by increasing the variance of the prior from the previous posterior distribution by a scale of k_2. The resulting parameters for the Gamma prior distribution were calculated as:

$$a = mean(\sigma)^2/(variance(\sigma) * k_2)$$

$$b = (variance(\sigma) * k_2)/mean(\sigma)$$

Similar to the results for the mean prior, the error in the stable dataset increased when the value of k_2 was increased (Figure 7-30). As expected the error reduces for datasets that exhibit a changing variance, since the posterior P_{pk} can follow these changes. Additionally as k_2, increases the error also reduces for the datasets that exhibit a changing mean, the reasons for which are explained in Section 7.6.2.7.3.

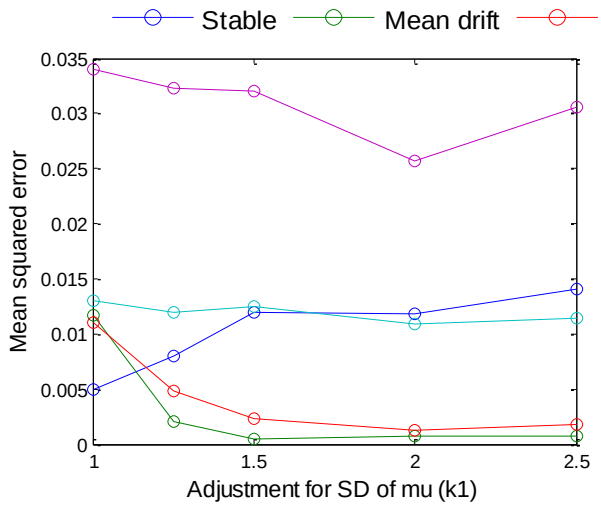


Figure 7-29: MSE when prior for μ is widened

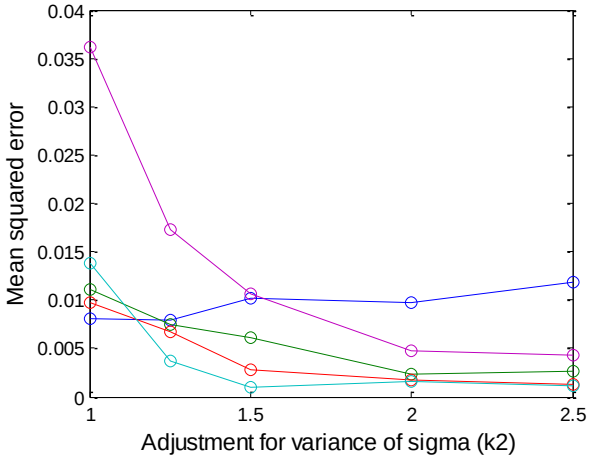


Figure 7-30: MSE when prior for σ is widened

7.6.2.7.3 How does the prior for σ affect the posterior for μ?

When the underlying mean of the data changes, the accuracy of the posterior median for P_{pk} is seen to increase as the prior for σ is made wider (Figure 7-30). However on closer inspection of the data, it can be observed that the accurate estimate for P_{pk} is a result of counteracting errors in the estimates of both the posterior mean and standard deviation. The reason can be illustrated when the prior variance is increased by a scale of two and the mean exhibits a shift change after the sixth update (Figure 7-31).

The prior for μ is narrow, hence the posterior μ does not reflect the changes in the mean of the data. However, this change in the data mean is represented by an overall increase in the posterior variability of the data, so the posterior median for σ is larger than the standard deviation of the data. As a result, the mean is underestimated, the standard deviation is overestimated and the calculated P_{pk} value is close to the underlying capability of the data.

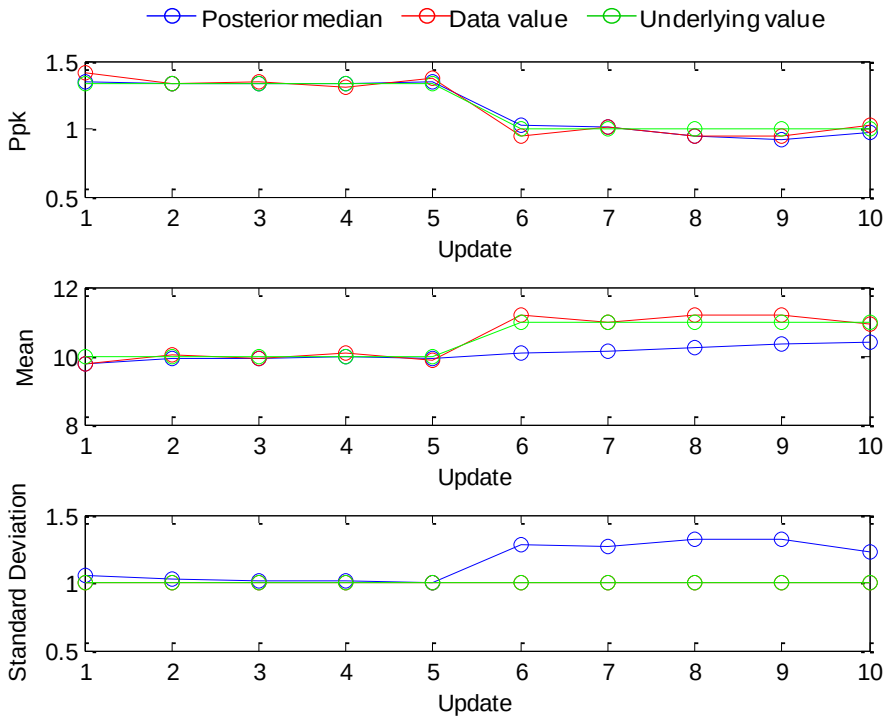


Figure 7-31: Posterior medians from ten updates, when k_2=2

When the priors for both the mean and standard deviation are widened (Section 7.6.2.7.4), the posterior for μ reflects the change in the mean. Hence the overall variability in the data is not seen to increase and the posterior for σ remains close to the actual variation in the data. Therefore to accurately estimate the posterior median of P_{pk} when the mean is changing, it is required to widen the prior distributions of both μ and σ.

7.6.2.7.4 Adjust both priors simultaneously

The optimal adjustment to the widths of the prior distributions is investigated by simultaneously varying the values of k_1 and k_2 and calculating the MSE (Figure 7-32). The results suggest that setting k_1 and k_2 to 1.5 reduces the error for the datasets that exhibit a changing mean or standard deviation, but do not cause a large increase in the error for the stable data. Setting k_1 and k_2 to 1.5 allows the posterior to follow changes in the mean and standard deviation more closely but still filter out the sample to sample variation (Figure 7-33). Therefore the recommended values for k_1 and k_2 are both 1.5.

