A FRAMEWORK FOR ESTABLISHING AN EXPERIMENTAL DESIGN APPROACH IN INDUSTRIAL DATA MINING

W. H. van Blerk

(Student Number: 11098651

Thesis submitted in fulfilment of the requirements for the degree

PHILOSOPHIAE DOCTOR

in

Operational Research

in the

School of Information Technology

in the

FACULTY OF ECONOMIC SCIENCES AND INFORMATION TECHNOLOGY

at the

NORTH-WEST UNIVERSITY

(VAAL TRIANGLE CAMPUS)

Supervisor: Prof. P.D. Pretorius

November 2016

DECLARATION

I declare that the dissertation titled

A framework for establishing an experimental design approach in industrial data mining

is my own work, and that all the resources used or quoted have been duly acknowledged by means of in-text citations and complete references, and that I have not previously submitted the dissertation for a degree at any other university.

……

W.H. van Blerk

November 2016

Vanderbijlpark

H C Sieberhagen
SATI no 1001489
Hettie.Sieberhagen@nwu.ac.za

Translator and Editor
082 3359846
018 2994554

CERTIFICATE
of
LANGUAGE EDITING

issued on 17 November 2016

I hereby declare that I have edited the language of the PhD thesis

A FRAMEWORK FOR ESTABLISHING AN EXPERIMENTAL DESIGN APPROACH IN INDUSTRIAL DATA MINING

by

W. H. van Blerk

STUDENT NUMBER: 11098651

submitted in fulfilment of the requirements for the PhD degree in the School of Information Technology at the Vaal Triangle Campus of the North-West University

The responsibility to accept recommendations and effect changes remains with the author.

H C Sieberhagen
SATI no 1001489
ID 4504190077088

Potchefstroom
17 November 2016

ACKNOWLEDGEMENTS

To my wife and son, without their unselfish support during this journey, this endeavour would never be possible. Thank you.

There is a plethora of angles to data analysis applying an analytical technique. To find the best fit to purpose is extremely difficult. Without Professor Phillip Pretorius's guidance in this respect, it would have been close to impossible to find a purposeful approach. As an expert, he constantly provided out of the box thinking to keep me focused and on track.

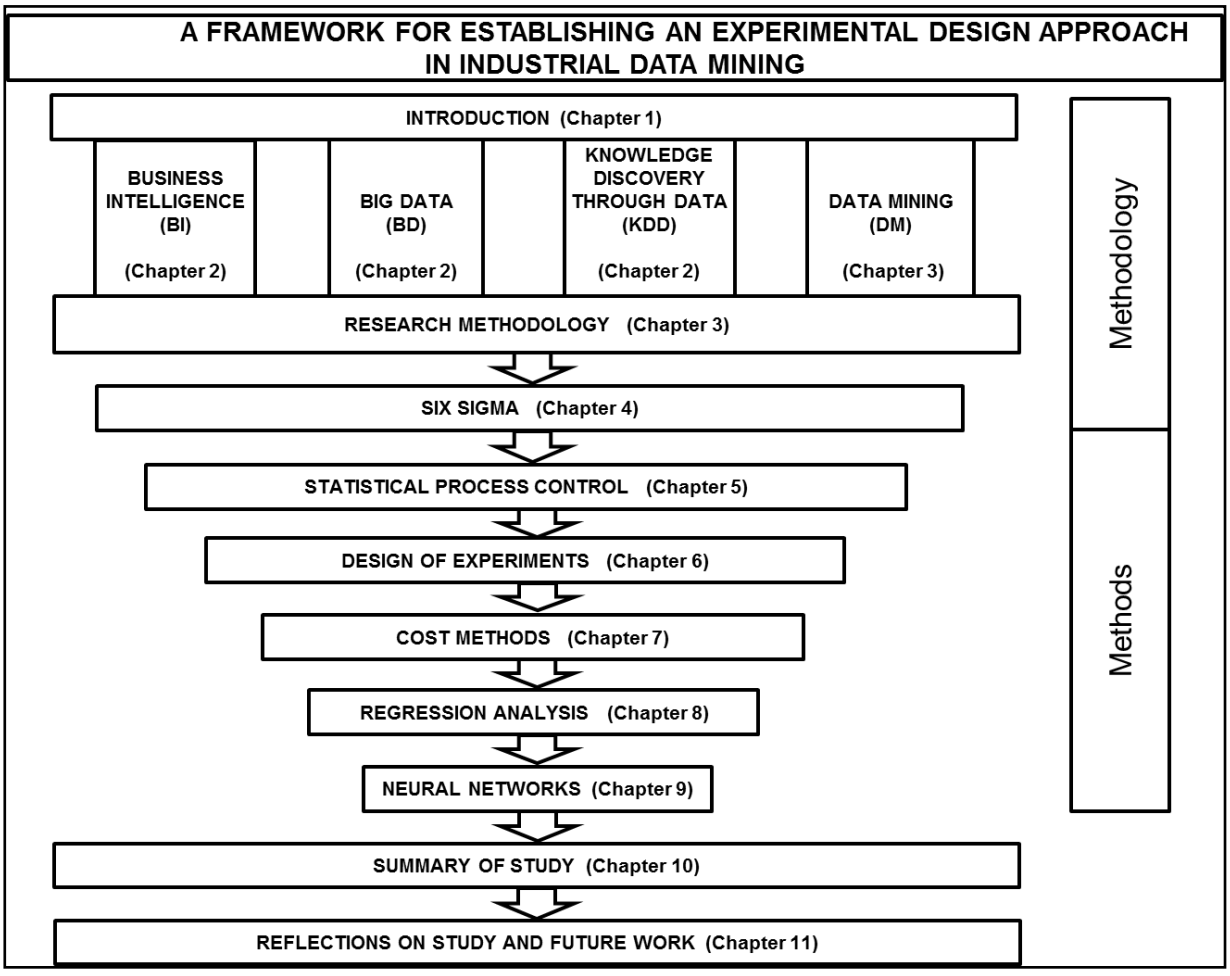
To Mrs Martie Esterhuizen from the NWU library that humbly volunteered her personal time assisting me in finalising the technical referencing for this document.

To Mrs Aldine Oosthuyzen from the NWU in assisting me in finalising the formatting for this document. An expert indeed.

To Mrs Hettie Sieberhagen for the language editing for this document.

ABSTRACT

This research was conducted due to a need in a specific industrial environment to provide a structured problem solving approach, which accommodate DOE within a framework, assisting analysts and management for strategic decision making for process improvement. Building an experimentation model within an industrial manufacturing environment is the base for the developed framework that consists of two main components; one being methodologies, showing the high level non-analytical portion of the framework, and two; selected statistical methods, including DOE, which represent statistical techniques for scientific data analysis following a sequential statistical technique selection for the data analytical process. The proposed framework and the sequential data analytical process is presented in the diagram below.



This framework is the blueprint for this research which was applied to a case study to evaluate the pragmatic relevance of this framework. Specific goals were set for the research that were aligned with the proposed framework. These goals are;

To accommodate DOE as a Data Mining Technique in an Industrial Data Mining environment.

To enhance the awareness of expanding DOE as a statistical approach to complement existing methods and methodologies used for Data Mining.

To validate the integrity of captured data through the refining process to determine upper and lower operating conditions required by DOE, any abnormal data points will be exposed.

To focus on Industrial Data Mining, and concentrate on process data, applying DOE rather than generic, traditional Data Mining techniques.

To develop a methodology to accommodate the use of DOE as a Data Mining technique to determine impacts of variables on process outcomes through experimenting with data within current databases.

The main purpose by meeting the above goals at the end of the study shows that the analytical process illustrates:

That the proposed framework is generic, applicable to this case study and for any data analytical process.
That the focus is on process improvement with experimentation as a process improvement basis.
Shows an alternative perspective to data analytics by utilizing historic data within databases by applying experimentation to reduce the impact of experimentation cost.

Applying the proposed framework for process optimisation studies in any company where needed should enhance process improvement, because this research is about following a new experimental analysis design approach that is generic for any process development and improvement, irrespective of the product rendered. The framework and techniques used in this research are applicable within any processing plant where multiple variables affect product quality. The proposed model for process development could not be tested because the company has shut its operations in South Africa but the concept for the proposed DOE methodology proved to be representative for the period upon which the model was developed and tested, based on all the different the comparative results between the predictive model and the validation period.

TABLE OF CONTENTS

LIST OF DIAGRAMS

LIST OF FIGURES

LIST OF TABLES

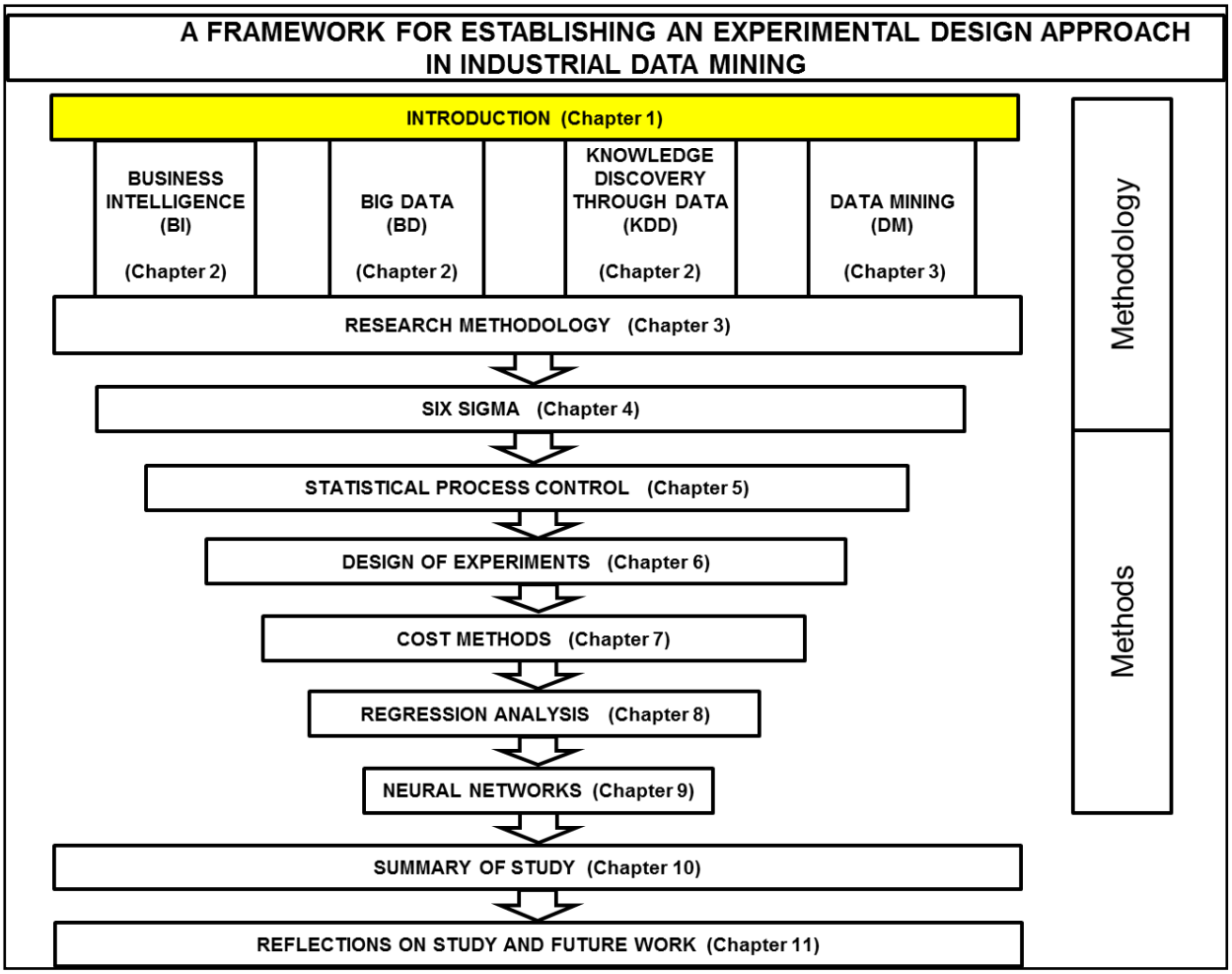
LIST OF GRAPHS

LIST OF ABBREVIATIONS

OR	Operational Research
DOE	Design Of Experiments
MR	Multiple Regression
BI	Business Intelligence
BD	Big Data
DMAIC	Define Measure Annalise Improve Control
SPC	Statistical Process Control
DM	Data Mining
KDD	Knowledge Discovery through Data
RA	Regression analysis
NN	Neural Networks
SS	Six Sigma
Normal like	Look like a normal distribution, but is not a normal distribution. It is close to a bell-shaped distribution but not normal distributed yet or is on its way. It is visual evaluation of data distribution.

CHAPTER 1

INTRODUCTION



1.1 PRE-AMBLE

There was a need in a specific industrial environment to provide a structured problem solving approach, which accommodates DOE within a framework, assisting analysts and management for strategic decision making for process improvement. Building an experimentation model within an industrial manufacturing environment is the base for the developed framework that consists of two main components; one being methodologies, showing the high level non-analytical portion of the framework, and two; selected statistical methods, including DOE, which represent statistical techniques for scientific data analysis following a sequential statistical technique selection for the data analytical process.

The proposed framework is generic to any data analytical process that focuses on process improvement with experimentation as a process improvement basis. It also shows an alternative perspective to data analytics by utilizing historic data within databases by applying experimentation to reduce the impact of experimentation cost.

Data Mining (DM), traditionally concentrates on large databases, millions of records or data in some form or another, which with normal exploratory techniques is difficult to access. It is even more difficult to find meaning from these databases.

DM from an industrial environment perspective may have changed the perspective of large databases that need to be examined by the DM techniques to a more focused approach that is process specific in order to find the true contributors to quality variation and the challenge of operating parameters.

Knowledge Discovery in Databases (KDD), a wider approach than DM, suggests that DM is a step in the process of knowledge discovery. Wright (1998) acknowledges that KDD is a growing field that focuses on non-trivial extraction of implicit, unknown and potentially useful information from data, where traditional DM focuses on pattern and trend recognition. Different approaches for KDD are: Probabilistic, Statistical, Classification, Data-Cleaning and Decision Trees. These approaches are typical for data filtering and structuring to be accessed by DM approaches.

Fayyad *et al.* (1996:39) acknowledge that KDD is the process for knowledge discovery where DM is part of this process. They describe the KDD process as a nine-step process:

1. Define the goal of the KDD process from the customer's perspective.
2. Create the target data set.
3. Clean and pre-process data.
4. Reduce number of variables effectively through data reduction process.
5. Match KDD goal to specific DM methods.
6. Analyse explore and select model
7. DM: Search for patterns and trends.
8. Interpret results.
9. Act on discovered knowledge.