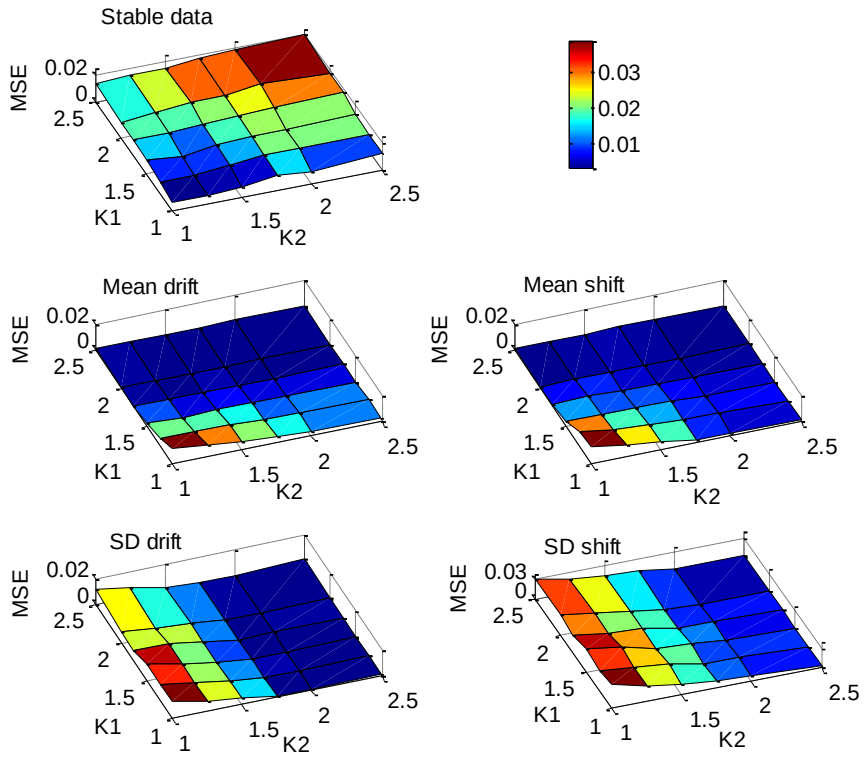


Figure 7-32: MSE when the priors for μ and σ are both varied, change

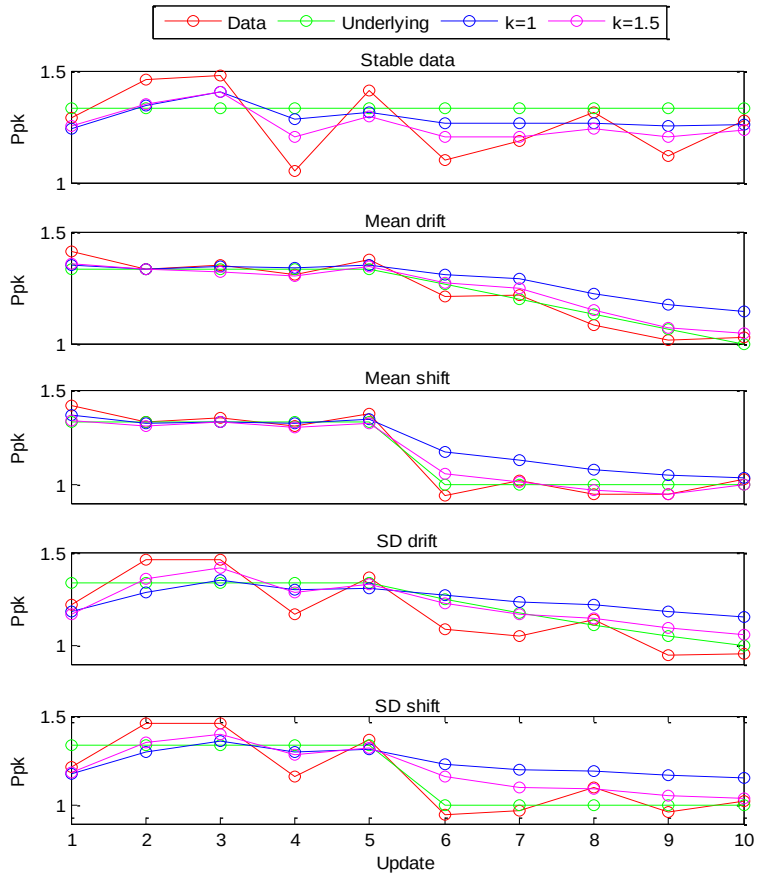


Figure 7-33: Results from ten updates, showing the posterior P_{pk} medians when there is no adjustment for the priors and when k1=k2=1.5.

When the widths of the prior distributions are increased, the resulting posteriors are also wider. The interquartile ranges of the posterior distributions for P_{pk} with and without the adjustment were compared. It can be seen that after ten updates the posterior with the adjustment is approximately 1.5 times wider (Figure 7-34). Consequently if the posterior distribution is wider, it will be less likely that the whole distribution is above the target for capability (e.g. 1.33) and so the process may not be labelled capable in a Bayesian sense (Section 7.4).

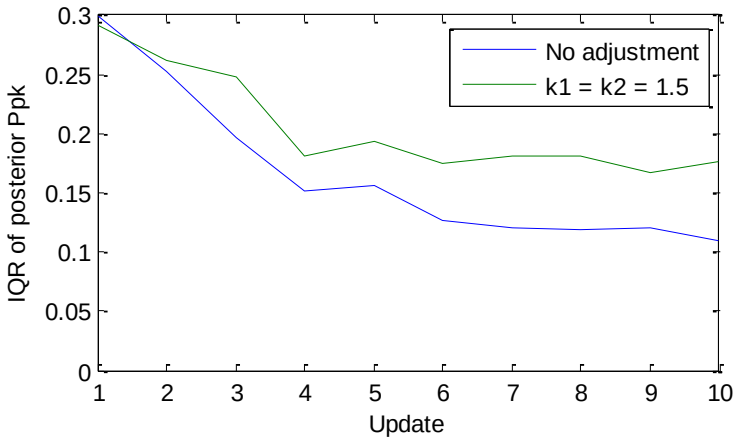


Figure 7-34: Interquartile range of posterior P_{pk}, with and without the adjustment for the prior variances

Whether or not the adjustment is preferred depends on the goal of the analysis. The adjustment may be preferred if the aim is to sequentially estimate the median P_{pk} of a process, because this method has been found to be more accurate at detecting changes in the mean and standard deviation. However, when the goal of the analysis is to determine the level of certainty that the process is capable, i.e. whether the process is capable in a Bayesian context, and if it can be assumed that the process is stable over time, then applying the adjustment may not be necessary and will result in a wider than necessary posterior distribution for P_{pk}.

7.6.3 Application to process data

The Bayesian sequential P_{pk} proposed in this chapter was applied to two sets of industrial process data that exhibited behaviour similar to a normal distribution. For each variable there is an upper specification limit that must be conformed to. P_{pk} values from these variables are calculated every month to track the capability of the process and to highlight any potential changes. Therefore the required output from the Bayesian sequential P_{pk} is the median of the posterior for P_{pk} and hence the adjustment for the prior variances was used, with k_1 and k_2 set to 1.5. Data was collected over a period of eight months, sample sizes varied between 13 and 35. The initial prior distributions were set based on the historical mean and variability of the data.

The mean for variable one appears to be stable over time but the standard deviation reduces, resulting in an increasing P_{pk} (Figure 7-35 and Figure 7-36). The posterior median appears to smooth out the variation in the P_{pk} values calculated directly from the data, but still highlights the increasing trend. If the process is drifting over time, for example due to the drift in a measurement probe, it is likely that the change will happen smoothly and in this case the sequential P_{pk} is able to reflect this change. From the

third month, the whole posterior distribution is above 1.33, so the process is considered to be highly capable.

For the second set of process data, the mean reduces over time and the standard deviation appears to vary from between months (Figure 7-37, Figure 7-38). The results for the sequential P_{pk} are similar to those of the standard P_{pk}, but the sequential method smooths out the noise in the standard deviation that occurs around month seven. The median posterior P_{pk} is below the target of 1.33 between months two to five, suggesting that the process should be targeted for improvement.

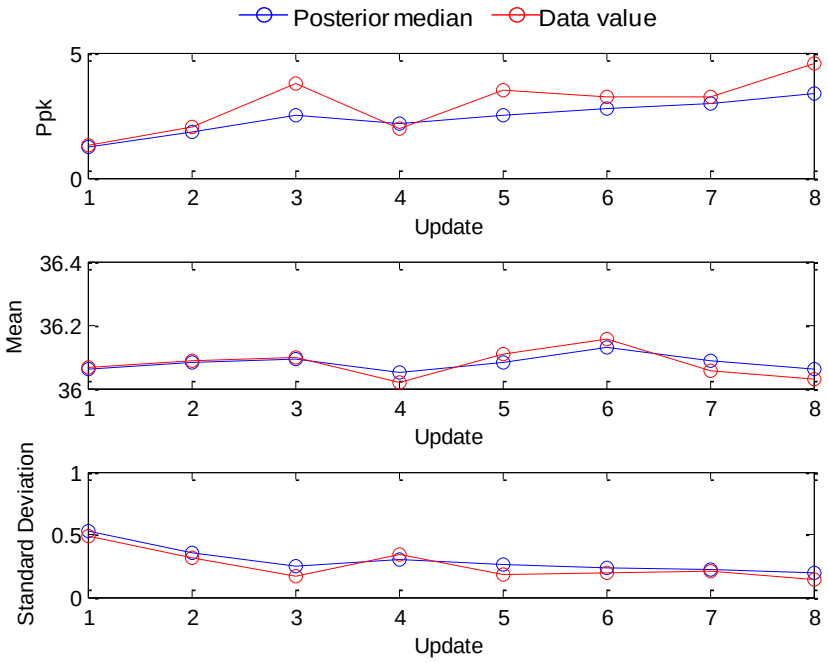


Figure 7-35: Bayesian sequential P_{pk} applied to variable one

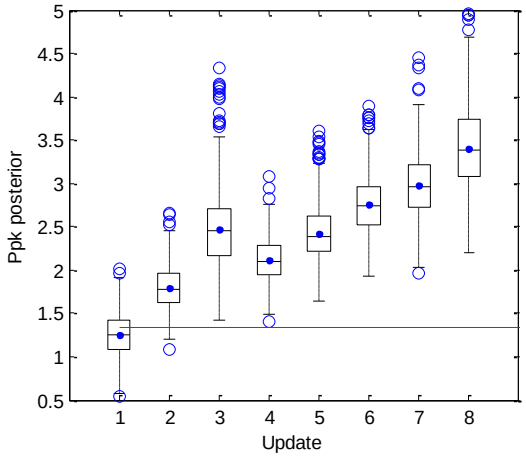


Figure 7-36: Posterior P_{pk} for variable one, showing the target of 1.33

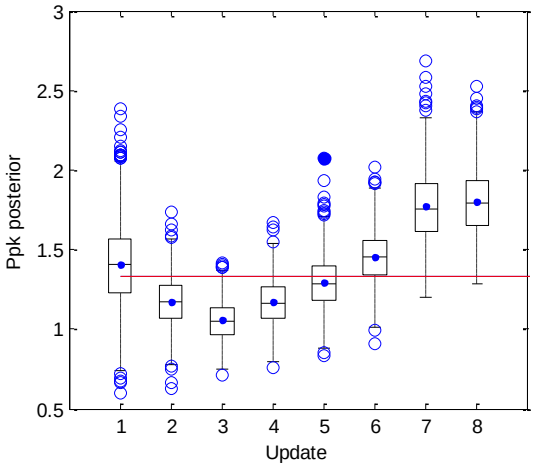


Figure 7-37: Posterior P_{pk} for variable two, showing the target of 1.33

7.6.4 Summary of Bayesian sequential P_{pk}

The Bayesian Sequential method for calculating the posterior distribution for P_{pk} has been developed to transfer information from one update to the next, but with a greater weighting being given to the new set of data. Use of this method required prior distributions to be specified for the mean and standard deviation of the data. It was found that in general the use of a wide uninformative prior distribution has little effect on the resulting posterior distribution for P_{pk}. Hence when the prior information is uncertain, a wide prior should be selected and the posterior distributions will reflect the information in the data.

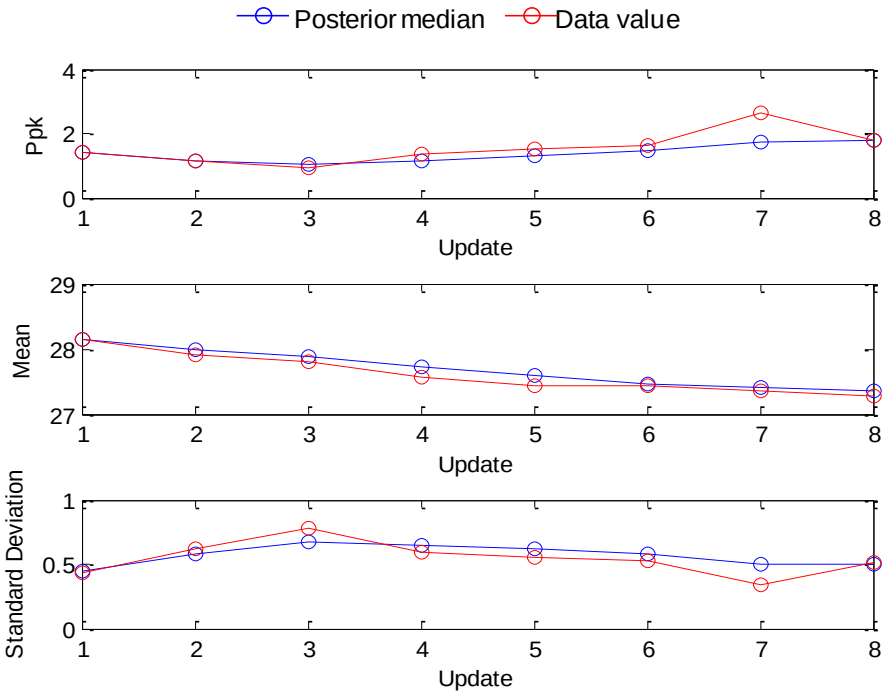


Figure 7-38: Bayesian sequential P_{pk} applied to variable two

This method has been found to be effective at filtering out the variation caused by taking small samples of data, both with simulated and industrial process data. However, to detect changes in the mean and standard deviation of the process, it is necessary to increase the width of the prior distributions to increase the influence of the new data. An adjustment of the variation of the prior distributions of 1.5 has been found to provide a balance between smoothing the noise and detecting actual changes.

Using the median of the posterior distribution, the Bayesian sequential P_{pk} has the potential to provide a reliable estimate of the capability of the process when the capability is tracked over time using small sample sizes.

7.7 Conclusions and further work

Bayesian methodologies are an alternative to the classical or frequentist statistical methods that are traditionally used. The major difference with Bayesian statistics is that parameters are considered to be stochastic variables, which are described by a distribution rather than a point estimate. As a result, Bayesian methods are more complex to implement but can provide a higher level of certainty in the results that are obtained.

In this chapter, three areas of application to pharmaceutical process development and manufacture were studied. Bayesian methods in experimental design allowed the use of prior information to optimise the design and maximise the information gained. When experiments are run sequentially, the design can be updated with the new information that is obtained, allowing for the design space to move towards the optimal operating space of the process.

Following experimental work, Bayesian modelling can be used to identify the optimal operating space in which the process will be run, i.e. the region in which there is a high level of certainty of meeting the required quality specifications. By calculating the posterior predictive distribution of the process outputs, the overall Bayesian reliability of meeting a specification limit can be found. Then the design space can be visualised to assess how changing the process inputs can determine the quality of the outputs.

When a manufacturing process has been established, the capability of meeting the required specification limits can be estimated though the use of process capability indices, including P_{pk}. By calculating the posterior distribution of P_{pk}, the level of certainly that the P_{pk} value is above the target for capability can be calculated, which indicates the certainty that the process is capable. A novel methodology has been proposed to extend the Bayesian process capability to sequentially update the posterior distribution for P_{pk} when data is analysed in sequential blocks. This method could be implemented at AstraZeneca to track the capability of process variables that are monitored monthly, allowing for the detection of adverse trends in the data but filtering out the variation that is seen when sample sizes are small. The methodology was shared with the Global Statistics Group at AstraZeneca, allowing the group to further develop and implement the methodology where appropriate.

As a future project, the Bayesian sequential P_{pk} methodology could be further investigated to determine how the method can handle different levels of variation in the data and different changes in the underlying distribution of the data. Additionally it will be useful to determine the effect of the sample size for each update on the accuracy of

the estimated P_{pk} values. The method could also be further tested on industrial process data at AstraZeneca to assess how the results can be applied to determine areas in which process improvements are required.

This methodology could also be further developed to be applicable to data from distributions other than the normal. Currently it is assumed that the data follows a normal distribution, however this assumption is not always valid for all industrial process data. The Clements' method (Section 7.1.3.2) can be applied by fitting an alternative distribution to the data and then calculating the percentiles of the distribution that are used to calculate P_{pk}. The Clements' method could be applied to the Bayesian sequential P_{pk} method by fitting an alternative distribution, such as the Gamma, and then finding posterior distributions for the parameters of the chosen distribution. Then by sampling from these posterior distributions, the percentiles of the chosen distribution could be calculated for each sample and used to determine the posterior distribution for P_{pk}.

8 Conclusions

This final chapter concludes the thesis by reviewing the outcome of each chapter and then capturing the contributions, both for AstraZeneca and more generally.

8.1 Chapter Three: Modelling Methodologies

In Chapter Three, modelling methodologies were investigated to identify techniques that are appropriate to be applied to data from pharmaceutical manufacturing processes. Specific challenges of industrial process data include large multivariate data sets, batch data and non-linear relationships between variables.

When the data is multivariate, principal component analysis (PCA) can be applied to analyse the trends between variables and samples, while partial least squares (PLS) allows a prediction model to be developed between process inputs and outputs. Batch data presents the additional challenge of a 3-dimensional structure, which can be unfolded at an observational or batch level and analysed with PCA and PLS. An alternative prediction method for batch data is case based reasoning, which is capable of handling non-linear trends in the data.

A further method for representing non-linear relationships between variables is artificial neural networks, which is considered a ‘black box’ modelling approach. Each time the training algorithm is run the resulting network is unique and hence stacked neural networks are applied to capture all of the trends in the data that may be represented by different individual networks. In the literature, applications of the above techniques include process monitoring, optimisation and control in processes including fermentation, polymerisation and drying.

8.2 Chapter Four: Modelling of Filter Drying Times

The objective of Chapter Four was to develop a prediction model to estimate the duration of a drying process that is run by AstraZeneca, using data collected early in the process. Four modelling methods were investigated (Table 8-1). The most accurate results were found for the neural network models, suggesting that the data collected throughout the process can be captured by a small number of variables without loosing information about the expected drying times. The results also indicated that non-linear models may be required to represent the relationship between the process variables and the drying time. However the linear regression model was most accurate for one of the dryers, suggesting that with sufficient data a linear model may be suitable.

The multi-way PLS models produced similar results to the linear regression models, however the additional complexity of capturing and handling multivariate data means that linear models are preferred. Case based reasoning produced the poorest predictions of the drying time, suggesting that this method is unable to handle the noise that is contained within the data collected from this drying process.

	Linear	**Non-linear**
Non-batch data	Multiple linear regression	Stacked neural networks
Batch data	Multi-way partial least squares	Case based reasoning

Table 8-1: Summary of modelling techniques

8.3 Chapter Five: Multivariate Analysis of Particle Size Distribution Data

The literature review in Chapter Five presented a comparison of particle size distribution (PSD) measurement methods, focusing on sieving, microscopy and laser diffraction. Laser diffraction is the preferred method for the analysis of pharmaceutical powders due to the fast and precise results that are obtained. Microscopy is also required to visualise the particles and determine whether the assumption of spherical particles is met.

The objective of the PSD study at AstraZeneca was to characterise the PSD of a product and to understand the causes of variation between samples. The application of PCA highlighted the differences between samples manufactured on the two processing plants and also indicated the presence of unusual samples within the data set. PLS and stacked neural networks were investigated to represent the relationship between the manufacturing conditions and the PSD, indicating that the drying time or precipitation temperature may potentially be linked to the PSD. However not enough variation was observed in the normal operating range of the process to demonstrate any firm relationships and hence experimental work would be required to further investigate the process.

8.4 Chapter Six: Process Capability Indices for Non-Normal Data

Process capability indices (PCIs) are used to calculate the risk of a process breaching a specification limit and are applied to both in-process measurements and quality testing on the final product. The use of PCIs allows for several sets of specification

limits to be compared across a process and priority areas for improvement to be determined. The standard process capability metric, P_{pk}, relies on the assumption that the data satisfies the normal distribution. Many sources of variation in an industrial process can cause the shape of the resulting data to differ from the normal distribution.