The above process clearly shows that DM and KDD is part of a holistic approach to knowledge discovery, which is an integral part of the proposed framework with the focus on DOE as a useful addition to the analytical approach.

This research will concentrate on categorizing Design of Experiments (DOE) as an effective DM technique. The power of quantitative and qualitative benefits of using this technique will also be realized for strategic decision makers.

1.2 PROBLEM STATEMENT

Of all the traditional statistical techniques used for Data Mining, Design of Experiments (DOE) is not used as a accepted Data Mining technique. DOE is utilised to scientifically determine how an input affects outputs, and then to use this knowledge to optimize processes.

Management has generally underutilised the use of statistical techniques to uncover nuggets in large databases. These buried nuggets in databases, which appear in many forms. General guidelines what to look for are:

- Association patterns between variables and outcomes, which were previously unknown. Identifying associations is not easy; it needs a focused and well-defined search methodology to identify patterns.
- Do not automatically disqualify Outliers as “bad” data points when identified through exploratory analysis. It could be the unexpected discovery of valuable information. Treat non-normal data points as part of natural variation and analyse them accordingly.
- Group associated data types for segmenting the area of research; this will assist in reducing the risk of cross-pollinating variables.

Typical statistical techniques used for mining data are:

- Pareto Analysis
- Multiple Regression Analysis
- Scatter Diagrams
- Cluster Analysis
- Discriminant Analysis
- Line Graphs
- Simple Regression Analysis
- Box Plots
- Bar Graphs
- Factor Analysis
- Principal Components
- Canonical analysis

- Classification trees
- Distribution fitting
- Multidimensional scaling
- Process capability analysis
- Correspondence analysis

The objective is to develop a framework that focuses primarily on the industrial process applications in a manufacturing environment. The proposed approach is different from the traditional Experimental design approach in that historical data accumulated in traditional databases are used to determine effects of variables on different outputs.

1.3 DELIMINATION OF FIELD

The application field for this study is a specific process from a local company showing how the Experimental design approach was done through empirical methods and methodologies. It represents the environment for this research as well as the management structure that supports this environment. Here the quantitative research boundaries are defined for this study.

1.3.1 Research environment

This research was done at a local company, which forms part of five international manufacturing plants worldwide. These plants are all a blue print of one another, therefore any processing changes or product enhancements are globally implementable. This group of companies is a listed as one company on the New York stock exchange. This group of companies is the single largest producer of their product worldwide and they maintain a dominant footprint in this market. The core business is to provide their product to the steel industry for electrical arc furnaces to smelt steel. There are many products that can smelt steel but this product is currently the only cost effective product for an electrical arc furnace to smelt steel.

The research will concentrate on process data for the manufacturing of this product at a local facility. There are five sequential manufacturing processes to produce this product, where one only contributes to dimensional specifications. The remaining four processes are batch processes that flow sequentially from raw material processing to machining. A determining factor for this research is the long processing time of three months to process one product. This manufacturing process does not allow the privilege of evaluating results quickly for every corresponding process parameter change. For this

reason, the risk of producing non-conforming products when operational changes are made, is high and very costly.

The proposed framework is an analytical process starting from database exploration to statistical techniques application, specifically including of designs of experiments (DOE). The reference database is process data between 2005 and 2013, specific to one of the four production processes. Subsequently 2005 – 2009 is used for the DOE application, and 2010 - 2013 will be used as a verification and validation period for results.

1.3.2 Empirical analysis

The research will only concentrate on processing data for a specific product manufactured in a South African facility. There are five processes, where one is a machining operation that only contributes to dimensional specifications. The proposed framework will be tested on historical process data obtained between 2005 and 2009 for at least one of the remaining four process departments; 2010 - 2013 will be used as verification of results before normal experimentation parameters are set for future experimentation. Once the verification process has proved that experimentation on historical data is viable and statically accurate, the approach may be adopted by similar processing environments.

1.3.3 Management structure

The management structure for the company within which the research is done is reflected in Diagram 1.1. Although the nine management functions are treated as a unit which ensures the company's strategic direction, this research focuses only on the Quality Assurance function, and within this function particularly on process optimization and new product development. The referenced database is applicable to this function as well.

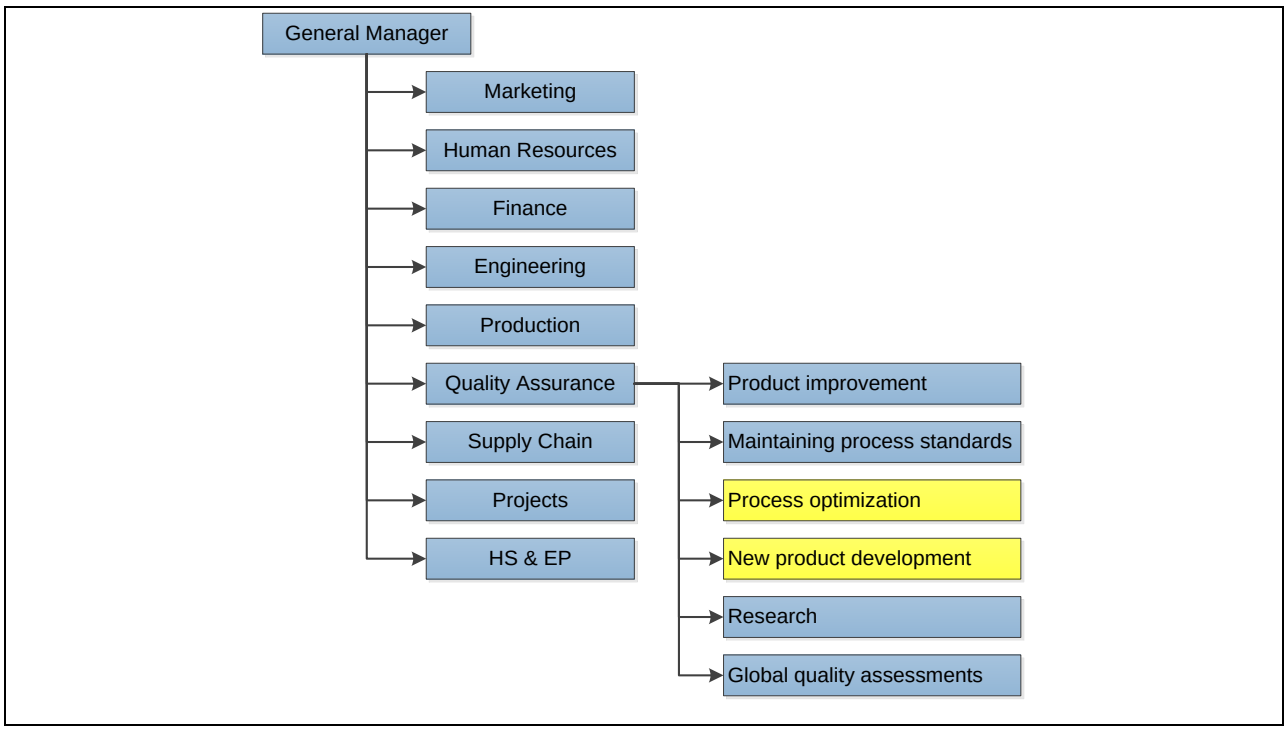


Diagram 1.1: Company management structure

This product contributes ± 97% of the total revenue, management reports are well entrenched and highly developed. A sophisticated database is the foundation from which these reports operate. A database designed by "J D Edwards one world Enterprise Software" provides data sets to the global manufacturing industry. This software accommodates Finances, Manufacturing, Shipping and billing, Sales order processing, Purchase order processing, Marketing, Product processing data, Operational processing data and Global product processing data interface.

Applicable management reports generated from this software that relate to the quality function are: Monthly/weekly/quarterly/annual product processing reports, Product quality trends, Product performance within all market segments related to this specific product and the Monitoring of different product grades used in different markets.

Even though this product market is the largest of the various products produced, the quality aspect of this product is crucial for the existence of this company, and also for the consumers. The database for this segment of the company must be explored for understanding before it is used blindly, in order for data analysis to improve product quality and or product processing. Although the main goal of this study is to provide a framework within the quality function of the company to assist in improving product

quality, the integrity of processing data within the referenced database is of equal importance.

1.4 CRITICAL TERMS

Although a plethora of terms exist, only terms that were critical for the development of the framework were selected for this research.

1.4.1 Business intelligence

BI has evolved into a multi-dimensional data processing environment where technologies, methodologies, processes and different architectures are utilised for value added process analysis for management strategic decision-making. Unless businesses have an integrated system to analyse and control this ever-increasing volume of data, the accumulated data will be only data and nothing else. Companies that utilise a well-designed BI structure and BI tools to manage accumulated data in a structured way to provide management with meaningful information to make informed decisions, will experience a high competitive advantage amongst their peers in similar industries.

Koch (2015:56) states that by 2016, 70% of high-performing companies will be integrating real-time predictive analytics into their business operations. This gives a new dimension to BI in that interactive predictive analysis will become a strategic competitive advantage for companies, which needs to adapt to constant market changes and new operational challenges. Kopčeková *et al.* (2013:43) and Lasi (2013:387) clearly indicate that BI is a broad category of data analytical methods and data base technologies for gathering, storing, analysing, and providing access to data to assist management in all levels of an organization to improve their business decisions. The importance of a stable, integrated data warehouse is becoming increasingly critical for BI as a supportive system for management. BI provides this environment that is crucial for process improvement analysis.

1.4.2 Knowledge discovery through data

Knowledge discovery through data (KDD) focuses on the development of methods and analytical techniques for making sense of data for strategic management decision making. This enhances the goal of KDD to extract high-level knowledge from low-level data of large data sets.

Horníková *et al*. (2011:12) state that KDD was developed first, and data mining later. It makes sense in that it seems that originally management focused on knowledge within specified processes and business segments.

1.4.3 Data mining

Data mining (DM) concentrates mainly on process industry, but is also relevant in other sectors, like the service industry. Management and analysts, through DM, challenge operating conditions, environment, raw materials, process changes and traditional analytical methodologies to investigate alternative operating conditions.

Data mining is an integrated process of various data analysis disciplines and methodologies, and is not a stand-alone analytical discipline that provides a business solution to management. Petre (2013:27) describes DM as a methodology which compares to a typical DM approach such as define the problem, get data, transform the data, determine which analytical technique is appropriate to the problem, analyse the data, review results then implement selected results.

DM in general is projected as a framework consisting of four major stages: data accumulation, product family classification, design retrieval and modification (Yu & Zhang, 2014:2).

Web data mining became a new extension to DM for mining WEB based data. Because of the vast amount of data and multiple data sources, a big disadvantage is discovering questionable data source whose integrity cannot be proven. “The bigger the better” in data accumulation must be handled with caution, and for this reason internal databases are better managed and controlled for data integrity.

1.4.4 Big data

Big Data (BD) is a collection of databases and data generated internally from all processes within an organization, it also includes other potential data sources that are not so obvious to recognise, or not seen as data. Weber (2013:20) refers to these data sources as dark data that are internal data collected in the course of doing business, often in archives or generally not accessible. These may include internal unstructured data that comprise memos, reports, client notes, call centre recordings, meetings, videos, machine data that are internal or external to the company, databases external to

the company, high volume data due to the usability provided by mobile devices and sensors, the ability to create new data is convenient, easy and simple. The connectivity of these devices with other people, machines, and networks allows data to be easily shared and replicated, further increasing the volume. High velocity data with the same usability and connectivity may turn weekly/daily analyses, assessments, and feedback into real time. Think of how quickly we can forward e-mails, pictures, tweets, and and a vast variety of data in many shapes and sizes, especially external to the company.

1.4.5 Six Sigma

Six Sigma (SS) has become an integral part of modern business because the recognition of waste within all processes and functions in an organization is a priority to eliminate for survival. Six Sigma follows a strict statistical system and a management philosophy acknowledged by Otero *et al*. (2012:934) and Surange (2015:283), that is defined as DMAIC (Define Measure Analyse Improve Control), which is strict and focused, and makes use of specific quality tools. Kučerová and Fidlerová (2014:148) mention these quality tools to be cause-effect charts and statistical process control charts.

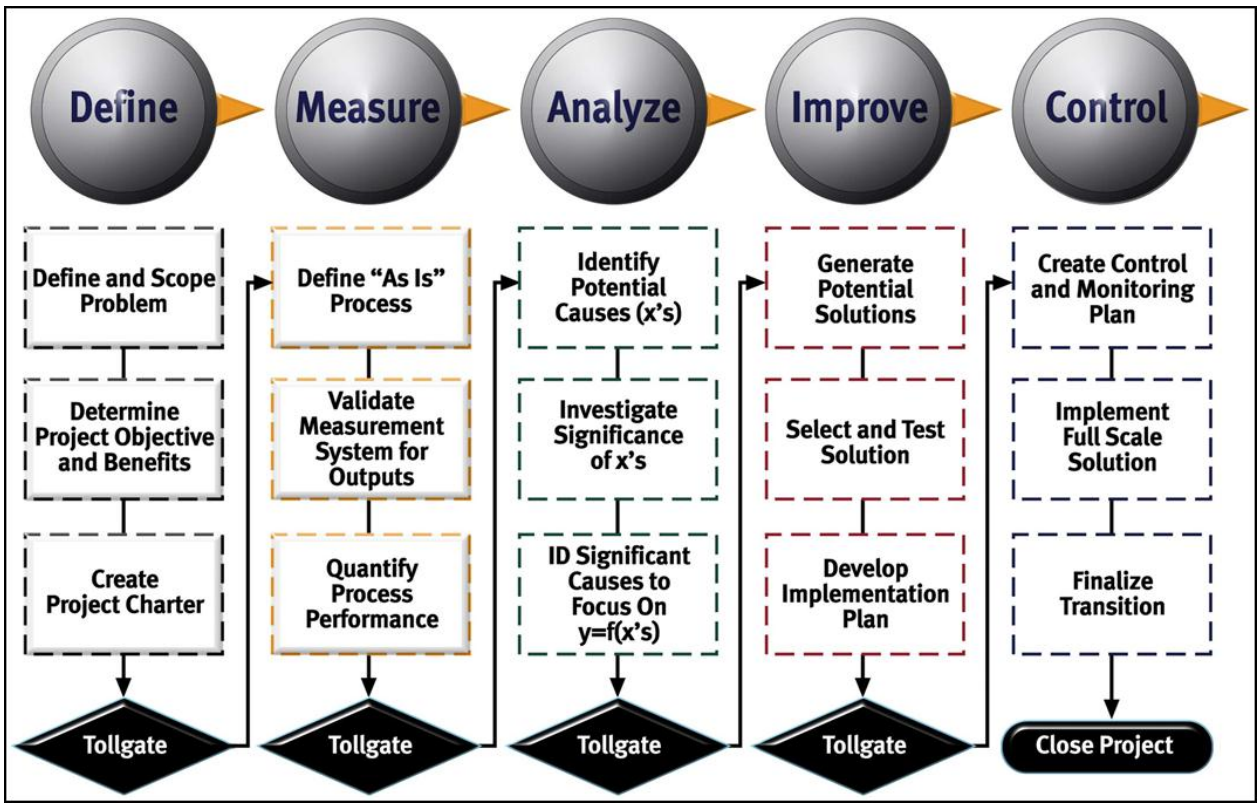


Diagram 1.2: DMAIC methodology

Six Sigma is based on the DMAIC methodology, which is a structured data analytical process for process improvement. Refer to Diagram 1.2 that shows the five steps for improvement process. Applying the case study the DMAIC methodology flows through the chapters as follows:

D (Define)	Chapter 3: Research Methodology Chapter 4: SS
M (Measure)	Chapter 5: SPC
A (Analize)	Chapter 6: DOE
I (Improve)	Chapter 7: Cost Methods
C (Control)	Chapter 5: SPC Chapter 8: MR Chapter 9: NN

1.4.6 Statistical process control

Render *et al.* (2012:623) describe Statistical Process Control (SPC) as a statistical tool to help set standards as well as monitor, measure, and correct quality problems. Quality control charts are the basis for SPC, which is used to measure and control processes. Developing of SPC charts follows five basic steps:

- **Select the process to be measured.** This process can be quantitative or qualitative.
- **Statistically measure all appropriate variables applicable to the identified process.**
- **Select the appropriate SPC chart** then apply it to the selected process.
- **Track the variable through the SPC chart** and observe for patterns like, trends, shifts, clusters, non-normal variation in new values.
- **Only make adjustments** if necessary.