Distribution free PCIs have been investigated by simulating samples of data from known non-normal distributions. When no data is outside of the specification limits, calculating the percentiles of the data that correspond to the mean and three standard deviations of the normal distribution produced more accurate results than P_{pk}. However for distributions with a long tail towards the specification limit, the P_{pk} metric exhibits a smaller sampling variation. When there is some data outside of the limits, calculating the capability using the percentage of data outside of the limits was shown to produce the most accurate results and low sampling variation. This method was enhanced by calculating the percentiles of the data that correspond to the specification limits, providing more information about the process capability when there is only one data point outside of the limits.

8.5 Chapter Seven: Bayesian Methods in Pharmaceutical Process Development

Bayesian statistics is a branch of statistics whereby model parameters are considered to be random variables with an associated statistical distribution rather than being fixed values. In addition, Bayesian statistics allows for prior information to be combined with new information from data. A literature review was presented in Chapter Seven to give examples of applications of Bayesian statistics in pharmaceutical development. For example, when an experimental design is used to determine the optimal operating conditions for a process, prior information can be quantified and used to develop a design that will maximise the information that is gained from the results. Experiments can also be run sequentially to reduce the number of runs that are required. To optimise the process inputs, Bayesian modelling allows the certainty of the model parameters to be captured, so that an operating space can be found in which there is high certainty of meeting the process quality metrics.

Bayesian methods are also applicable to process capability calculations. Using a Bayesian model, the uncertainly associated with estimating the mean and standard deviation can be quantified to determine the probability that a process is capable. A novel Bayesian process capability methodology was developed for the case when process capability is analysed sequentially as more data is collected. This method was applied to processes at AstraZeneca, for which the process capability is assessed

monthly but a small number of batches are manufactured every month. Older data was used to find prior distributions for the mean and standard deviation, and combined with new data to estimate the posterior distributions. It was found that the prior distributions required widening to ensure that the posterior distributions captured the information from the new data. Following the investigation and adjustment of the prior distributions, the Bayesian sequential P_{pk} was able to filter out variation that was observed between samples and follow changes in the mean and variance of the process data.

8.6 Thesis Contributions

The contributions from this thesis have enabled AstraZeneca to gain more knowledge and understanding of their manufacturing process and have provided examples of the application of statistical methodologies to industrial process data. In addition, novel methodologies were developed for calculating the capability of a process.

8.6.1 Contributions to the Industrial Sponsor

Four case studies were presented relating to processes that are run by AstraZeneca, with the contributions listed below.

Predicting the duration of a drying process

- Analysis of the measurements collected during the drying process to determine which variables are most related to the drying time.
- Identified how the differences between the three dryers resulted in different trends in the temperature and flow rate measurements
- Development and implementation of a linear model to predict the drying time early in the drying process, allowing plant managers to make decisions of when the next batch will be required and when the filter should be cleaned.

Particle size distribution study

- Captured a baseline of the current particle size distribution of the product being made.
- Highlighted that there are differences between material manufactured on two different plants, but not between the main and recovery process streams.
- Quantified the differences in results when samples were analysed on different laser diffraction instruments.
- Determined that the current operating range of the process is not wide enough to indentify a relationship between the process variables and the resulting PSD.

- Initiated regular process capability analysis reports to track the capability of the process and to highlight any potential issues.
- Investigated distribution free capability indices that allow for a meaningful process capability estimate to be obtained when the data does not satisfy the normal distribution.
- Development of a Bayesian methodology to allow the process capability to be calculated monthly when a small sample size is collected every month.
- Methods and results were shared with the Global Operations Statisticians Forum at AstraZeneca, to support the development of a best practice for process capability analysis

8.6.2 General Contributions

The general contributions of the thesis include the assessment of how various statistical techniques can be applied to industrial process data and then the development of novel methods for quantifying the process capability.

- A comparison of modelling techniques for predicting the duration of a drying process, demonstrating how a number of modelling methodologies perform when applied to process data and assessment of the quantity of data was required to predict the duration of the drying process.
- Demonstration of how multivariate analysis techniques can enhance the analysis of particle size distribution data, allowing for the size of the data set to be reduced.
- Detailed comparison of distribution free process capability indices to determine the accuracy and sampling error when applied to data sampled from various statistical distributions, allowing for more robust process capability indices to be calculated.
- A novel extension to a process capability metric by calculating the percentile of the data that corresponds to the specification limit, increasing the accuracy of the capability estimate when few samples are outside of the limit.
- Literature review of the application of Bayesian statistics to the development of pharmaceutical manufacturing processes.
- Development of a novel methodology to utilise Bayesian statistics to sequentially calculate the process capability sequentially, allowing the capability to be estimated when data is collected in blocks of small sample sizes.

8.7 Future work

There exist a number of opportunities to build upon the work in this thesis, which are described in this final section.

8.7.1 Models of drying time

The models of the duration of the drying process could be improved by collecting more data for batches with long drying times. In particular the data set for dryer one only contained three batches with drying times longer than 70 hours, which may not be enough to fully reflect the trends in the input variables that indicate batches with long drying times. Fewer batches are observed with longer drying times, since the filters are generally cleaned when drying times start to increase, so it is required to monitor the process over time to identify that batches that should be added to the data set. Focus should be given to the multiple linear regression and neural network models, since the investigation concluded that it is not necessary to use multivariate methods to handle the data collected throughout the drying process. Additionally these models require fewer input data points and hence are more straightforward to implement.

An additional piece of work is to implement the prediction models in the plant's online control system, so that the predicted drying time can be generated as soon as the required data is generated and the information used to make decisions based on when the dryer is expected to become available and when a filter clean in required. Online implementation would require some logic to be derived to measure the inputs variables automatically. For the wash flow, rate this would require measuring the time and receiver water level at the start and end of the wash, based on the rate of change in the receiver level. The temperature peak after agitation could be measured as the maximum temperature over a specified time after the agitator has been used.

8.7.2 Optimisation the particle size distribution

There is an opportunity to further study the particle size distribution of the product and potentially adjust the manufacturing and milling conditions to optimise the PSD for the formulation process. The formulation process could be investigated to determine the relationship between the PSD and the critical quality attributes, to improve the process yield and therefore identify the optimal PSD. The relationship could be assessed by developing a PLS model between the PSD and the yield and other quality indicators of the formulation process. This model could be used to identify the range of the PSD that is expected result in an optimal formulation process. It may also be necessary to include as inputs variables relating to the formulation process that affect the outcome of the process.

Further investigation of the relationship between the manufacturing process and the PSD could allow the optimal processing conditions to be determined. It was concluded in Chapter Five that the process may be required to be run outside of the normal operating range to determine such a relationship and hence small scale laboratory work may be required, focusing on the variables that were identified as the most likely to have a relationship with the PSD. A design of experiments approach could be implemented to measure the effects of a number of input variables on the PSD.

8.7.3 Development of Bayesian sequential P_{pk}

The Bayesian method for sequential process capability calculations requires further analysis to determine the effectiveness of the method. The Bayesian sequential P_{pk} should be tested using data from a number of variables and over several months to determine how capable the method is of highlighting a potential change in the process capability whilst filtering out sample variation in the data. If necessary, adjustments may be required to improve the effectiveness of the method. For example the adjustment of the width of the prior distribution could be re-assessed with process data.

In addition, to enhance the effectiveness of the capability estimates, a sequential method is required that does not rely on the assumption that the data satisfies the normal distribution. An option could be to find an alternative distribution that fits more closely to the shape of the data, similar to the Clements' method discussed in Section 6.1.3.2. For the parameters of the selected distribution, prior distributions would be defined and combined with process data to determine the posterior distribution. Then by sampling from the posterior distributions of the parameters, percentiles of the selected distribution could be sampled and applied to produce samples of the posterior for P_{pk}.

Appendix 1: Principal Component Analysis

Calculating principal components from eigenvectors

Principal components were originally derived from the eigen decomposition of the covariance matrix, $\mathbf{X}^T\mathbf{X}$ (Wold et al, 1987a). The eigenvectors ($\mathbf{e}_1$, $\mathbf{e}_2$,…, $\mathbf{e}_k$) are listed in decreasing order of their corresponding eigenvalues. The eigenvectors make up the principal components. The eigenvalues show the proportion of variation in the data that is explained by each principal component. So for component i, with eigenvector $\mathbf{e}_i$ and eigenvalue λ_i, the proportion of explained variance is found by:

$$\lambda_i \Big/ \sum_{j=1}^{K} \lambda_j$$

Eigenvectors are mutually orthogonal so the PCs define uncorrelated directions of variation. The retained PCs are always selected in order of decreasing eigenvalues.

The loadings matrix consists of the eigenvectors that are to be retained in the model, so if A PCs are to be retained, then $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_A]$. It is necessary to constrain the magnitude of each loading vector, otherwise any multiple of the eigenvector is possible, so $||\mathbf{p}_i|| = 1$.

The scores matrix, **T**, is calculated as the product of the data **X** and the loadings **P**, so $\mathbf{T} = \mathbf{XP}$. The columns of **P** and **T** are orthogonal, since **P** consists of the eigenvectors of **X**, so $\mathbf{p}_i^T\mathbf{p}_j = 0$ and $\mathbf{t}_i^T\mathbf{t}_j = 0$, for $i \neq j$.

R^2X and Q^2

R^2X is the cumulative variance explained by the retained principal components (Eriksson *et al*, 2006):

$$R^2X = 1 - \frac{RSS}{SSX} \qquad \text{Equation 0-1}$$

Where:

$$RSS = \sum_{i=1}^{N}\sum_{j=1}^{K}\left(\mathbf{E}_{ij}\right)^2 \qquad \text{Equation 0-2}$$

$$SSX = \sum_{i=1}^{N}\sum_{j=1}^{K}\left(\mathbf{X_{ij}}\right)^{2}$$ Equation 0-3

The Q^2 measures the proportion of the variation that can be predicted by a component. As each component is added to the model, the SPE (Section 1.1.1.5) from the cross-validation is compared to the residual sum of squares (RSS) from the PCA model with one fewer component. The RSS is found from the error matrix **E**. If the SPE for a PCA with A components, SPE_A, is larger than the RSS for a PCA with A-1 components, RSS_{A-1}, then adding the A^{th} component does not increase the predictive power of the model (Wold et al 1987a). The Q^2 for the A^{th} component is given by:

$$Q_A^2 = 1 - \frac{SPE_A}{RSS_{A-1}}$$ Equation 0-4

For the first PC, the RSS is found directly from the scaled and centred **X** matrix.

The cumulative Q^2 for a PCA model with A components is given by:

$$Q^2(\text{cumulative}) = 1 - \prod_{i=1}^{A}\left(\frac{SPE_i}{RSS_{i-1}}\right)$$ Equation 0-5

Appendix 2: Partial Least Squares

PLS algorithm

Partial least squares is used to find a linear relationship between a set of K predictor variables ($\mathbf{X} \in \Re^{NxK}$) and a single or set of M response variables ($\mathbf{Y} \in \Re^{NxM}$), with N observations.

Firstly weight vectors $\mathbf{w_1}$ and $\mathbf{v_1}$ are found such that $\mathbf{t_1}$ and $\mathbf{u_1}$ have the maximum possible covariance (cov($\mathbf{t_1}$, $\mathbf{u_1}$) = $\mathbf{t_1}^T\mathbf{u_1}$), where:

$\mathbf{t_1} = \mathbf{X}\mathbf{w_1}$ Equation 6

$\mathbf{u_1} = \mathbf{Y}\mathbf{v_1}$ Equation 7

These vectors will describe the direction of maximum variation in the predictor and output variables. The vectors $\mathbf{t_1}$ and $\mathbf{u_1}$ are linear combinations of the original variables. A further constraint is that $\| \mathbf{w_1} \| = \| \mathbf{v_1} \| = 1$.

A predictive model is built between **X** and **Y** by finding the inner relationship between the latent variables:

$\mathbf{u_1} = \mathbf{b_1}\mathbf{t_1} + \mathbf{e_1}$, Equation 8

where $\mathbf{b_1}$ is found by ordinary least squares regression and $\mathbf{e_1}$ is the error vector.

Then the outer relationship is found that links the latent variables to the original variables.

$\mathbf{X} = \mathbf{t_1}\mathbf{p_1}^T + \mathbf{E_1}$ Equation 9

$\mathbf{Y} = \hat{\mathbf{u}}_1\mathbf{q}_1^T + \mathbf{F_1} = \mathbf{b_1}\mathbf{t_1}\mathbf{q}_1^T + \mathbf{F_1}$ Equation 10

Again $\mathbf{p_1}$ and $\mathbf{q_1}$ are calculated by least squares regression.

To calculate the second latent variable, the contribution from the first is removed from **X** and **Y**:

$\mathbf{X_2} = \mathbf{X} - \mathbf{t_1}\mathbf{p_1}^T$ Equation 11

$\mathbf{Y_2} = \mathbf{Y} - \mathbf{b_1}\mathbf{t_1}\mathbf{q_1}^T$ Equation 12

The second latent variables are found in the same way as the first. This process is repeated as additional latent variables are calculated. If A latent variables are chosen to be included, then **X** and **Y** are estimated as:

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t_i p_i^T} + \mathbf{E}$$ Equation 13

$$\mathbf{Y} = \sum_{i=1}^{A} \mathbf{\hat{u}_i q_i^T} + \mathbf{F}$$ Equation 14

R^2Y and Q^2

R^2Y denotes the proportion of variation in the response variables that is explained by the retained latent variables:

$$R^2Y = 1 - \frac{RSS}{SSY}$$

where:

$$RSS = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\mathbf{F_{ij}}\right)^2$$

$$SSY = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\mathbf{Y_{ij}}\right)^2$$

Q^2 denotes the proportion of variation than can predicted on new data by the retained latent variables:

$$Q^2 = 1 - \frac{SPE}{SSY}$$

The SPE is found from cross-validation (Section 1.1.2.2).

Appendix 3: Components of Variance Calculation

For a study with several repeated measurements from a number of batches, the variance components for batch and repeatability are found from the mean square (MS) in the analysis of variance table, as follows:

Variance component for batch = MS_{batch} / number of repeats

Repeatability = MS_{error}

Bibliography

Ahmad, S., Abdollahian, M., Zeephongsekul, P. 2008. Process capability estimation for non-normal quality characteristics using Clement, Burr and Box-Cox methods. *Anziam J*, 49: 642-665.

AIAG, 2002. *Measurement Systems Analysis*. 3rd ed. Detroit, Michigan: DaimlerChrysler, Ford Motor, General Motors.

Allen, T. 1997. *Particle size measurement*. 5th ed. London: Chapman and Hall.

am Ende, D., Bronk, K. S., Mustakis, J., O'Connor, G., Santa Maria, C. L., Nosal, R., Watson, T. J. N. 2007. API quality by design examples from the Torcetrapib manufacturing process. *Journal of Pharmaceutical Innovation*, 2: 71-86.

Anis, M. Z. 2008. Basic process capability indices: A expository review. *International Statistical Review*, 76 (3): 347-367.

Anjard, R. P., Hagerty, D., Griffith, G. K., Liu, S-T., Mustonen, E. M., Pasfield, D. A. 1991. C_{pk} applications – Uses and abuses. *Microelectron Reliability*, 31 (6): 1123-1125.

Banks, D. 1993. Is industrial statistics out of control? *Statistical Science*, 8 (4): 356-377

Beekman, A., Shan, D., Ali, A., Dai, W., Ward-Smith, S., Goldenberg, M. 2005. Micrometer-scale particle sizing by laser diffraction: Critical impact of the imaginary component of refractive index*. Pharmaceutical Research*, 22 (4): 518-522.

Behazadi, S. S., Prakasvudhisarn, C., Klocker, J., Wolschann, P., Viernstein, H. 2009. Comparison between two types of artificial neural networks used for validation of pharmaceutical processes. *Powder Technology*, 195: 150-157

Bendell, A., Disney, J., McCollin, C. 1999. The future role of statistics in quality engineering and management. *The Statistician*, 48 (3): 299-326.

Berry, D. A. 2006. Bayesian clinical trials. Natures Reviews: Drug Discovery, 5: 27-36.

Bicheno, J. 2004. *The New Lean Toolbox*. Buckingham: PICSIE Books.

Bosquillon, C., Lombry, C., Preat, V., Vanbever, R. 2001. Comparison of particle sizing techniques in the case of inhalation dry powders. *Journal of Pharmaceutical Sciences*, 90 (12): 2032-2041.

Box, G. 1994. Statistics and quality improvement. *Journal of the Royal Statistical Society. Series A,* 157 (2): 209-229.

Box, G. E. P., Cox, D. R. 1964. An analysis of transformations. Journal *of the Royal Statistical Society. Series B*. 26 (2): 211-252.

Box, G., Kramer, T. 1992. Statistical Monitoring and Feedback Adjustment: A Discussion. *Technometrics*, 34 (3): 251-267.

Brooks, S. P. 1998. Markov Chain Monte Carlo method and its application. *The Statistician*, 47 (1): 69-100.

Brueggemeier, S. B., Reiff, E. A., Lyngberg, O. K., Hobson, L. A., Tabora, J. E. 2012. Modeling-based approach towards quality by design for the Ibipinabant API step. *Organic Process Research and Development*, 16: 567-576.

Brülls, M., Folestad, S., Sparén, A., Rasmuson, A. 2003. In-situ near-infrared spectroscopy monitoring of the lyophilization process. *Pharmaceutical Rresearch*, 20 (3): 494-499.

Burt, J. L., Braem, A. D., Ramirez, A., Mudryk, B., Rossano, L., Tummala, S. 2011. Model-guides design space development for drug substance manufacturing process. *Journal of Pharmaceutical Innovation*, 6: 181-192.

Chaloner, K., Verdinelli, I. 1995. Bayesian Experimental Design: A Review. *Statistical Science* 10 (3): 273-304.

Chang, P. L., Lu, K. H. 1994. PCI calculations for any shape of distribution with percentile. *Quality World Technical* (September): 110-114.

Chen, H. 1994. A multivariate process capability index over a rectangular solid tolerance zone. *Statistica Sinica*, 4: 749-758.

Chen, J-P., Ding, C. G. 2001. A new process capability index for non-normal distributions. *International Journal of Quality and Reliability Management*, 18 (7): 762-770.