Xie and Kruger (2012:3) stipulate that SPC has been introduced into the general manufacturing industries for monitoring process performance and product quality and to monitor the general process variation that is caused by a few key process variables.

SPC is used as a statistical control technique from a process perspective. For this research, it also assists in screening the critical few independent variables for DOE analysis.

1.4.7 Multiple regression

Multiple Regression (MR) analysis has been a critical part of statistical techniques used through the years, specifically when trying to find relationships amongst independent variables and a dependent variable. The basis for regression analysis is to fit models for a dependent variable as a function of one or more independent variables. Regression analysis complements designed experiments in predicting so far the behaviour of the dependent variable through selected independent variables. In this research, multiple regression analysis will be applied to compare multiple regression analysis to designed experiment model regression.

1.4.8 Design of experiments

Design of Experiments (DOE) is a statistical technique, which comprises data driven to enable designers or experimenters to determine the effects of many factors that affect output results of any process. It is very useful in assisting in robust design that accommodates most uncontrollable independent variables and fixes them prior to going into production.

DOE was popularised by Box, Hunter, *et al*. (1978), Box and Draper (1969) and Mason *et al.* (1989) who primarily discuss designs with many factors that estimate process outcome effects with a minimum number of observations. DOE is a powerful statistical technique that focuses on evaluating contributions from independent variables (factors) to the effect on output variables (responses) based on a scientific approach by purposefully changing input variables to evaluate the impact on process outcomes in a controlled experimentation environment. Different designed scenarios with factors create an exploration of the process analysed. The analyst gains valuable process knowledge through this exploration process for process optimization and improvement. Care should be taken to ensure that the DOE design is correct and applicable to the

process analysed. If not, the opposite effect may be experienced, e.g. increased process variation, increased process cost and reduced market share. This summarized view for process improvement through DOE emphasize the importance of exploring new processes, of controlled experimentation, the effects of variables on process outcomes, of DOE as a scientific approach, the importance of correct DOE design and of gained process knowledge for the analyst (Antony, 2014:36; Cano *et al*. 2012:201; Launsby & Schmidt, 1991:2; Peterka, 2008a; Peterka, 2008b; Sundararajan, 2010:67).

Launsby and Schmidt (1991:2) describe it as:

"Experimental design as a scientific approach of purposefully change inputs to a process to evaluate the changes in the outputs".

Phadke (1989:11) refers to variation contributors as noise factors that are classified as External (Variation external to the process like, weather, raw materials, changing in seasons), Unit-to-unit variation (measured quality characteristic differences between processed products) and Deterioration (equipment mechanical deterioration causing erratic results or unexplainable trends). These factors cannot be controlled by the designer, but can only be statistically measured to evaluate the effect on the response variable.

Khuri and Cornell (1987:1) believe that most exploratory investigations involving DOE have a twofold purpose:

1. Quantifying the relationships between response variables and settings for experimental factors that affect the response variables.
2. Determining the settings of these experimental factors that produce the best values for the response variable.

Common terminology used for Experimental design studies is vital for understanding the scientific approach when analysing and communicating the design, analysing process and results. Mason *et al*. (1989:92) define these terminologies as:

- Block: Groups of homogeneous experimental runs.
- Confounding: One or more effects that cannot be attributed to a single factor or interaction.

- Covariate: An uncontrollable variable that influences the response but is unaffected by any other experimental factors.
- Design: Complete specification of experimental runs.
- Effect: Change in the average response between two factor-level combinations.
- Experimental region: Window of experimentation upon which design is built.
- Factor: A controllable factor that is thought to influence the response.
- Interaction: Existence of joint factor effects in which the effect of one factor depends on the level of the other factor.
- Level: Specific experimental value for a factor.
- Repeat tests: Two or more observations for the same experimental factor levels.
- Replication: Repetition of an entire experiment under two or more sets of conditions.
- Response: Outcomes or result of experiment.
- Test run: Running a single combination of a selected factor level to compare its responses to actual yield responses.

1.4.9 Scatter graphs

StatTrek (2014) defines a scatter plot as a kind of mathematical diagram utilising Cartesian coordinates to show values for bi-variates for a set of data using horizontal and vertical axes.

The more closely the data points get when plotted to make a linear line, the higher the correlation and the stronger the linear relationship. Correlation represents two ways, a negative or positive correlation between the data points plotted. The dependency of variables in a plotted data point determines the relationship of the data to be analysed.

1.4.10 Neural networks

The ability of Neural Networks (NN) to learn by example is one of the many features that enable the analyst to model data and establish accurate rules governing the underlying relationship between various data attributes. Neural network uses training algorithms, which can automatically learn the structure of the data presented by the

analyst. This unique analytical feature of neural networks makes it a popular DM technique for analysts as a predictive model.

1.5 GOALS OF THE STUDY

1.5.1 Primary goals

- To accommodate DOE as a Data Mining Technique in an Industrial Data Mining environment.
- To enhance the awareness of expanding DOE as a statistical approach to complement existing methods and methodologies used for Data Mining.
- To validate the integrity of captured data through the refining process to determine upper and lower operating conditions required by DOE, any abnormal data points will be exposed.
- To focus on Industrial Data Mining, and concentrate on process data, applying DOE rather than generic, traditional Data Mining techniques.
- To develop a methodology to accommodate the use of DOE as a Data Mining technique to determine impacts of variables on process outcomes through experimenting with data within current databases.

1.5.2 Secondary goals

- To reduce the risk of costly process failures owing to experimenting into the unknown by firstly dividing the historical dataset into two periods, where period one is statistically analysed to predict the validation period of the historical dataset by applying the DOE methodology approach. Financial risk plays a key role in strategic decision making; when this risk is scientifically reduced, management will be more willing to include process development as a key strategic focus.
- To improve data integrity by utilising the methodology; validating the integrity of data will consequentially lead to data credibility.
- To provide an alternative statistical approach for data mining in an industrial environment for process development/improvement through screening of independent variables during the define stage, and not finalising the critical variables during the traditional improvement stage.

- To enhance and categorise DOE as an effective data mining technique to complement existing methods and methodologies used for statistical data analysis for process improvement.

1.6 IMPORTANCE OF THE STUDY TO THE FIELD OF OR

Hamdy (2007:2) describes the OR (Operational Research) field as studies that consists in building a model of the physical situation which means that a OR model can be defined as a simplified representation of a real-life system. The complexity of a real system results from large amount of variables that control the behaviour of a system. Also, the assumed real world is abstracted from the real world by concentrating on the dominant variables that control the behaviour of the real system.

Against this OR background this study do make a positive contribution to the OR field. A framework was introduced that integrates methodologies and methods that includes literature and a case study by both for model building to refine the present situation (case study), identifying all controlling variables that controls process behaviour within the case study and identify the major variables from all variables that significantly impact process behaviour.

A brief overall summary of study contributions that were identified as important to this study that also fits within the OR field :

- To experiment with historical data based on real process data.
- Full and fractional design scenarios. This allows the analyst not to have a one-dimensional analytical approach but to evaluate which design fits data best.
- No risk of costly process failures due to experimenting into the unknown.
- To provide an alternative statistical approach for Data Mining in an industrial environment for screening independent and dependent variables for DOE model usage.

When this study was completed, the lessons learned, see section 11.2 is an important part of OR because this is part of the learning curve shared with future analists.

1.7 LAYOUT OF CHAPTERS TO FOLLOW

Flow Diagram 1.3 shows the flow of the chapters for this research.

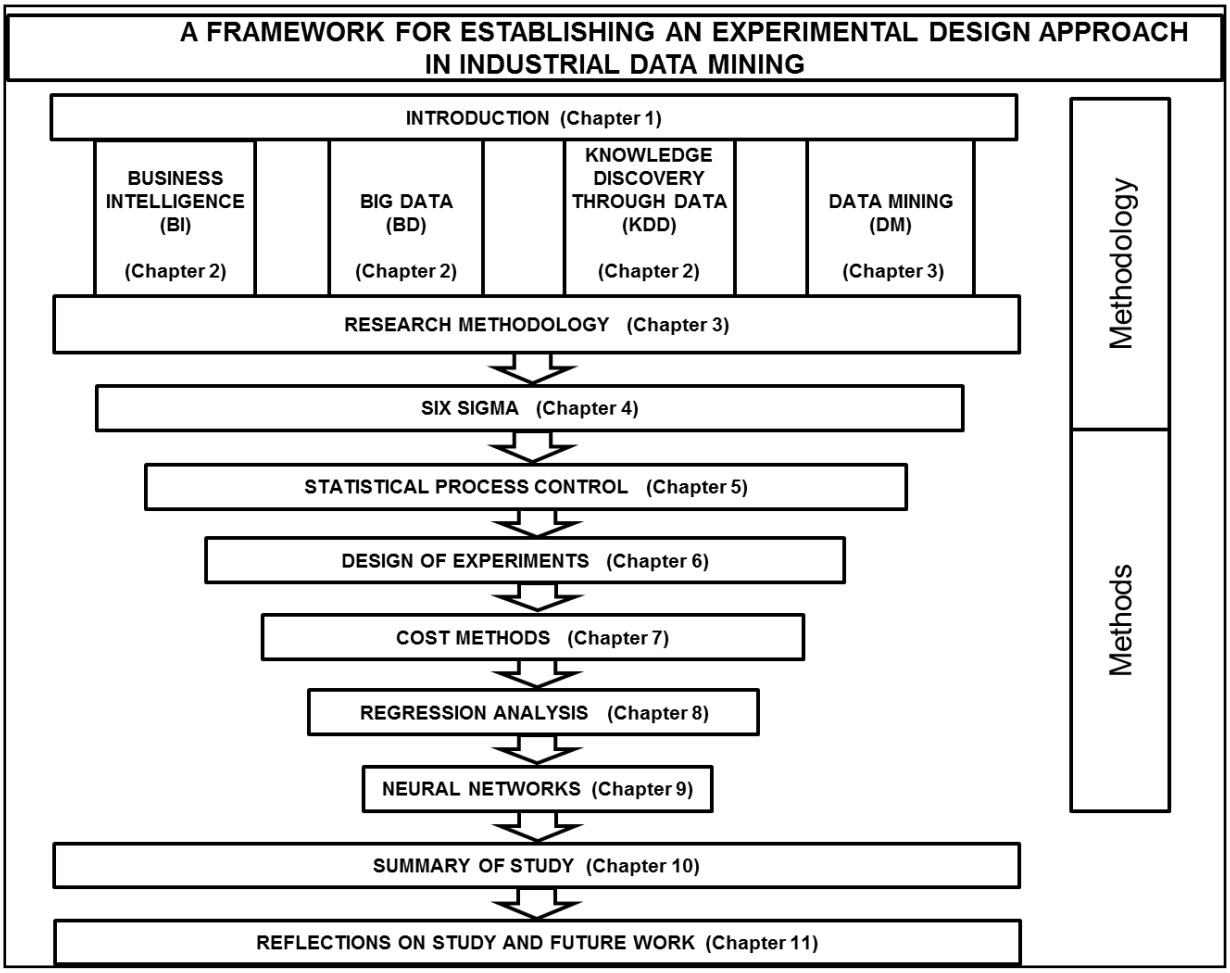


Diagram 1.3: Outlay of chapters

CHAPTER 1: INTRODUCTION

This chapter represents the research proposal, outlining the scope for this research.

CHAPTER 2: BUSINESS CONTEXT

Big data (BD), Knowledge discovery through data (KDD) and Business Intelligence (BI) are discussed as the main data frameworks that form an integral part data analytics. Although the focus is on Data Mining (DM) for this study, it is discussed in chapter 4; BD forms an integral part in complementing DM in terms of analysing large databases. BI and KDD focus on database transformation where BI is the broad category of applications and technologies for gathering, storing, analysing data to assist management, and KDD focuses on the development of methods and analytical techniques for making sense of data for strategic management decision making.

CHAPTER 3: RESEARCH METHODOLOGY

There is not only one data mining methodology to fit all analytical situations. For this study, the following goals should form part of a data mining methodology: Purpose driven, Data driven, Domain driven, Six Sigma driven and DOE driven. These goals will give focus and direction to the analytical process as well as strengthen the proposed framework.

CHAPTER 4: SIX SIGMA

Six Sigma is a methodology based on a structured process for continued improvement and refers to a data-driven improvement cycle used to design, measure, analyse, improve and control businesses processes and designs. For this research, Six Sigma is discussed as a proposed methodology to complement existing methods used for DM.

CHAPTER 5: STATISTICAL PROCESS CONTROL

Statistical process control (SPC) is a statistical technique that focuses on statistically controlling a process, which is also part of DMAIC (design, measure, analyse, improve and control) processes of Six Sigma. For this research, SPC will be discussed as a variable selection screening technique on input data to determine which independent and dependent variables are usable for DOE and regression analysis.

CHAPTER 6: DESIGN OF EXPERIMENTS

For this research, Design of Experiments (DOE) is discussed as an approach within the data transformation process that focuses on applying DOE to historical data to determine future process improvement. This approach serves as a screening process, using historic process data and not the typical approach by subjectively selecting design parameters based on experience and personal preferences or theories.