Chen, K. S., Pearn, W. L. 1997. An application of non-normal process capability indices. *Quality and Reliability Engineering International*, 13: 355-360.

Cheng, C. S. 1997. A neural network approach for the analysis of control chart patterns. *International Journal of Product Research*, 35 (3): 667-697.

Cheng, S. W., Spiring, F. A. 1989. Assessing process capability: A Bayesian approach. *IIE Transactions*, 21 (1): 97-98.

Chew, W., Sharratt, P. 2010. Trends in process analytical technology. *Analytical Methods*, 2: 1412-1438.

Chou, C-J., Chen, L-F. 2012. Combining neural networks and genetic algorithms for optimising the parameter design of the inter-metal dielectric process. International Journal of Production Research, 50 (7): 1905-1916.

Coit, D. W., Jackson, B. T., Smith, A. E. 1998. Static neural network process models: considerations and case studies. *International Journal of Product Research*, 36 (11): 2953-2967.

Colosimo, B. M., del Castillo, E. 2007. Bayesian Process Monitoring, Control and Optimization. Taylor and Francis Group.

Congdon, P. 2003. *Applied Bayesian Modelling*. Wiley.

Cook, D. F., Ragsdale, C. T., Major, R. L. 2000. Combining a neural network with a

genetic algorithm for process parameter optimization. *Engineering Applications of Artificial Intelligence*, 13: 391-396.

Coulson, J. M. 2007. *Chemical Engineering. Volume 2, Particle technology and separation processes*. 5th ed. Amsterdam, Boston: Butterworth-Heinemann.

Cui, Y., Song, X., Chuang, K., Venkatramani, C., Lee, S., Gallegos, G., Venkateshwaran, T., Xie, X. 2012. Variable selection in multivariate modelling of drug product formula and manufacturing process. Pharmaceutical Technology, 101 (12): 4597-4607.

Dassau, E., Zodak, I., Lewin, D. R. 2006. Combining six-sigma with integrated design and control for yield enhancement in bioprocessing. *Industrial Engineering Chemical Research,* 45: 8299-8309.

De Villiers, M. M. 1995. Influence of cohesive properties of micronized drug powders on particle size analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 13 (3):191-198.

Deleryd, M. 1998. On the gap between theory and practice of process capability studies. *International Journal of Quality and Reliability Management*, 15 (2): 178-191.

Deleryd, M. 1999. A pragmatic view on process capability studies. *Int. J. Production Economics*, 53: 319-330.

Deming, W. E. 1986. *Out of the Crisis*. Cambridge: Massachusetts Institute of Technology Press.

Deshpande, P. B., Makker, S. L., Goldstein, M. 1999. Boost competitiveness via six sigma. *Chemical Engineering Process*, September 1999: 65-70.

Does, R. J. M. M., Trip, A. 2001. The impact of statistics in industry and the role of Statisticians. *Austrian Journal of Statistics*, 30 (1): 7-20.

Doherty, S. J., Lange, A. J. 2006. Avoiding pitfalls with chemometrics and PAT in the pharmaceutical and biotech industries. *Trends in Analytical Chemistry*, 25 (11): 1097-1102.

Dube, M. C., Penlidis, A., Reilly, P. M. 1996. A systematic approach to the study of multicomponent polymerisation kinetics: The butyl acrylate / methyl methacrylate / vinyl acetate example. IV. Optimal Bayesian design of emulsion terpolymerization experiments in a pilot plant reactor. *Journal of Polymer Science: Part A: Polymer Chemistry*, 34: 811-831.

DuMouchel, W., Jones, B. 1994. A simple Bayesian modification of D-optimal design to reduce dependence on an assumed model. *Technometrics*, 36 (1): 37-47.

Efron, B., Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37 (1): 36-48.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., Wold, S. 2006. Multi- and Megavariate Data Analysis. Part 1 Basic Principles and Applications.

Umetrics 2006, Umeå, Sweden.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., Wold, S. 2008. Design of Experiments: Principles and Applications. Umetrics 2008, Umeå, Sweden. pp 217

Fahmy R., Kona, R., Dandu, R., Xie, W., Claycamp, G., Hoag, S. W. 2012. Quality by design 1: Application of failure mode effect analysis (FMEA) and Plackett-Burman design of experiments in the identification of "main factors" in the formulation and process design space for roller-compacted Ciprofloxacin Hydrochloride immediate-release tablets. *AAPS PharmSciTech*, 13 (4): 1243-1254.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. 2013. Regression Models, Methods and Applications. Springer Berlin Heidelberg.

FDA, 2004. Guidance for Industry. PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.

FDA, 2009. Guidance for Industry: Q8(R2) Pharmaceutical Development.

FDA, 2011. Guidance for Industry. Process Validation: General Principles and Practices.

Ferreira, A. P., Lopes, J. A., Menezes. J. C. 2007. Study of the application of multiway multivariate techniques to model data from an industrial fermentation process. *Analytica Chimica Acta*, 595: 120-127.

Frank, I. E. 1990. A nonlinear PLS model. *Chemometrics and Intelligent Laboratory Systems,* 8: 109-119

Gabrielsson, J., Jonsson, H., Trygg, J., Airiau, C., Schmidt, B., Escott, R. 2006. Combining process and spectroscopic data to improve batch modelling. *AIChE Journal*, 52 (9): 3164-3172.

Garcia-Munoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., Murphy, G. 2003. Troubleshooting of an industrial batch process using multivariate methods. *Industrial Chemical Engineering Research*, 42: 3592-3601.

Gelfand, A. E., Smith, A. F. F. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85 (410): 389-409.

Geman, S., Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721-741.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57 (6): 1317-1339.

Gilmour, S. G., Mead, R. 1995. Stopping rules for sequences of factorial designs. *Applied Statistics*, 44 (3): 343-355.

Grau, D. 2010. New capability indices for non normal process. *Communications in Statistics – Theory and Methods*, 39 (16): 2913-2929.

Grossberg, S. 1988. Nonlinear neural networks, principles, mechanisms and

architectures. *Neural Networks*, 1: 17-61.

Guh, R-S. 2007. On-line identification and quantification of mean shifts in bivariate processes using a neural network-based approach. *Quality and Reliability Engineering International*, 23: 367-385.

Haaland, D. M., Thomas, E. V. 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 60: 1193-1202.

Hahn, G. J., Hill, W. J., Hoerl, R. W., Zinkgraf., S. A. 1999. The impact of six sigma improvement - A glimpse into the future of statistics. *The American Statistician*, 53 (3): 208-215.

Hair, J. F., Sarstedt, M., Hopkins, L., Kuppelweiser, V. G. 2014. Partial least squares structural equation modeling (PLS-SEM). European Business Review, 26 (2): 106-121.

Haskell, R. J. 1998. Characterization of submicron systems via optical methods. *Journal of Pharmaceutical Sciences*, 87 (2): 125-129.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97-109.

Himmelblau, D. M. 2000. Applications of artificial neural networks in chemical engineering. *Korean Journal of Chemical Engineering*, 17 (4): 373-392.

Hines, P., Nick, R. 1997. The seven value stream mapping tools. *International Journal of Operations & Production Management*, 17 (1): 46-64.

Hosen, M. A., Hussian, M. A., Mjalli, F. S. 2011. Control of polystyrene batch reactors using neural network based model predictive control (NNMPC): An experimental investigation. *Control Engineering Practice*, 19: 454-467.

Huang, J., Kaul, G., Cai, C., Chatlapalli, R., Hernandez-Abad, P., Ghosh, K., Nagi, A. 2009. Quality by design case study: An integrated multivariate approach to drug product and process development. International Journal of Pharmaceutics, 382: 23-32.

Hussain, M. A. 1999. Review of the applications of neural networks in chemical process control - simulation and online implementation. *Artificial Intelligence in Engineering*, 13: 55-68.

Hutcheson, G. D., Sofroniou, N. 1999. The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models. SAGE Publications.

Iacocca, R. G., Burcham, C. L., Hilden, L. R. 2010. Particle Engineering: A strategy for establishing drug substance physical property specifications during small molecule development. *Journal of Pharmaceutical Sciences*, 99 (1): 51-75.

ICH Q8 Pharmaceutical Development; US Department of Health and Human Services, Food and Drug Administration, Centre for Drug Evaluation and Research (CDER): Rockville, MD, 2005.

ICH Q10: Pharmaceutical Quality System. 2007.

Jackman, S. 2009. *Bayesian Analysis for the Social Sciences*. John Wiley and Sons, Chichester. 1st ed.

Johnson, N. L. 1961. A simple theoretical approach to cumulative sum control charts. *Journal of the American Statistical Association*, 56 (296): 835-840.

Jones, B., Lin, D. K. L., Nachtsheim, C. J. 2008. Bayesian D-optimal supersaturated designs. *Journal of Statistical Planning and Inference*, 138: 86-92.

Kaiser, C., Pototzki, T., Ellert, A., Luttmann, R. 2008. Applications of PAT-process analytical technology in recombinant protein processes with escherichia coli. *Engineering Life Science*, 8 (2): 132-138.

Kane, V., E. 1986. Process capability indices. *Journal of Quality Technology*, 18 (1): 41-52.

Karim, M. N., Hodge, D., Simon, L. 2003. Data-based modeling and analysis of bioprocesses: Some real experiences. *Biotechnology Progress*, 19: 1591-1605.