CHAPTER 7: COST METHODS

This chapter discusses the cost for producing nonconforming products when determining the best experimental run as each experimental run outcome deviates from the process outcome target. Process target is the process performance target set for process control. The cost implication for each experimental run is evaluated against the deviation from the process target. Evaluating the theoretical optimum DOE variable

combination with its nonconforming cost avoidance impact compared to product quality outcomes for each run may result in a more pragmatic decision.

CHAPTER 8: REGRESSION ANALYSIS

In this research, multiple regression (MR) analysis is applied to compare multiple regression analysis to designed experiment model regression. Multiple regression serves as a comparative benchmark for DOE because regression analysis complements designed experiments by also explaining the behaviour of the dependent variable through selected independent variables.

CHAPTER 9: NEURAL NETWORKS

The aim of this chapter is to compare the application of Neural networks (NN) to the statistical methods applied using the same database. NN is a black-box DM technique.

CHAPTER 10: SUMMARY OF STUDY

This chapter describes how the five primary goals for this study set out in Chapter 1 were achieved, with the appropriate links referencing to the associated chapters discussing how these goals were met through this research.

CHAPTER 11: REFLECTIONS ON STUDY AND FUTURE WORK

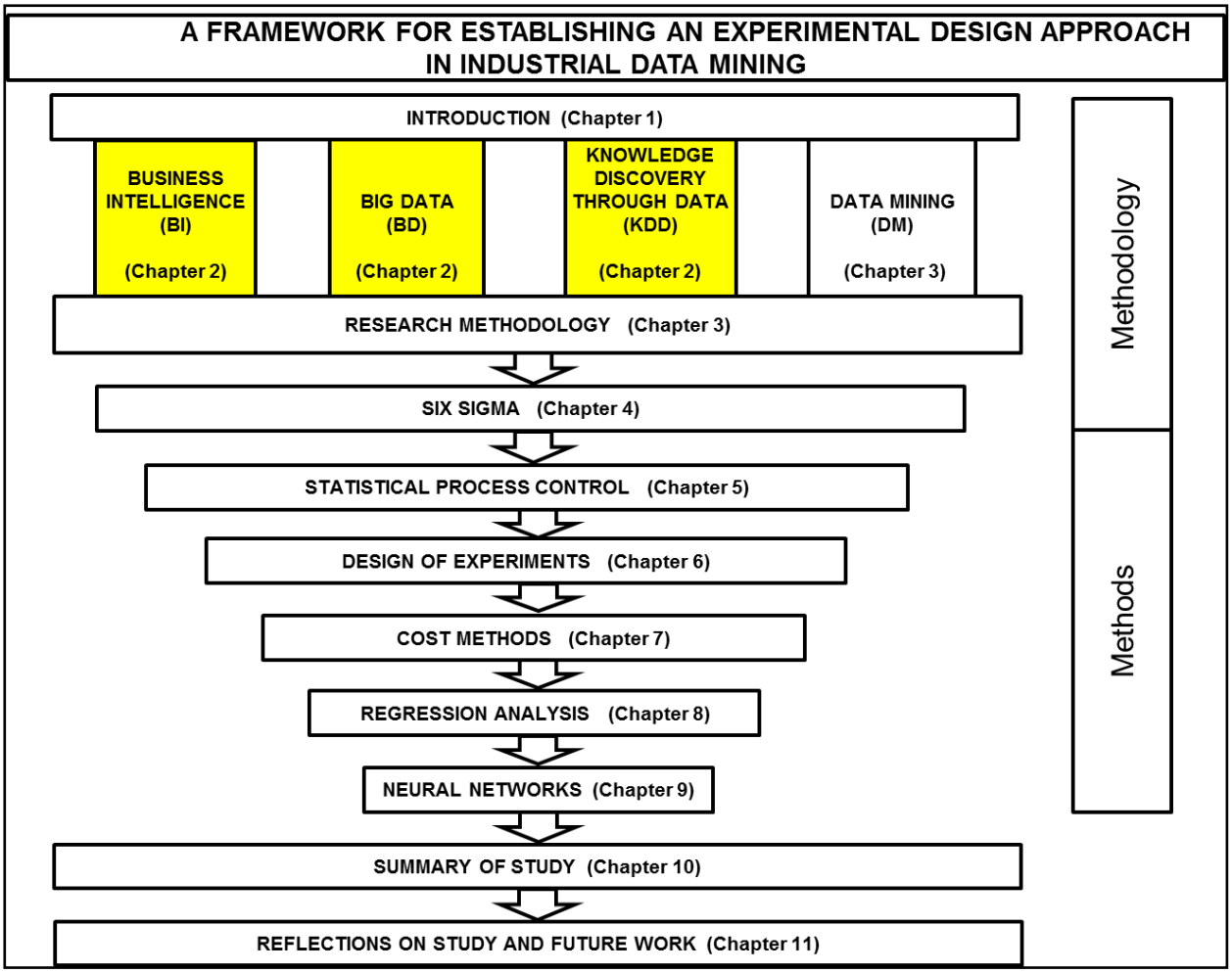
In this chapter, a summary of the research is presented as well as future potential projects that are aligned to this research, presented under future work.

1.8 SUMMARY

This chapter outlines how the title for this research, “A FRAMEWORK FOR ESTABLISHING AN EXPERIMENTAL DESIGN APPROACH IN INDUSTRIAL DATA MINING**”** was addressed by defining the purpose for this research, at which company the research was conducted and the goals defined to achieve for this study. From a broader angle, critical terms as well as the chapter layouts for this research are presented to give structure to the research. Chapter 2 covers the business context for selected frameworks for this research.

CHAPTER 2

BUSINESS CONTEXT



In this chapter, Business Intelligence (BI), Knowledge Discovery through Data (KDD) and Big Data (BD) are discussed in paragraphs 2.1, 2.2 and 2.3 respectively. These three business contexts are discussed with the goal of introducing different methodologies, from data sourcing to decision implementation. These methodologies gives a theoretical database perspective of managing data for analytical purposes. The importance of these methodologies for this framework lies with the importance of data preparation before the analytical part for the framework.

2.1 BUSINESS INTELLIGENCE

Kumar (2012:357) supports Lasi (2013:387) in confirming that data warehousing and Business Intelligence (BI) complement each other in a business environment for assisting management to make meaningful decisions. Because BI is data warehouse

driven and specifically focused on online analytical processing (OLAP), management can dynamically analyse, summarize and create online reporting to assist them in daily operational issues.

Jha (2014:1) emphasizes that BI provides data in a structured layered process by data extraction from operations into a data warehouse then to be analysed through parametric and non-parametric analytical tools by management for strategic decisions. BI does not only provide a platform for data analysis but also data for predictive analysis. This is important in the sense that strategic decisions are focused on the future events that have not yet happened rather than the day-to-day operational process dynamics (Mikroyannidis & Theodoulidis, 2010:559).

Hermida *et al.* (2013:411) add the World Wide Web, in short (WEB), as an additional angle to available data and the use thereof in a global environment. The WEB provides an unlimited source of data for management but needs well-structured data warehouses to accommodate and manage WEB information. This is an important dimension to BI because the business environment today is global, and to stay competitive their data support systems must be globally accessible and available for management. This dimension is especially true for multinational organizations competing on a global basis.

Unstructured data are not necessarily bad data, such data are simply not user friendly for direct access through analytical tools. BI will not fix this issue by itself, Alazmi (2012:8) explains the success of BI implementation to avoid 80% of unstructured data is to focus on data handling policies, warehouse design standards, architecture, systems, and to train staff to provide complete coverage of system needs that will assist in organising data for easy analytical processing. BI is the data vehicle for management and therefore must be well designed so that data analysts do not waste time to provide management with good information.

KOPČEKOVÁ *et al.* (2013:44) summarise the advantages of BI for businesses and companies as providing strategic direction to achieve strategic objectives, introducing decision-making flexibility to managers, segregating facts from fiction to shorten the strategic decision making process, identifying competitive advantages and culminating various data sources for data analysis to assist in the decision making process in order to provide meaningful inferences from outputs to management.

Tembhurkar *et al.* (2014:132) describe five important stages to transform data into BI successfully. These are the collection of raw data from business enterprise, data cleaning through search engines and filtering processes, data warehousing, implementation of BI tools, and analysing outputs. Transforming data into value added BI is not easy but must follow a systematic approach and well-designed scientific methods of analysing data for management strategic decision making.

Alazmi and Alazmi (2012:10) agree with Hermida *et al.* (2013:411) that the web (internet) has revolutionized the ability for accessing additional information from an **external source** to complement the internal generated data. Without a well-developed BI structure to absorb these challenges, businesses will lose their competitive edge.

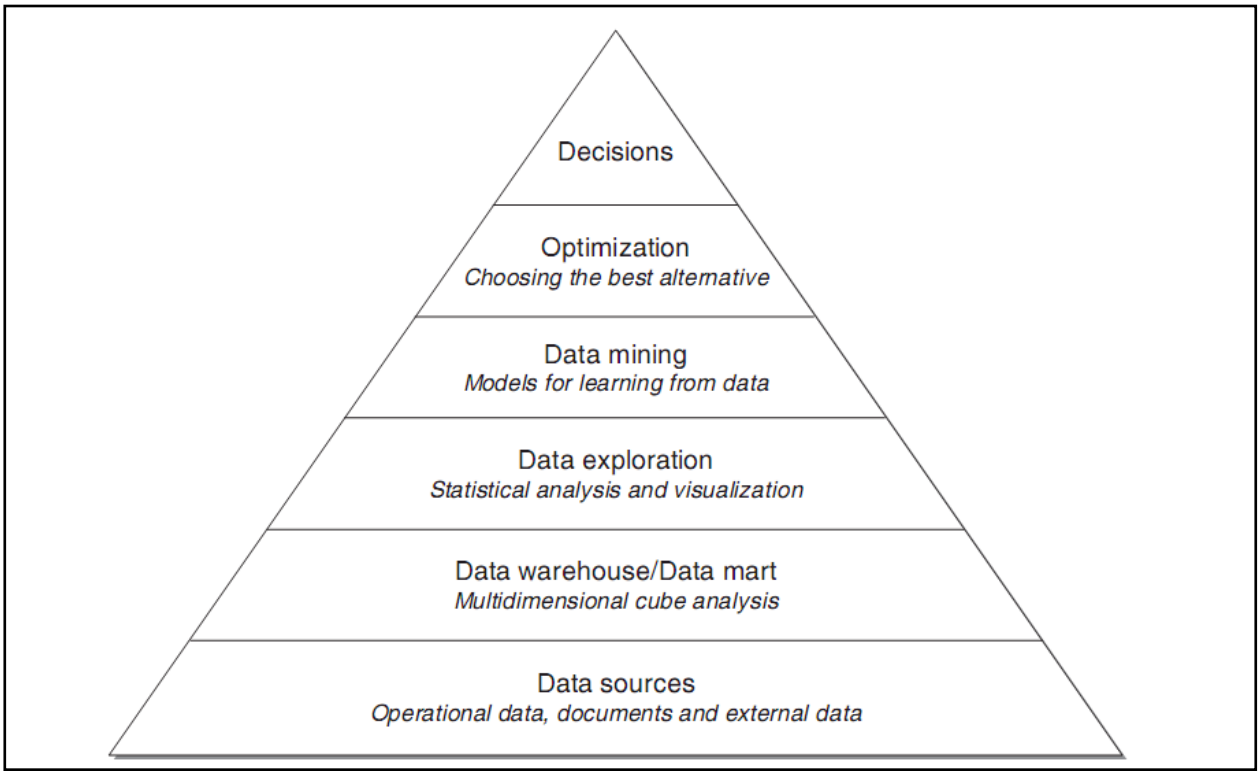


Figure 2.1: Main components of BI.

Figure 2.1 by Vercellis (2009) illustrates a typical business architecture, with the main components of BI and methodologies used for BI from data sources to decisions. This figure summarises BI for data usage, data analysis and possible methodologies to use for exploratory data analysis.

OLAP is the backbone for BI for system users to acces data in real time and online. This will only be possible if the BI architecture is designed to accommodate real time users

(Azhar *et al*. 2010:92; Bowman, 2009:13; Forsman, 1997:8; Turban & Aronson, 2001:147).

Figure 2.2 graphically shows the dependency between OLAP and BI to the benefit of management to make informative decisions. Although BI is the holistic component of the proposed framework for strategic management thinking and decision-making, BI and OLAP are an integrated system that allows users to carry out high-level data analyses with the information in data warehouses. The central element of business intelligence architecture represents data warehouses; business intelligence and data warehousing are mentioned on an interchangeable basis. Data warehouses are a critical component in providing useful data for BI in a holistic environment.

Traditionally BI was data warehouse driven, comprising three layers, namely queries and reports, online analytical processing, and data mining. Data were stored sequentially onto a master file in the form of magnetic tapes, and disk storage was then developed so that data could be stored and accessed directly more effectively. Modern BI developed from the traditional three-layered data warehouse to a data warehouse that consists of transactional and non-transactional data transformed for querying, reporting and data analysis when needed. These modern three layered data warehouses are summarised as follows:

First layer of analysis (Querying and reporting)

Online, dynamic querying using a computer to obtain real time answers to user questions. Reporting creates standard, real time reports through querying, describing specific report components and features.

Second level of analysis (On-line analytical processing)

On-line analytical processing (OLAP) allows analysts to conduct real time data analyses with the assistance of fast and interactive access to information in data warehouses. The dimensionality within the OLAP application usually reflects the different dimensions of an organisation.

Third level of analysis (Data mining)

The data mining process focuses on analysing and finding patterns in large amounts of data in order to support strategic decisions. This process not only involves applying

scientific techniques to data, but also starts with business and data understanding, data preparation, and then selecting the right modelling techniques for evaluation and implementation.

Figure 2.2 shows a typical BI architecture, which provides the initial building blocks for IT management. OLAP evolved because of the integrated nature of business, and the provider of data for statistical analysis to a critical architecture for management to have data available on or off-line, for decision-making.