Kaye, B. H., Alliet, D., Switzer, L., Turbitt-Daoust, C. 1999. The effect of shape on intermethod correlation of techniques for characterizing the size distribution of powder. Part 2: Correlating the size distribution as measured by diffractometer methods, TSI-amherst aerosol spectrometer, and Coulter counter. *Particle and Particle Systems Characterization*, 16: 266-273.

Kirdar, A. O., Conner, J. S., Baclaski, J., Rathore, A. S. 2007. Application of Multivariate Analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnology*, 23: 61-67.

Köksal, G., Batmaz, I., Testik, M. C. 2011. A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38: 13448-13467.

Korakianiti, E., Rekkas, D, 2011. Statistical thinking and knowledge management for quality-driven design and manufacture in pharmaceuticals. *Pharmaceutical Research*, 28: 1465-1479.

Kosanovich, K. A., Dahl, K. S., Piovoso, M. J. 1996. Improved process understanding using multiway principal component analysis. *Industrial Chemical Engineering Research*, 35: 138-146.

Kotz, S., Johnson, N. 1993. *Process Capability Indices*. Chapman and Hall: London.

Kourti, T. 2006. The process analytical technology initiative and multivariate process analysis, monitoring and control. *Analytical and Bioanalytical Chemistry*, 384: 1043-1048.

Kourti, T., MacGregor J. F. 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28: 3-21.

Kovach, J., Stringfellow, P., Turner, J., Cho, B. R. 2005. The house of

competitiveness: The marriage of agile manufacturing, design for six sigma, and lean manufacturing considerations. *Journal of Industrial Technology*, 21 (3): 1-10.

Kumar, S. 2004. *Modelling and Abnormal Change Detection in Multivariate Signals and Systems using Subspace Projection Techniques*. Newcastle University.

LeBlond, D. 2008. Data, variation, uncertainty and probability distributions. *Journal of GXP Compliance*, 12 (2): 30-41.

LeBlond, D. 2009. Understanding hypothesis testing using probability distributions. *Journal of Validation Technology*, Winter 2009: 1-17.

Lee, P. M. 2012. *Bayesian Statistics: An Introduction*. 4th Edition. London: Arnold.

Levinson, W. A. 2010. Statistical Process Control for Real-World Applications. Hoboken: Taylor and Francis.

Liker, J. K. 2004. *The Toyota Way*. Madison, WI: CWL Publishing Enterprises.

Lippmann, R. P. 1987. An introduction to computing with neural nets. IEEE ASSP Magazine, April 1987: 4-22.

Liu, Y., Hourd, P., Chandra, A., Williams, D. J. 2010. Human cell culture process capability: a comparison of manual and automated productions. *Journal of Tissue Engineering and Regenerative Medicine*, 4: 45-54.

Lopes, J. A., Costa, P. F., Alves, T. P., Menezes, J. C. 2004. Chemometrics in bioprocess engineering: process analytical technology (PAT) applications. *Chemometrics and Intelligent Laboratory Systems*, 74: 269-275.

Lopes, J. A., Menezes, J. C. 2003. Industrial fermentation end-product modelling with multilinear PLS. *Chemometrics and Intelligent Laboratory Systems*. 68: 75-81.

Lourenco, V., Lochmann, D., Reich, G., Menezes, J. C., Herdling, T., Schewitz, J. 2012. A quality by design study applied to an industrial pharmaceutical fluid bed granulation. European Journal of Pharmaceutics and Biopharmaceutics, 81: 438-447.

Lunn, D. 2012. Introduction *to Bayesian Analysis Using WinBUGS*. Course notes, Cambridge University, May 2012.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D. 2000. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10: 325 337.

Lunney, P. D., Cogdill, R. P., Drennen, J. K. 2008. Innovation in pharmaceutical experimentation part 1: Review of experimental designs used in industrial pharmaceutics research and introduction to Bayesian D-optimal experimental design. *Journal of Pharmaceutical Innovation*, 3: 188-203.

Ma, Z., Merkus, H. G., de Smet, J. G. A. E., Heffels, C., Scarlett, B. 2000. New developments in particle characterization by laser diffraction: Size and shape. *Powder Technology*, 111: 66-78.

Ma, Z., Merkus, Henk. G., Merkus, J. G., de Smet, J. G. A. E., Verheijen, P. J. T.,

Scarlett, B. 1999. Improving the sensitivity of forward light scattering technique to large particles. *Particle and Particle Systems Characterization*, 16: 71-76.

MacGregor, J. F., Yu, H., Munoz, S. G., Flores-Cerrillo, J. 2005. Data-based latent variable methods for process analysis, monitoring and control. *Computers and Chemical Engineering*, 29: 1217-1223.

MacMurray, J. C., Himmelblau, D. M. 1995. Modeling and control of a packed distillation column using artificial neural networks. *Computers and Chemical Engineering*, 19 (10): 1077-1088.

Maltesen, M. J., van de Weert, M., Grohganz, H. 2012. Design of experiments-based monitoring of critical quality attributes for the spray-drying process of Insulin NIR spectroscopy. *AAPS PharmSciTech*, 13 (3): 747-755.

Mattila, M., Saloheimo, K., Koskinen, K. 2007. Improving the robustness of particle size analysis by multivariate statistical process control. *Particle and Particle Systems Characterization*, 24: 173-183.

Mercier, S. M., Diepenbroek, B., Dalm, M. C. F., Wijffels, R. H., Streefland, M. 2013. Multivariate data analysis as a PAT tool for early bioprocess developments data. Journal of Biotechnology, 167: 262-270.

Mockus, L., Lainez, J. M., Reklaitis, G., Kirsch, L. 2011a. A Bayesian approach to pharmaceutical product quality risk quantification. *Informatica*, 22 (4): 537-558.

Mockus, L., LeBlond, D., Basu, P. K., Shah, R. B., Khan, M. A. 2011b. A QbD case study: Bayesian prediction of lyophilization cycle parameters. AAPS PharmSciTech, 12 (1): 442-448.

Mohammed, K-J. R., Zhang, J. 2012. Reliable optimisation control of a reactive polymer composite moulding process using ant colony optimisation and bootstrap aggregated neural networks. Neural Computing and Applications, 23: 1891-1898.

Montague, G. A., Martin, E. B., O'Malley, C. J. 2008. Forecasting for fermentation operational decision making. *Biotechnology Progress* 24 (5): 1033-1041.

Montgomery, D. C., Peck, E. A., Vining, G. G. 2012. Introduction to Linear Regression Analysis. Hoboken, New Jersey: John Wiley and Sons.

Muteki, K., Swaminathan, V., Sekulic, S. S., Reid, G. L. 2011. De-risking pharmaceutical tablet manufacture through process understanding, latent variable modelling and optimization technologies. APPS PharmSciTech, 12 (4): 1324-1334.

Nabifar, A., McManus, N. T., Vivaldo-Lima, E., Penlidis, A. 2010. A sequential iterative scheme for design of experiments in complex polymerizations. *Chemical Engineering Technology*, 33: (11): 1814-1824.

Nabifar, A., McManus, N. T., Vivaldo-Lima, E., Penlidis, A. 2011. Diagnostic checks and measures of information in the Bayesian design of experiments with complex polymerizations. *Macromol. Symp.*, 302: 90-99.

Nascimento, C. A. O., Giudici, R., Guardani, R. 2000. Neural network based approach for optimization of industrial chemical processes. *Computers and Chemical Engineering*, 24: 2303-2314.

Nave, D. 2002. How to compare six sigma, lean and the theory of constraints. *Quality Progress*, 35 (3): 73-78.

Neal, R. M. 2003. Slice sampling. *The Annals of Statistics*, 31 (3): 705-741.

Neogi, D., Schlags, C. E. 1998. Multivariate statistical analysis of an emulsion batch process. *Industrial Chemical Engineering Research,* 37: 3971-3979.

Neshat, N., Mahlooji, H., Kazemi, A. 2011. An enhanced neural network model for predictive control of granule quality characteristics. Scientia Iranica E, 18 (3): 722-730.

Nomikos, P., MacGregor, J. F. 1994. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40 (8): 1361-1375.

Nomikos, P., MacGregor, J. F. 1995a. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1): 41-59.

Nomikos, P., MacGregor, J. F. 1995b. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, 30: 97-108.

Ntzoufras, I. 2009. *Bayesian Modelling Using WinBUGS*. John Wiley & Sons, New Jersey.

Orr, N. A. 1982. Calibration and standardisation of particle sizing methods. *Analytical Proceedings*, 19: 368-370.

Paliwal, M., Kumar, U. A. 2009. Neural networks and statistical techniques: A review of applications. Expert Systems with Applications, 36 (1): 2–17.

Palmer, K., Tsui, K-L. 1999. A review and interpretation of process capability indices. *Annals of Operations Research*, 87: 31-47.

Pearn, W. L., Wu, C-W. 2005. Process capability assessment for index Cpk based on Bayesian approach. *Metrika*, 61: 221-234.

Peterson, J. J. 2004. A posterior predictive approach to multiple response surface optimization. *Journal of Quality Technology*, 36 (2): 139-153.

Peterson, J. J. 2008. A Bayesian approach to the ICH Q8 definition of design space. *Journal of Biopharmaceutical Statistics*, 18 (5): 959-975.