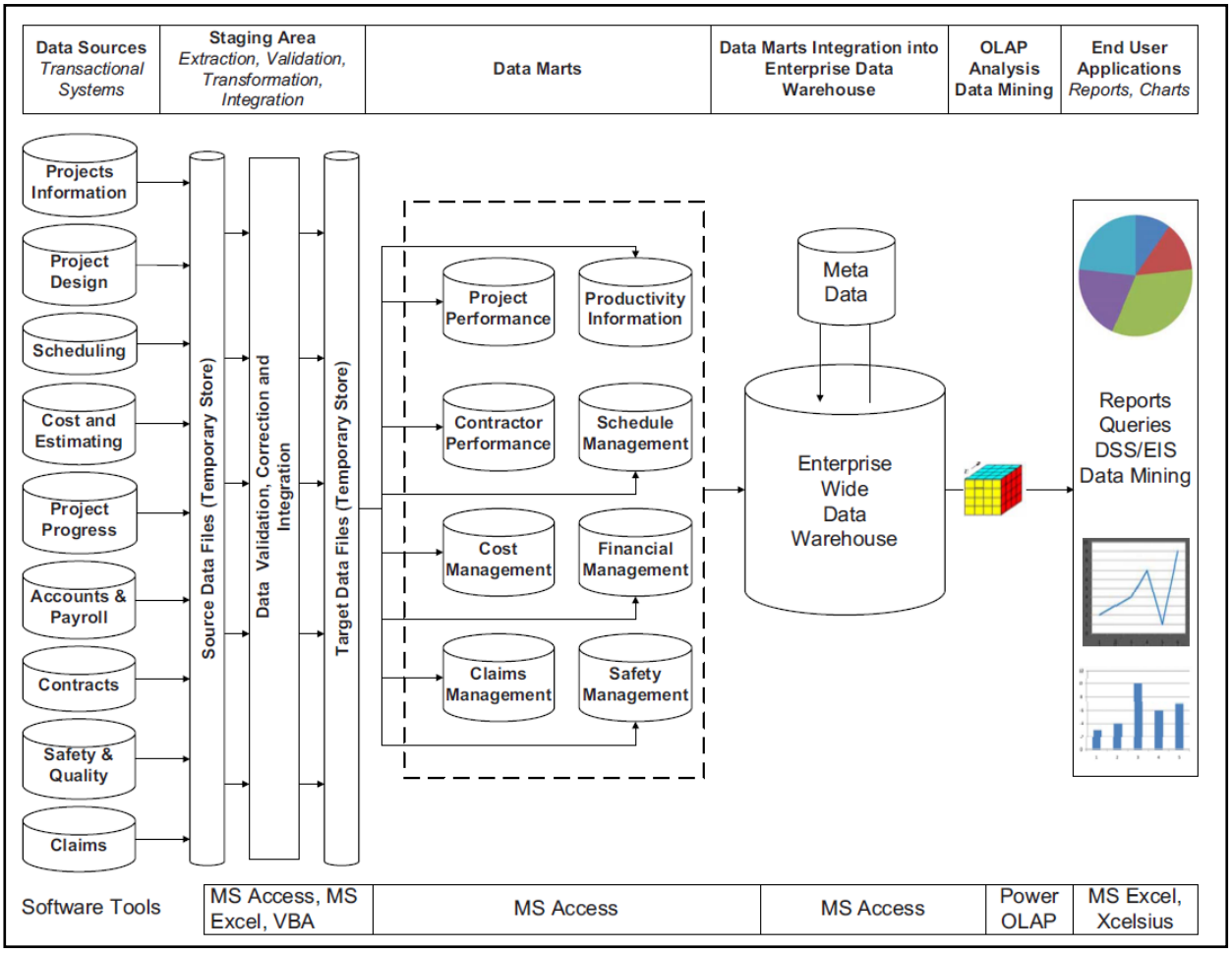


Figure 2.2: Typical business intelligence architecture.

In reference to Figure 2.2, Azhar *et al.* (2010:92) illustrate how OLAP operates within a data base architecture. It shows the quantitative analytical process through OLAP from data sources to end user application.

Figure 2.2 shows how important it is that all multi sourced data required for analytical data mining within an enterprise must be structured to ease the quantitative analytical

process. When comparing the main components of BI illustrated in Figure 2.1, the similarities are clearly visible.

According to Turban and Aronson (2001:147), no agreement exists on activities considered as OLAP. Usually OLAP activities are generating queries, requesting ad hoc reports, conducting statistical analyses, and building Decision Support Systems (DDS) and multimedia applications. To facilitate OLAP, it is useful to work with the data warehouse and with a set of OLAP tools, which can be query tools, excel spreadsheets, data mining tools and data visualisation tools.

The general business model is a multi-dimensional model to provide managers and analysts with data from all sectors within an organization, from production to services. For this reason, data for all these sectors are contained in the OLAP databases. Managers must be able to analyse data across all business dimensions, at any level of aggregation, with equal functionality and ease. It refers to the ability to perform complex analytical calculations, in order to create information from very large and complex amounts of data which also include the time dimension. Madhuri (2013:330) focuses on the importance of OLAP in terms of data cleaning, the integration of data that are a part of data warehousing technology. For BI from a strategic management perspective, the focus is clearly on data warehouse design, online data accessibility for managers and data users for data analytics.

BI is a broad category of applications and technologies for gathering, storing, analysing, and providing access to data to assist management in all levels of an organization to improve their business decisions. Although BI is the holistic component of the proposed framework for strategic management thinking and decision-making, BI and OLAP are an integrated system that allows users to carry out high-level data analyses with the information in data warehouses.

Because the central element of business intelligence architecture represents data warehouses, business intelligence and data warehousing are mentioned on an interchangeable basis.

2.2 BIG DATA

Big data (BD) is a consequence of data explosion, experienced by all industries. The advancement in server design and software development has made managing BD

effectively easier. Sowmya *et al.* (2015:121) describe big data as a data mining evolution that became big data as accumulated data grow exponentially. Because of the large accumulated data on a daily basis in industry, the traditional DM process had to adopt the growing data explosion. Accumulated data stream, external to a company that is not managed effectively in terms of poor control and data integrity through a structured data accumulation process and storage architectures, will enhance the data explosion within a company.

The BD evolution has grown exponentially by profound IT developments through the years. Four main IT areas which have assisted in BD explosion are the reduction in data storage cost (which is negligible compared to processing total cost); increasing of computing processing speed; development of new machine-learning algorithms structuring data for managerial analytical purposes from unstructured data, and developing software that allows the end user to manage large portions of data in a structured way. Figure 2.3 by Perrons and Jensen (2015:218) illustrates that the reduction in processing cost of collecting and storing data, and the focus on maintaining data quality due to ensuring correct sampling methods have changed drastically towards analysing complete data bases with minimum sampling because of the high data processing speed and software development in data analysis.

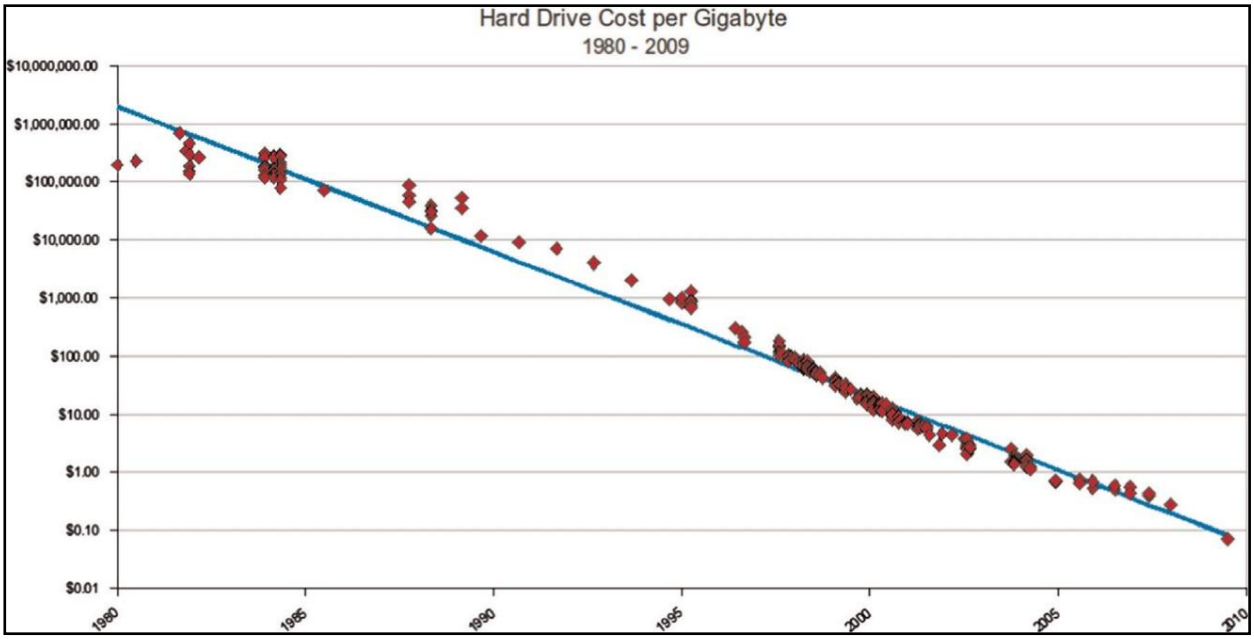


Figure 2.3: Cost of hard drive cost per Gigabyte

Graph 2.5 illustrates the advancement of hardware design accommodating the data explosion. It becomes progressively cheaper to store data for analysis purposes as time

progresses. For this reason, analysts are moving to “all data” analysis instead of sampling. Server cost portion of total cost is decreasing which makes data analysis even cheaper.

Critical drivers that emerged for data explosion, also named the 5Vs are: **Variety**: a plethora of different types of data exist that contain text, audio, web-data, video, graphs, quantitative, qualitative, process data, marketing data, financial data, and many more. **Volume**: data will keep on increasing exponentially and therefore continue to challenge the analyst. Managing the current data volume, Zhong *et al*. (2015:262) also acknowledge that the volume of data as a collection of data bases became so complex that traditional analytical techniques used by the analyst are not sufficient anymore, and therefore this has become an increased challenge for modern analysts. **Velocity**: the rate of data accumulation increases daily. As technology improves, the rate of incoming data streaming also increases. Storage and volume of data in modern business are a function of the daily data arrival rate. **Value**: organizations have a competitive advantage by making dynamic strategic decisions due to real-time data analysis. **Variability**: changes in database design, data structures, management reporting, management analytical requirements, data type changes and interpretation of data by analysts and management (Swan, 2015:469; Tyagi *et al*. 2015:16).

Datamation is another way of describing big data in terms of the value chain of an organization. Each process within the value chain contributes its data independently for management to optimize all related processes. For datamation BD is the culmination of all process data that constitute the value chain. Zhou *et al. (*2014:1629) also referto datamation as the optimization of the industrial value chain and in return improve the data operation for the complete related industrial chain.

A proposed framework for companies to use as a roadmap for implementing big data projects, characterized by the 5Vs that were discussed earlier, is presented by Diagram 2.1. It presents a detailed BD framework organized into three distinct phases: strategic groundwork, data analytics and implementation (Dutta and Bose, 2015:294). The 5Vs were put into context within a project implementation environment that shows a way of how big data projects should be managed. Big data are not only large databases for industrial processing industries to assist for analytical purposes, but also a collection of various data sources for project management to benefit from.

Diagram 2.1 shows that the benefits and challenges of BD implementation are not specific to a production or manufacturing environment. Even for service functions, these issues are similar because big data are not exclusive to manufacturing, but everywhere where large volumes of data are accumulated for strategic business decisions.

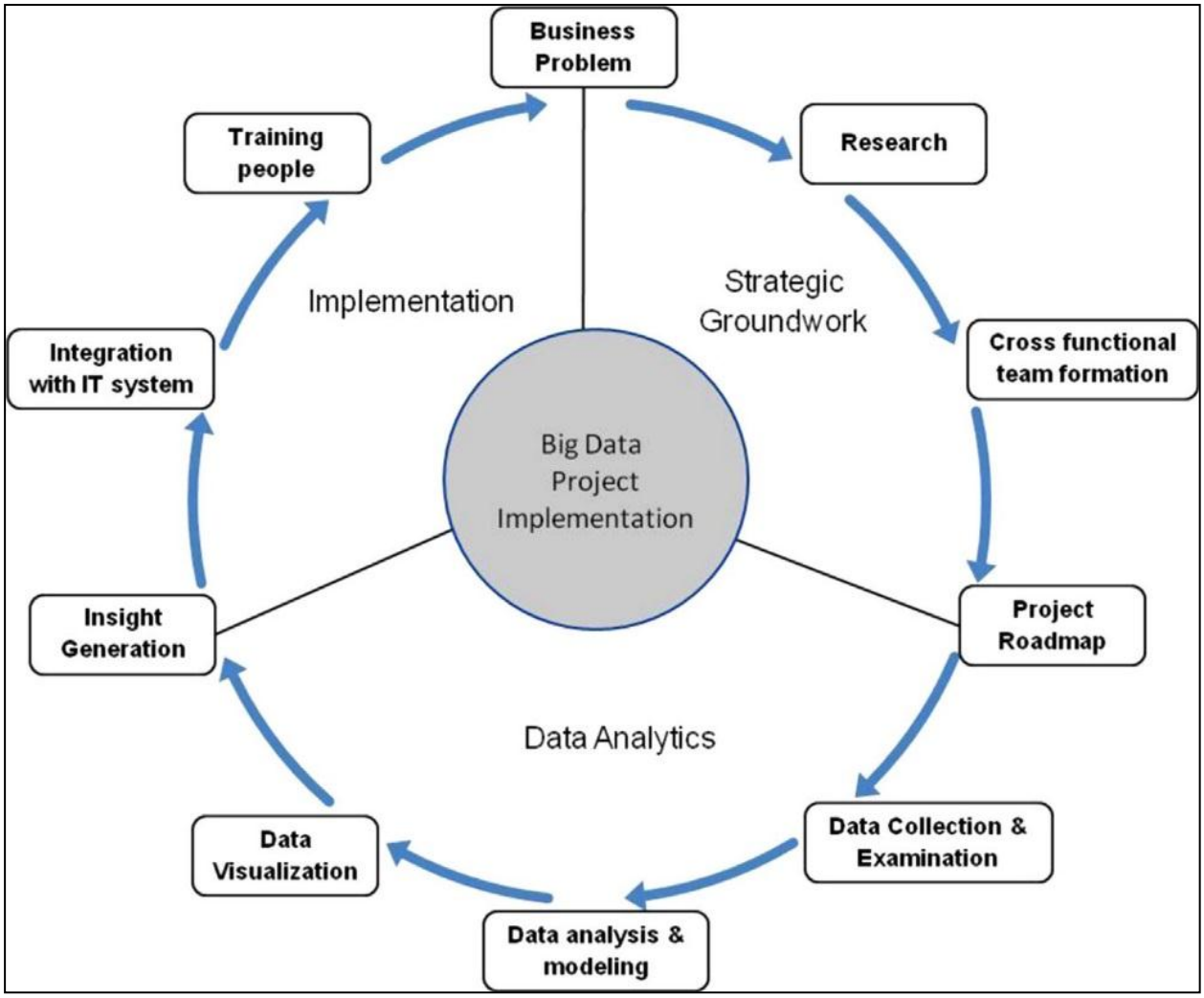


Diagram 2.1: Framework for implementation of Big Data projects in a firm

According to Nedelcu (2013:18), a survey done to identify benefits and challenges showed that the **top three biggest benefits** of Big Data are to detect product defects to boost quality, improve supply planning and improved defect detection in manufacturing/production environment. The **top three challenges** are the difficulty of building trust between data scientists analysing these databases and functional managers trusting information presented by analysts, determining which data are applicable for different business decisions, and the ability to handle large volumes of data velocity and variety of data.