Peterson, J. J., Miro-Quesada, G., del Castillo, E. 2009. A Bayesian reliability approach to multiple response optimization with seemingly unrelated regression models. *Quality Technology & Quantitative Management*, 6 (4): 353-369.

Polansky, A. M. 2001. A smooth nonparametric approach to multivariate process capability. *Technometrics*, 43 (2): 199-211.

Porter, L. J., Oakland, J. S. 1991. Process capability indices - An overview of theory and practice. *Quality and Reliability Engineering International*, 7: 437-448.

Puchert, T., Holzhauer, C.-V., Menezes, J. C., Lochmann, D., Reich, G. 2011. A new

PAT/QbD approach for the determination of blend homogeneity: Combination of on-line NIRS analysis with PC scores distance analysis (PC-SDA). *European Journal of Pharmaceutics and Biopharmaceutics*, 78: 173-182.

Ramprasad, Y., Patel, S., Ryali, S., Gudi, R. 2008. Prediction of batch quality indices using function space approximation and partial least squares. *2008 Americal Control Conference, Washington, USA*.

Read, E. K., Park, J. T., Shah, R. B., Riley, R. B., Brorson, K. A., Rathore, A. S. 2009. Process analytical technology (PAT) for biopharmaceutical products: Part 1. Concepts and applications. *Biotechnology and Bioengineering*, 105 (2): 276-284.

Rhodes, M. J. 1998. *Introduction to Particle Technology*. New York: John Wiley.

Richardson, J. F., Harker, J. H., Brackhurst, J, R., Coulson, J. M. 2002. Coulson and Richardson;'s Chemical Engineering. Vol 2, Particle Technology and Separation Process. 5th ed. Oxford: Butterworth-Heinemann.

Roberts, S. W. 1959. Control charts tests based on geometric moving averages. *Technometrics*, 1 (3): 239-250.

Rocha, W. F. C., Rosa, A. L., Martins, J. A., Poppi, R. J. 2010. Multivariate control charts based on net analyte signal and near infrared spectrscopy for quality monitoring of Nimesulide on pharmaceutical formualtions. *Journal of Molecular Structure*, 982: 73-78.

Sarkar, D., Doan, X., Ying, Z., Srinivasan, R. 2009. In situ particle size estimation for crystallization processes by multivariate image analysis. *Chemical Engineering Science*, 64: 9-19.

Seibert, K. D., Sethuraman, S., Mitchell, J. D., Griffiths, K. L., McGarvey, B. 2008. The use of routine process capability for the determination of process parameter criticality in small molecule API synthesis. *Journal of Pharmaceutical Innovation*, 3: 105-112.

Shaiu, J-J. H., Chiang, C-T., Hung, H-N. 1999. A Bayesian procedure for process capability assessment. Quality and Reliability Engineering International, 15: 369-378.

Shekunov, B. Y., Chattopadhyay, P., Tong, H. H. Y., Chow, A. H. L. 2007. Particle size analysis in pharmaceutics: Principles, methods and applications. *Pharmaceutical Research*, 24 (2): 203-227.

Shewhart, W. A. 1931. *Economic Control of Manufactured Product*. D. Van Nostrand. New York.

Shi, Z., Zaborenko, N., Reed, D. E. 2013. Latent variables-based process modelling of a continuous hydrogenation reaction in API synthesis of small molecules. Journal of Pharmaceutical Innovation, 8: 1-10.

Simões, S., Sousa, A., Figueiredo, M. 1996. Dissolution rate studies of pharmaceutical multisized powders - a practical approach using the Coulter method. *International Journal of Pharmaceutics*, 127: 283-291.

Spiring, F. A. 1995. Process capability: A total quality management tool. *Total Quality Management*, 6 (1): 21-34.

Sridhar, D. V., Seagrave, R. C., Bartlett, E. B. 1996. Process modelling using stacked neural networks. *AIChE Journal*, 42 (9): 2529-2539.

Stockdale, G. W., Cheng, A. 2009. Finding design space and a reliable operating region using a multivariate Bayesian approach with experimental design. *Quality Technology & Quantitative Management*, 6 (4): 391-408.

Stoumbos Z. G., Reynolds, M. R., Ryan, T. P., Woodall, W. H. 2000. The state of Statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, 95 (451): 992-998.

Sukthomya, W., Tannock, J. 2005. The training of neural networks to model manufacturing processes. *Journal if Intelligent Manufacturing*, 16: 39-51.

Taguchi, G. 1987. System of Experimental Design. New York: White Plains.

Tang, L. C., Than, S. E. 1999. Computing process capability indices for non-normal data: a review and comparative study. *Quality and Reliability Engineering International*, 15: 339-353.

Tinke, A. P., Carnicer, A., Govereanu R., Scheltjens, G., Lauwerysen, L., Mertens, N., Vanhoutte, K., Brewster, M. E. 2008. Particle shape and orientation in laser diffraction and static image analysis size distribution of micrometer size rectangular particles. *Powder Technology*, 186: 154-167.

Tinke, A. P., Govoreanu, R., Weuts, I., Vanhoutte, K., De Smaele, D. 2009. A review of underlying fundamentals in a wet dispersion size analysis of powders. *Powder Technology*, 196: 102-114.

Tolve, A. 2009. How lean manufacturing can cut costs [online]. Available from: http://social.eyeforpharma.com/uncategorised/how-lean-manufacturing-can-cut-costs [accessed 31/07/2014].

Tomba, E., De Martin, M., Facco, P., Robertson, J., Zomer. S., Bezzo, F., Barolo, M. 2013. General procedure to aid the development of continuous pharmaceutical process using multivariate statistical modelling – An industrial case study. International Journal of Pharmaceutics, 444: 25-39.

van de Ven, P., Bijlsma, S., Gout, E., Maarschalk, K, v. d. V., Thissen, U. 2011. A framework for efficient process development using optimal experimental designs. *Journal of Pharmaceutical Innovation*, 6: 24-31.

van den Berg, J., Curtis, A., Trampert, J. 2003. Optimal nonlinear Bayesian experimental design: an application to amplitude versus offset experiments. *Geophysics Journal International*, 155 (2): 411-421.

Vellido, A., Lisbao, P. J. G., Vaughan, J. 1999. Neural networks in business: a survey of applications (1992-1998). *Expert Systems with Applications*, 17: 51-70.

Vinzi, V. E . 2010. Handbook of partial least squares concepts, methods and applications. Berlin, New York: Springer.

Vivaldo-Lima, E., Penlidis, A., Wood, P. E., Hamielec, A. E. 2006. Determination of the relative importance of process factors on particle size distribution in suspension polymerisation using a Bayesian experimental design technique. *Journal of Applied Polymer Science*, 102: 5577-5586.

Watson, I., Marir, F. 1994. Case-based reasoning: a review. *Knowl Eng Rev,* 9: 327-354.

Wilcox, J. A. D., Wright, D. T. 1998. Towards pultrusion process optimisation using artificial neural networks. *Journal of Materials Processing Technology*, 83: 131-141.

Wold, S., Esbensen, K., Geladi, P. 1987a. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2: 37-52.

Wold, S., Geladi, P., Esbensen, K., Ohman, J. 1987b. Multi-way principal components-and PLS-analysis. *Journal of Chemometrics*, 1: 41-56.

Wold, S., Kettaneh-Wold, N., Skagerberg, B. 1989. Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 7: 53-65.

Wold, S., Sjöström, M., Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58: 109-130.

Wolpert, D. H. 1992. Stacked Generalization. *Neural Networks*, 5: 241-259.

Womack, J. P., Jones, D. T. 2003. *Lean Thinking*. 2nd ed. New York, NY: Free Press.

Woodall, W. H. 2000. Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32 (4): 341-350.

Wu, C-W., Pearn, W. L., Chang, C. S., Chen, H. C. 2007. Accuracy analysis of the percentile method for estimating non-normal manufacturing quality. *Communications in Statistics – Simulation and Computation*, 36 (3): 657-697.

Yacoub, F., Lautens, J., Lucisano, L., Banh, W. 2011. Application of quality by design principals to legacy drug products. Journal of Pharmaceutical Innovation, 6: 61-68.

Yang, Z., You, W., Ji, G. 2011. Using partial least squares and support vector machines for bankruptcy prediction. Expert System with Applications, 38: 8336-8342.

Yu, L. X. 2008. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. *Pharmaceutical Research*, 25 (4): 781-791.

Yu, W., Hancock, B. C. 2008. Evaluation of dynamic image analysis for characterizing pharmaceutical excipient particles. *International Journal of Pharmaceutics*, 361: 150-157.

Zahid, M. A., Sultana, A. 2008. Assessment and comparison of multivariate process capability indices in ceramic industry. *Journal of Mechanical Engineering*, 39 (1): 18-25.

Zhang, J. 1999. Developing robust non-linear models through bootstrap aggregated

neural networks. *Neurocomputing*, 25: 93-113.

Zhang, J., Martin, E. B., Morris, A. J., Kiparissides, C. 1997. Inferential estimation of polymer quality using stacked neural networks. *Computers & Chemical Engineering*, 21: s1025-s1030.

Zorriassatine, F., Tannock, J. D. T. 1998. A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing*, 9: 209-224.