The data explosion in the service environment is the same as in the manufacturing environment, the 5Vs are also as prominent as the manufacturing environment. Opresnik and Taisch (2015:176) express BD in the service environment; see Diagram 2.2, as servitization. The servitization concept refers to knowledge gained through services that complement manufacturing with a common goal to increase competitive advantage. Analysing service and industrial data independently will enhance the risk that the results will be biased and not holistic. Many companies however analyse only data specific to a problem and ignore the holistic approach because they believe that it is time-consuming and costly.

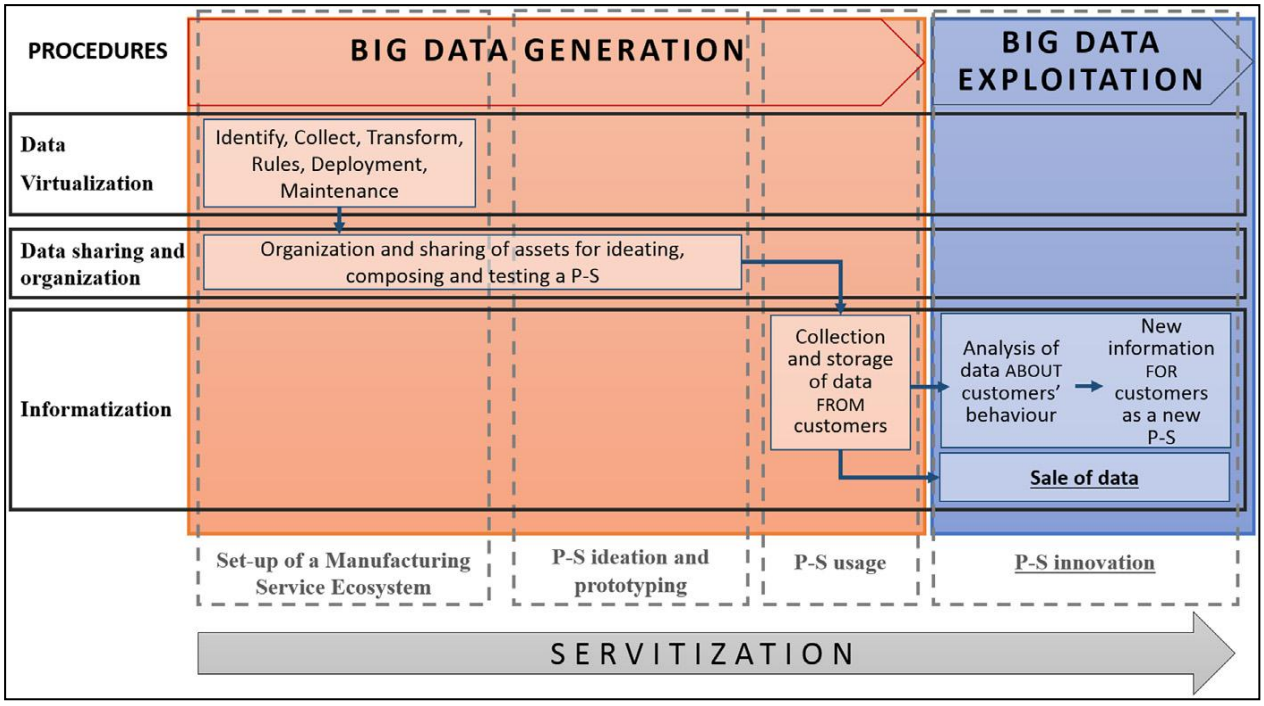


Diagram 2.2: Big Data Strategy within the context of servitization

Each compartment of Diagram 2.2 is not explained because it is referenced to show a different angle towards data generation through the transformation of products and processes through a manufacturing process from BD generation to BD exploitation in the service sector. This diagram also shows the importance of analysing (exploitation) data and not only generating large volumes of data.

Zhong *et al.* (2015:262) propose a framework that consists of six steps to manage and analyse big data from raw materials to finished goods. **Firstly**, get logistic data from all applicable databases for the manufacturing process that resides in several data warehouses relating to BD. **Secondly**, create a data warehouse for analytical purposes

by cleaning redundant, missing, irrelevant and non-added value data. **Thirdly**, compress data to remove all data identified by the data cleaning process. The new data warehouse should be smaller with reliable and applicable data. **Fourthly**, group applicable data sets within the compressed data warehouse together to fit the management needs for analysing data to enhance strategic decision-making. **Fifthly**, recognise patterns, trends and associations while executing data analysis. **Finally**, use these patterns and develop statistical and or mathematical predictive models for management to utilise for strategic decisions to maintain the competitive advantage.

The inclusion of statistical techniques to uncover the real value of data in large databases (BD) generally has been underutilised by management in the sense that management predominantly focuses on quick fixes. The framework described above should assist in structurally analysing BD irrespective of the statistical technique used.

According to Zhou *et al. (*2014:1631), China has entered the era of big data because of large amounts of structured and semi-structured data in storage generated by enterprises. This resulted in economic value to collect, analyse and to mine these massive amounts of data for a competitive advantage.

2.3 NOWLEDGE DISCOVERY THROUGH DATA

Feldkamp *et al.* (2015:50) refer to the primary goal of Knowledge discovery through data (KDD) to be the transformation of data into usable summarized forms for management and data users. In a general sense, this should be the goal of all analysts: to provide data in a summarized, condensed, factual format to allow management to focus on the core issues only, rather than be fragmented amongst various data bases to make sense of available data. Ayobami and Rabi (2012:231) focus on the development of methods and analytical techniques for making sense of data for strategic management decision making. Their view on KDD emphasizes the importance of selecting the correct analytical techniques for discovering knowledge. Focusing on technique driven analytics enhances the goal of KDD, to extract high-level knowledge from low-level data of large data sets.

KDD and DM are referred to as the same process. Although they refer to KDD and DM as synonymous, DM is not the same as KDD; however, the process of accessing data in both are very similar. Both have a structured approach for data analysis, but DM is based on raw unstructured high volume data, where KDD is a process where the

refined and structured data are used for online data analysis by management (Ayobami & Rabi, 2012:234; Lamont, 2012:8).

The KDD process focuses on different approaches for data extraction by KDD which is software driven and based on statistical analysis that includes probabilistic, statistical, classification, data cleaning and decision trees, but also data mining applications like neural networks and machine learning. Knowledge discovery cannot happen on its own and therefore is based on a structured data driven technique for analytical purposes. (Dunham, 2003:23; Fayyad *et al*. 1996:37; Purohit *et al.* 2012:457).

Brešić (2012:32) describes KDD as the discovery of new information and knowledge. New information is not necessarily information that has been recently added to a process or business, but can be latent information, never exposed until it has been discovered. Most of the time “new” knowledge is “old” knowledge that has been recently discovered. Modern KDD evolved into a multidisciplinary activity that utilises techniques such as machine learning, pattern recognition, statistics, data visualisation and high-performance computing with special emphasis on uncovering patterns, identifying outliers through exploratory analysis and structured experimenting (DOE) on databases.

Since the late nineties to the most recent definitions, the only difference has been the use of advanced software though OLAP to assist in analysing data for management. KDD is described as a methodology for data analysis, discovering patterns, relations between variables, and comparing KDD to DM (Fayyad, *et al.* 1996:; Mackinnon & Glick, 1999:256; Wright, 1998:94).

Maimon and Rokach (2010:2) describe KDD as a nine-step iterative and interactive process that integrates data mining as part of the knowledge discovery process (see Figure 2.4). These nine steps are summarized as: developing an understanding of the application domain, selecting and creating a data set on which discovery will be performed, pre-processing and cleansing of data, data transformation, choosing the appropriate DM task, choosing the DM algorithm, employing the DM mining algorithm, evaluation and using the discovered knowledge.

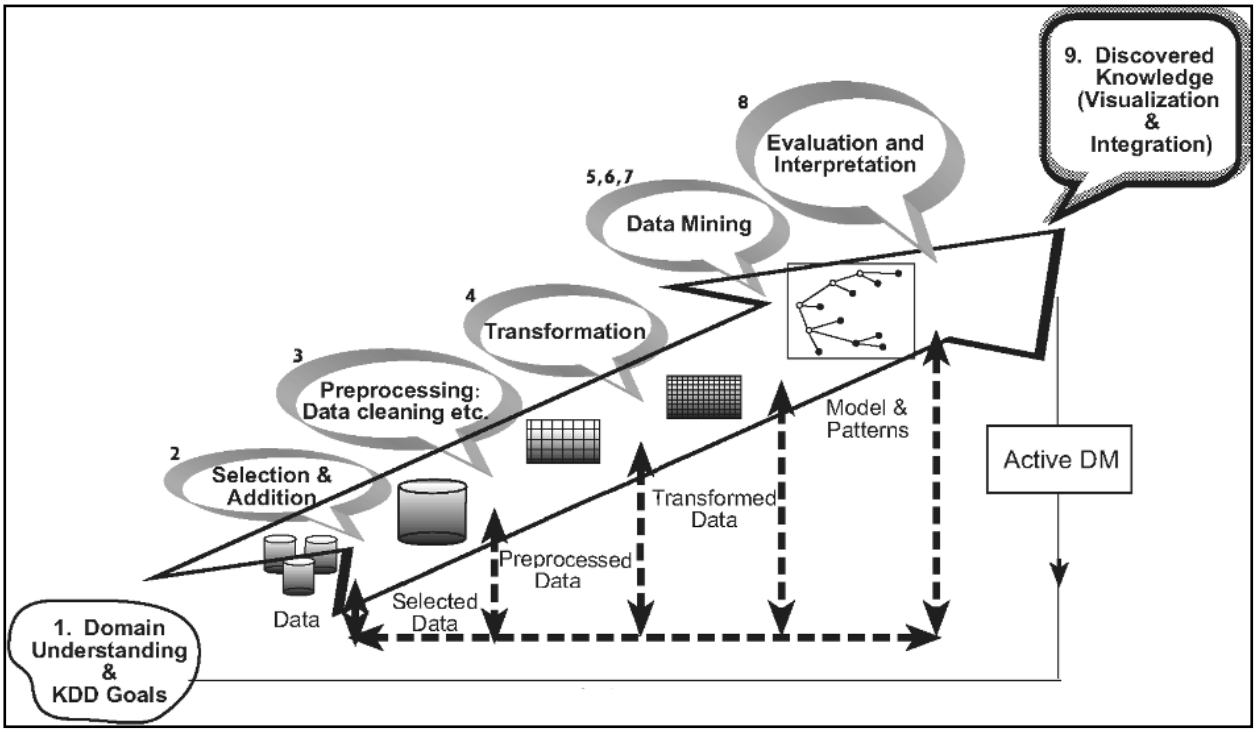


Figure 2.4: The Process of Knowledge Discovery in Databases

Erohin *et al.* (2012:428) describe the Modelling step in their KDD process to start with descriptive data mining to detect characteristic clusters in process and product data. By evaluation of the existing linkages between product and process instances, the correlations between clusters of process and product data are identified. These are the basis for the application of predicting data mining models in order to provide a structure mapping between product and process clusters.

Feldkamp *et al.* (2015:5) show in Figure 2.5 how KDD is gained through a simulated production type process from experimental design to knowledge discovery as well as how data mining forms an integral part of the KDD process.

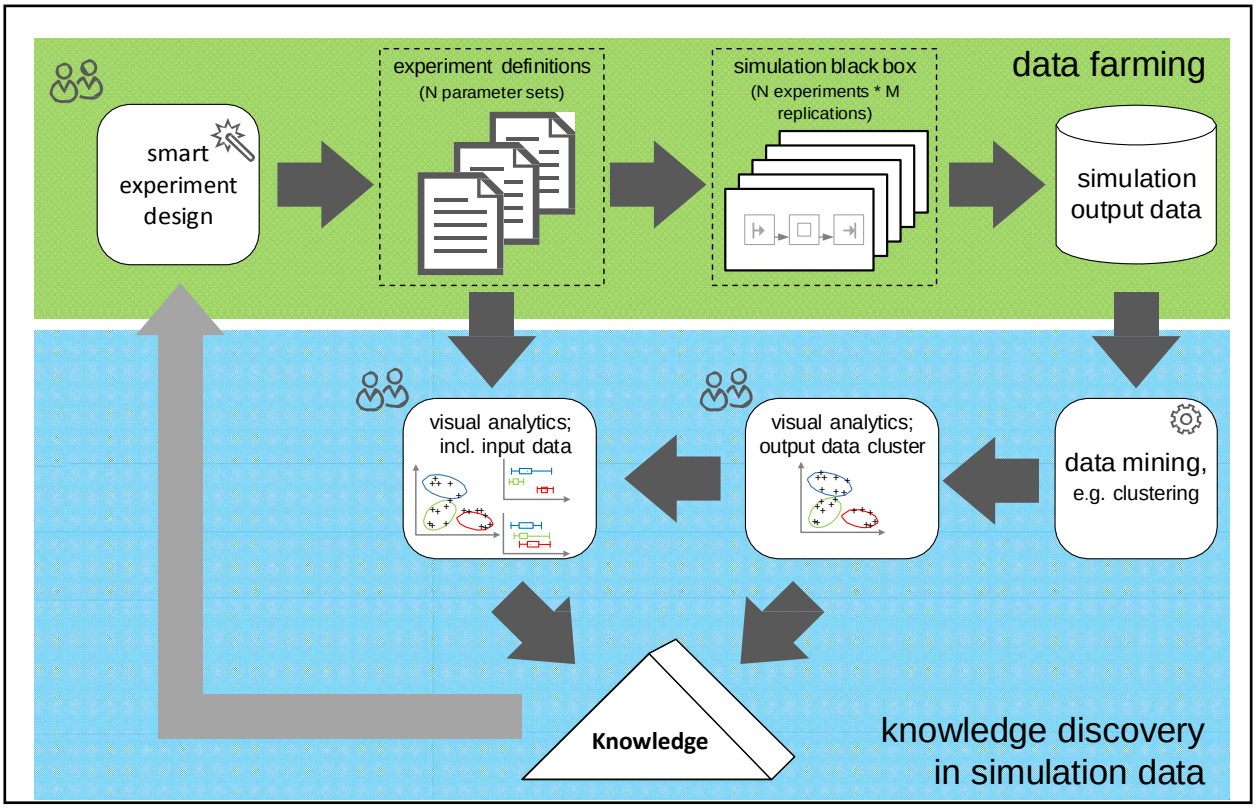


Figure 2.5: Knowledge discovery process in a discrete event manufacturing simulation

Figure 2.5 shows the importance of DOE in the KDD process cycle. Although this KDD cycle represents simulated data, DOE features as a prominent analytical tool in this environment for reaching desired results. Data warehousing helps set the stage for KDD in two ways, namely data cleaning and data access (Fayyad *et al.* 1996:40).

A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support.

Three authors provide process steps for KDD, recognising KDD as a procedural approach in analysing data. These are:

Wong and Chung (2007:364):

1. Select the application domain and determine the availability of information, selecting target data by identifying and defining data types.
2. Pre-process data by cleaning data for integrity to ensure data validity.

3. Extract knowledge by applying appropriate data mining techniques to uncover patterns and relationships not previously known.
4. Interpret and evaluate through removing redundant or irrelative patterns and relationships.
5. Translate data that fit the model into meaningful information for the end use.

Purohit *et al*. (2012:457):

1. Select data relevant to the analysis task from the database.
2. Remove noise and inconsistent data.
3. Combine multiple data sources.
4. Transform data into appropriate forms to perform data mining.
5. Choose an appropriate data-mining algorithm to extract data patterns.
6. Interpret the patterns into knowledge by removing redundant or irrelevant patterns.
7. Translate the remaining patterns into terms for human understanding.

Fayyad *et al*. (1996:42):

1. Develop an understanding of the application domain and the relevant prior knowledge, and identify the goal of the KDD process from the customer's viewpoint.
2. Create a target data set. Select a data set, or focus on subsets of variables or data samples, on which discovery is to be performed.
3. Clean and pre-process data. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
4. Reduce and project data: finding useful features to represent the data depending on the goal of the task − with dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representation for the data can be found.
5. Match the goal of the KDD process (Step 1) to a particular data mining method.

6. Choose the data patterns by exploratory analysis and model hypothesis selection.
7. Data mining: searching for patterns of interest in a particular representation, including classification rules or trees, regression, and clustering.
8. Interpret mined patterns, possibly returning to any of steps 1 through 7 for future iteration. This step can involve visualisation of the extracted patterns and models.
9. Act on the discovered knowledge. Use the knowledge directly, incorporating the knowledge into another system for further action, or simply document it and report it to interested parties.

KDD is proposed as a stepwise process, and emphasizes the importance of a structured approach for data analysis, through either a methodology, framework or steps. It guides the analyst and management when analysing data. KDD focuses on the development of methods and analytical techniques for making sense of data for strategic management decision making. This enhances the goal of KDD to extract high-level knowledge from low-level data of large data sets. This process emphasizes the importance of a structured approach in analysing data; a haphazard approach will only cause results that will not be positive or sustainable (Maimon & Rokach, 2010:3; Purohit *et al*. 2012:457; Wong & Chung, 2007:364).

KDD is a process that is interactive and iterative by nature and usually follows a defined structure for analysing data. A typical, generic, KDD process flows as follows:

- **Set goals and objectives before the process starts.** Understand customer needs before evaluating raw data and sources of data.
- **Quality of data is imperative for fact-based decisions**. Decide which data will be used and which will be discarded. This decision will be led by the detail and/or quality of data needed for achieving the goal or objectives.
- **Clean data.** This ensures that outliers do not clutter results. Experience and job knowledge are essential for success. The ideal is to have a database that contains clean data; this is a modern trend that follows from a well-designed architecture.
- **Design a structure or model prior to analysing data.** It gives the analyst a roadmap and an unstructured methodology to follow. By doing this, the model is prevented from growing to infinity.

- **Start the data analysis.** This is the core of the KDD process. It uses different techniques, and also fits the pre-designed model. Evaluate all outputs using an iterative process to refine the model.

KDD focuses on the development of methods and analytical techniques for making sense of data for strategic management decision making. This enhances the goal of KDD to extract high-level knowledge from low-level data of large data sets.

Different approaches for KDD include probabilistic, statistical, classification, data-cleaning and decision trees. These approaches typically are for data filtering and structuring to be accessed by data mining approaches.

2.4 SUMMARY

Business intelligence is the holistic data management of an enterprise for strategic decisions based on transforming data into valuable and actionable knowledge. This approach is pertinent particularly to all leadership positions, which make strategic decisions on a daily basis. Business intelligence tools assist management to make decisions on facts instead of instinct. Business intelligence is more than managing and storing of data, its design and applications touch a wide spread of managerial functions.

Unless businesses have an integrated system to analysee and control the ever-increasing volume of data, accumulated data will only be data and nothing else. Companies that utilise well-designed BI structures and BI tools in managing accumulated data in a structured way to provide management with **meaningful information to make informed decisions** will experience a high competitive advantage amongst their peers in similar industries.

Different viewpoints are an important characteristic of OLAP, also called multidimensionality. Multidimensional means viewing data in three or more dimensions through the application of different statistical methods ie; SPC, MR, DOE. Analysing data in multiple dimensions is particularly helpful in discovering relationships that cannot be deduced directly from the data itself. OLAP products and applications have been around for a long time, not in the current form, but have been recognised as a progressive booming approach to data storage and the providing of a platform for analytical processes. For this reason OLAP is currently becoming the biggest-ever growth of multidimensional applications.

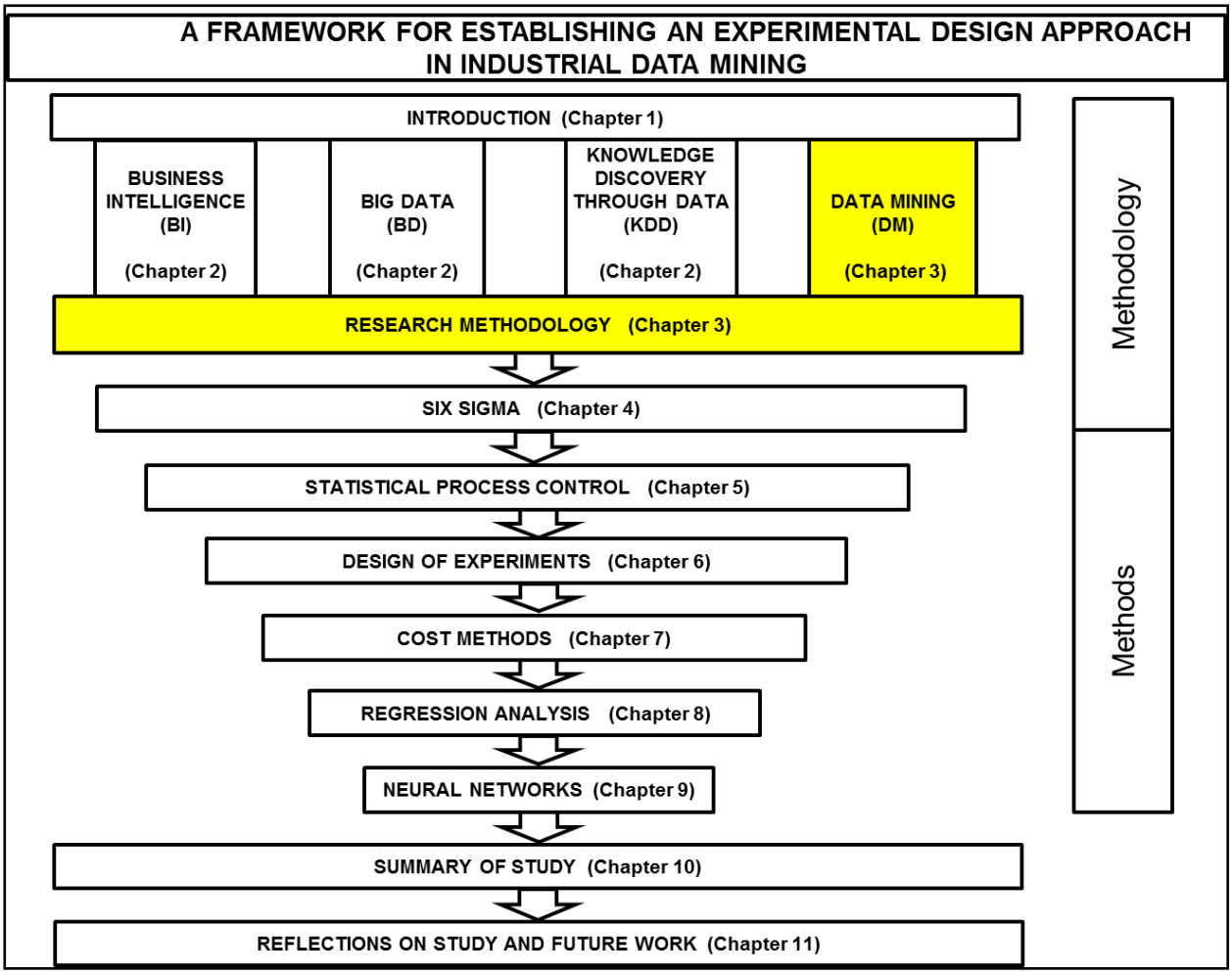
KDD evolved into a multidisciplinary activity that utilises techniques such as machine learning, pattern recognition, statistics, data visualisation and high-performance computing with special emphasis on uncovering patterns between customers and products, identifying outliers through exploratory analysis, and structured experimenting on (DOE) databases.

From a BD perspective, data gathered in databases for different applications in business, are transformed from these databases into information, then summarised into reports for decision-making by management. The quantity of data gathered daily in databases is too vast to manage effectively without the support of modern computer technology, computer-assisted software programmes and analysing techniques. Therefore, in managing these large data sets effectively, they are transformed into information and reduced into manageable quantities for analytical purposes for management.

The inclusion of statistical techniques to uncover the real value of data in large databases (BD) has generally been underutilised by management in the sense that management predominantly focuses on using statistics as a quick fix. Managing and statistically analysing large databases is not new, but traditionally these were not easily accessible. This obstacle is no more an issue with modern computer software/hardware available for analytical purposes. When analysing data, the only limitations are the creativity and innovativeness of the data analyst, and his/her and knowledge of statistical tools

CHAPTER 3

RESEARCH METHODOLOGY



The research philosophy focuses on the research approach and the research methodology focuses on how Data Mining (DM) is used in the study as well as the company with its related database for statistical analysis. Although DM is separated as a methodology inclusive with the research methodology, subsequent section 3.2 will discuss the integration of DM with DMIAC, BI, DOE, KDD and BD.

3.1 INTRODUCTION

The goal was not to exhaust all possible theories and philosophies but only those few that relate to the research framework for this study set out in the research approach.

Gray (2012:19) describes epistemology as a branch of a philosophy that assists researchers in determining the limits of human knowledge as well as in guiding the

philosophical background for deciding what kinds of knowledge are legitimate and adequate. Refer to Diagram 3.1.

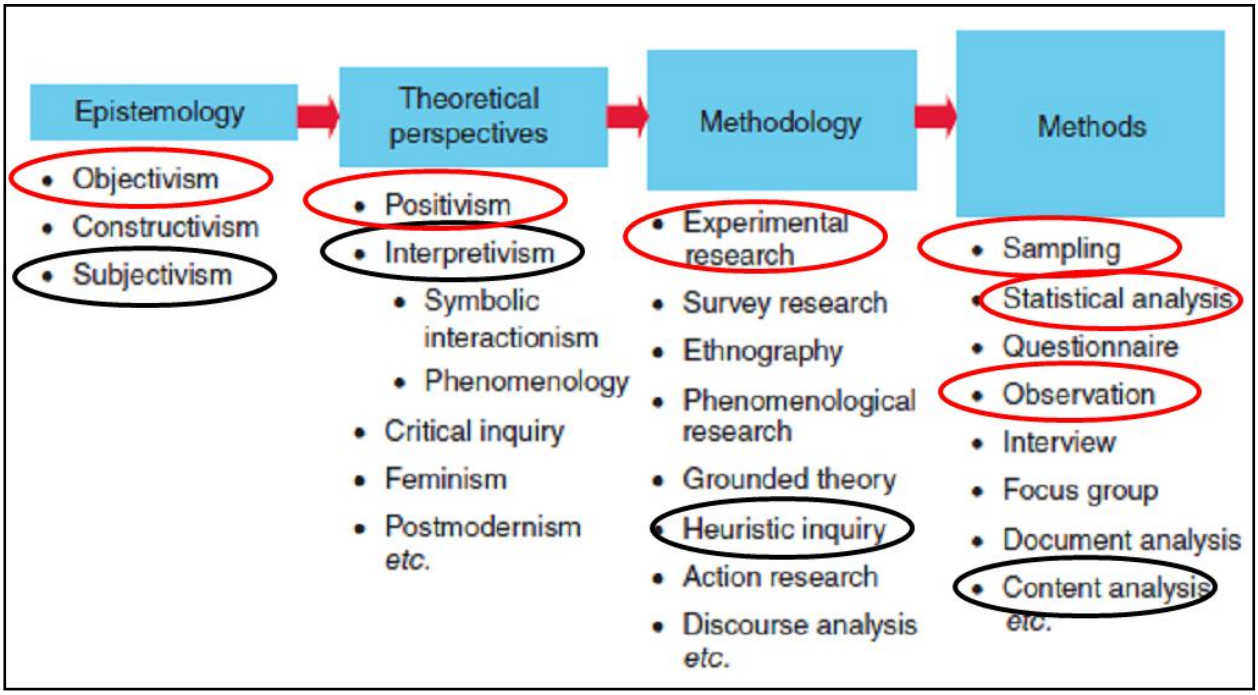


Diagram 3.1: Relationship between epistemology, theoretical perspectives, methodologies and research methods

Epistemology **firstly** assists in clarifying issues of research design in terms of choosing research tools as well as the type of evidence being gathered, from where, and how it is going to be interpreted. **Secondly**, knowledge of research philosophy will assist the researcher to recognize designs that will work for a given set of objectives, and which designs will not work.

Summarizing epistemology in Diagram 3.1 shows the relationship and separation between **literature study and empirical work for this study**. It emphasizes the main goal of epistemology, what knowledge is and how to structurally gain it for a research project. Not all sub items are included, but only the main items for this study are highlighted:

Epistemology: Objectivism (Empirical study), Subjectivism (Literature study)

Theoretical perspectives: Positivism (Empirical study), Interpretivism (Literature study)

Methodologies: DOE research (Empirical study), Heuristic inquiry (Literature study)

Methods: Sampling (Empirical study), Statistical Analysis (Empirical study), Content Analysis (Literature study), Observation (Empirical study)

From this summary it is clear that the empirical side links to positivism and the literature study side links to interpretivism.

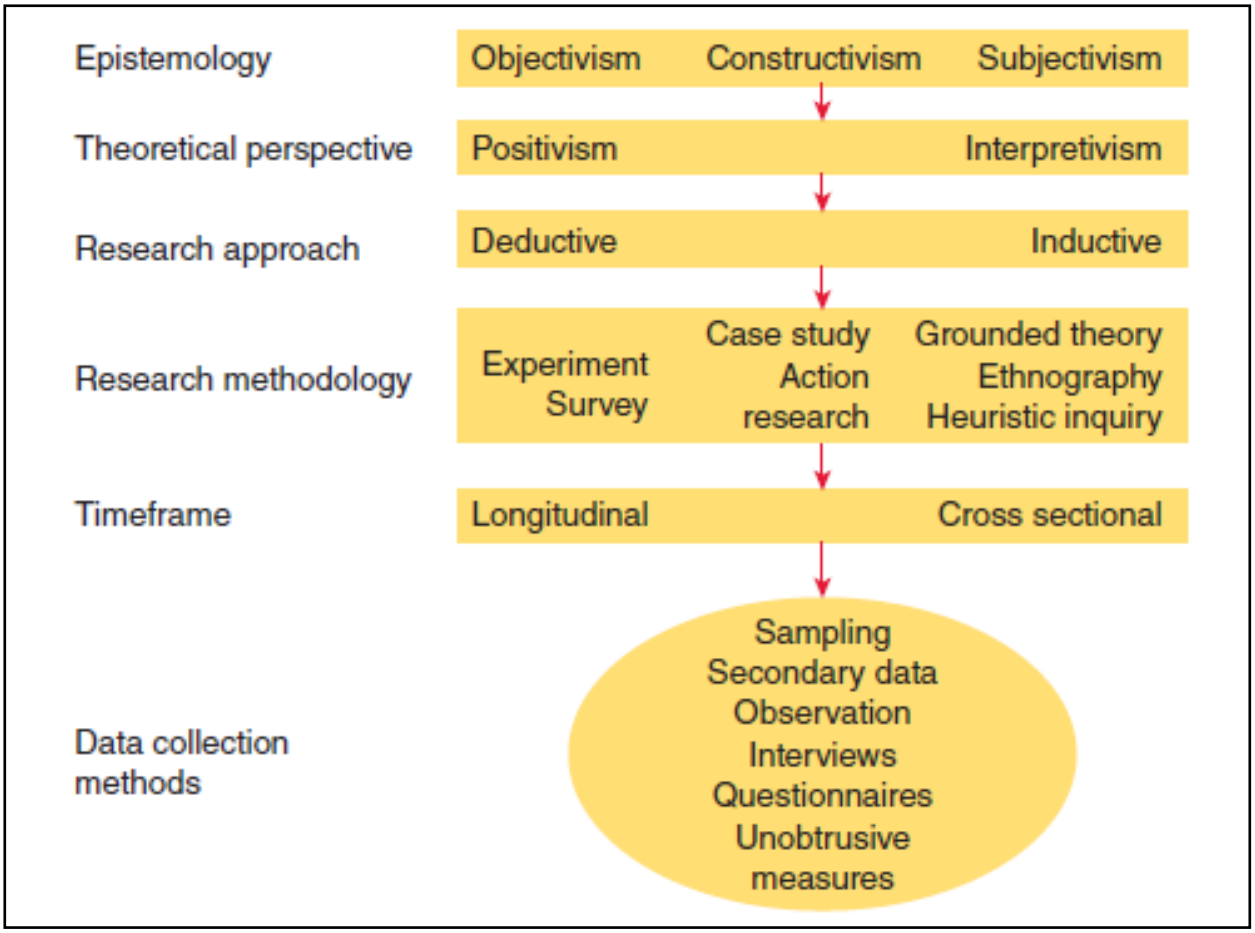


Diagram 3.2: The elements of the research process

Diagram 3.2 by Gray (2012:25) shows elements of a research process which assist in choosing a research methodology applicable for this research. Considering this research, both positivism and interpretivism relate to the goals of this study. Following Diagram 3.2 as a decision base for the research methodology referencing to the experimental design approach as a primary goal for this study, constructivism also seems to fit the research that links to the case study as a DMAIC project that is discussed in later chapters. See a summary below:

Epistemology: Constructivism

Theoretical perspective: Positivism and interpretivism

Research approach: Deductive and inductive

Research methodology: A case study: Secondary analytics for two time periods on observed data.

Timeframe: Longitudinal (weekly data in each period) and cross sectional (between periods).

Data collection methods: Sampling, secondary data or complete data sets of observed data.

Positivism has common elements across some authors that describe this research methodology (Creswell, 2008:9; Goes, 2013:5; Gray, 2012:20; Mackenzie & Knipe, 2006:14; McKinney, 2011:12; Parminter *et al.* 2003).

A summarized view and communality elements for positivism are

- Knowledge is factual, which is gained from observations designed by the study.
- Results are quantitative, research based, quantifiable and statistically analysed.
- Focused on natural and physical science and may predict patterns of behaviour.
- A quantitative methodology that involves experimental and control groups, as well as manages the logistics of measuring before and after results.
- Science research based on rationalistic and empiricist philosophy.
- Problem focused that provides pragmatic solutions for real problems.
- Combining qualitative and quantitative methods, experimental approaches, empirical models, and big data analytics to solve and provide analytical and empirical modelling options for management to assist in strategic decision-making.

Figure 3.1 depicts the worldview of research, how we approach our lives, what particular view we use, how we reflect our assumptions about the nature of the world, and how these might be investigated.

Because the quantitative basis of this research that flows from identifying different methodologies to a proposal for an experimental design approach for industrial DM not only involves one quadrant, see Figure 3.1 constructed by Parminter *et al.* (2003), adapted for this study.

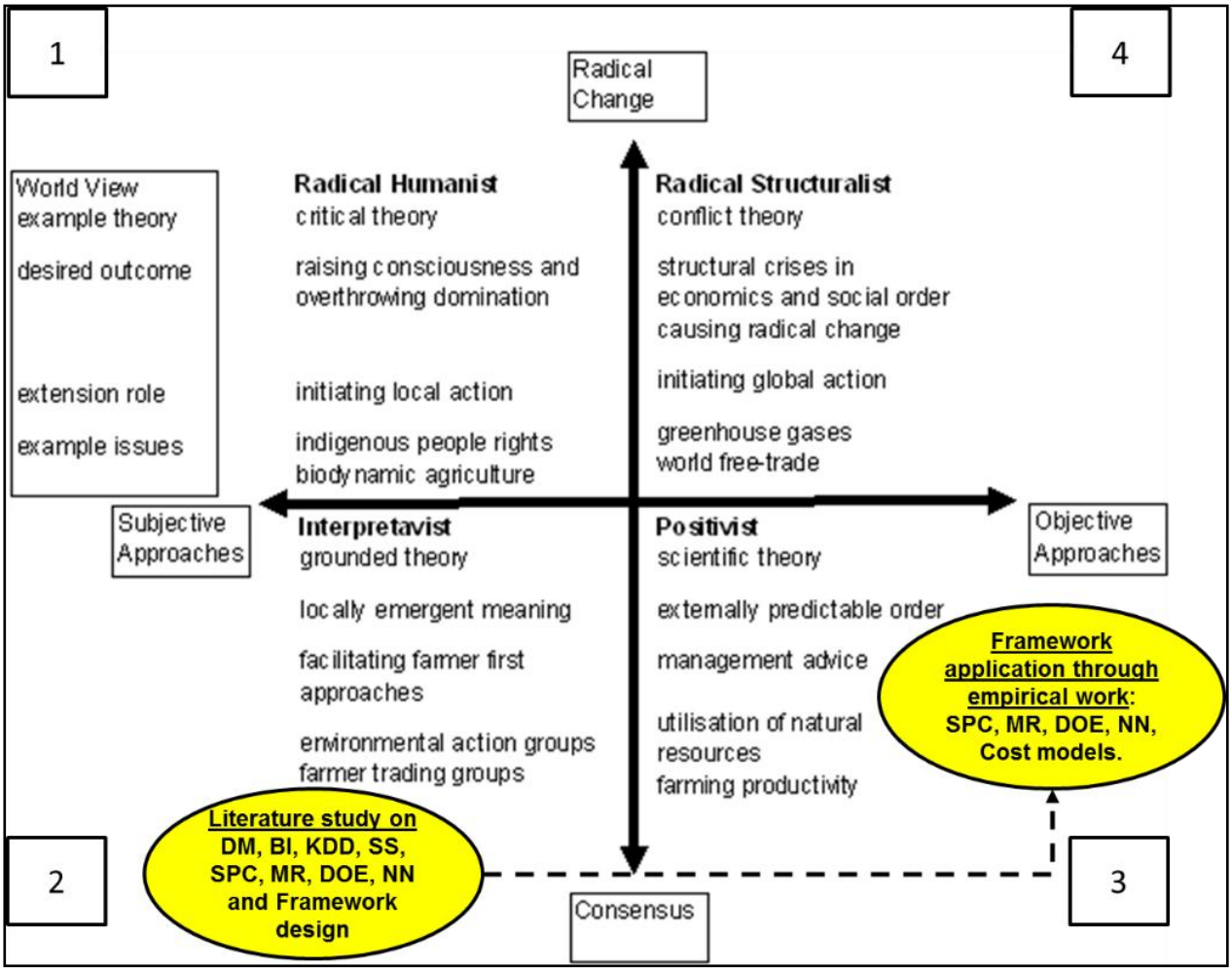


Figure 3.1: Dimensions and attributes of extension worldviews with framework design and application

From an interpretivist approach, quadrant 2, a literature study was done that included frameworks (DM, BI, KDD, DM) and analytical methods (SS, SPC, MR, DOE, NN) to improve a theoretical understanding of selected frameworks and analytical methods for this study. The application of these selected frameworks through analytical methods (SPC, MR, DOE, cost models) shifted the approach to a positivist approach.

The different discussions and research to design a framework fits in interpretivist quadrant 2. It is still subjective and consensual by nature. The application of the framework fits in the positivist quadrant 3. This approach is objective, based on science but still consensual. The Experimental design approach may fit in the radical structuralist quadrant 4 because the implementation is new and radical to its contribution to change. However, there is no guarantee that long-term implementation benefits are sustainable, because the context of the database may change with human intervention, like raw material changes and purchasing policy changes.

From a research methodology perspective, relating to diagrams 3.1, 3.2 and 3.3, this research comprises an interpretivist approach followed by a positivist approach and ends with a constructivism framework due to the Experimental design approach design. For this reason, not a single research approach is applicable to this research, but positivism seems to be the main approach because of the empirical portion of the research.

3.2 DATA MINING METHODOLOGY INTEGRATION

For the research methodology used, DM is discussed first, and then in subsequent sections the integration of the DM methodology with DMAIC, BI, DOE, KDD and BD as a roadmap for data analysists for this study is discussed.

3.2.1 Introduction

Change in operational conditions is dynamic and needs an iterative approach therefor the importance of an iterative analytical process for finding nuggets is evident. In this section from a research methodology perspective the iterative analytical process that contains six distinct methodologies and methods is discussed how it fits and integrates with each other to show an integrated approach for applying the proposed framework.

There are many definitions, perceptions, methodologies and frameworks describing data mining (DM). A general thread throughout all literature is that the foundation of DM is to discover nuggets in vast amount of data with specific analytical techniques. Horníková *et al*. (2011:12) describe DM as an interactive and iterative process of finding knowledge in experimental data sets. This iterative analytical process is typical for DM analysts because since the nineties analysists have experienced that a once off data analysis process is not effective because the change in operational conditions is dynamic and needs an iterative approach. Maimon and Rokach (2010:2) also echo this approach for KDD that expresses the importance of an iterative analytical process for finding nuggets.

Web data mining is a new extension to DM for mining WEB based data in the modern data accumulation era as described by Yu and Shan (2014:1503). Because of the vast amount of data and multiple data sources, a big disadvantage is questionable data integrity. “Bigger the better” in data accumulation, must be handled with caution because the perception that large amounts of data give accurate results is not always

true. WEB data transferred into internal databases are managed and controlled for data integrity through WEB content, structure and usage mining. This means that data used directly from the WEB are not preferred from an analyst perspective.

Yu and Zhang (2014:2) describe a data mining framework for DM as selecting and accumulating data for analysis, classification of data into product families, formulating a designed model and an iterative process to change or modify proposed design. Whether the DM process is described as a methodology by Petre (2013:27) or a framework, both follow a structured analytical process.

Data warehousing as described by Madhuri (2013:330) for BI also complements and sets a platform for Data mining as an analytical process for business applications to ultimately make factual scientific strategic decisions. Data mining is a natural progression from data warehousing. Large amounts of data are accumulated in data warehouses, which, if not analysed, will merely remain data with no value to management. With OLAP the data mining process is transformed into meaningful information through process specific data warehouse design for management that is available online with direct access.

Although Data mining seems to be the answer for large data analysing for business solutions, one of the issues encountered when analysing data, is missing data. Brown and Kros (2003:612) refer to these data as missing at random, data missing completely at random, non-ignorable missing data, and outliers treated as missing data. For this reason, before any data analysis commences, irrespective of the methodology (BI, KDD, DM, BD) followed, data integrity testing is a high priority. Missing data should be part of data cleaning before any data analysis should be attempted.

Data mining not only concentrates on the manufacturing industry, but is also relevant in other sectors, like the service industry. Operating conditions, environment, raw materials, process changes and traditional analytical methodologies will be challenged to validate alternative operating conditions through the DM process. Although DM has grown as a major discipline in IT for analysing industrial data, Thota and Rao (2013:50) confirm that DM is not confined to process data but has evolved into all functions of business where the need arises to analyse data. These function areas include service departments, manufacturing, text mining, multimedia, Web based data, etc. DM is not only for quantitative environments but also for qualitative environments, and is getting

increasingly popular due to the development of software for data analysis in the qualitative field. Each of these areas requires specific mining techniques and is developed as new areas are identified for data analysis.

DM can be used for applications ranging from business management, production control and market analysis to engineering design, and science exploration, marketing, advertising, sales, credit risk, finance, fraud detection and quality control. The most popular area for DM is in process quality to detect non-conforming consumer products. DM application is limitless due to the versatility of DM techniques developed for specific areas as new applications have emerged (Han & Kamber, 2006:1; Washio, 2007:241).

The earlier DM definitions are similar in that they all refer to a similar methodology with the main goal of extracting “nuggets”. The modern definitions still have a methodology as a basis but focus more on “how” data analysis is done. With the explosion of data in resources (WEB, cross functional global company data warehouses, Internet, IT clouds and the drive for more data by big companies), software to assist in analysing data has also become advanced. Therefore, although the basics stayed the same, the analysis and sustainable implementation of results have changed drastically (Adriaans & Zantinge, 1996:4; Berry & Linoff, 1997:5; Berson & Smith, 1997:332, 341; Bigus, 1996:5; Cabena *et al.* 1997:12; Global Intel, 1998; Hand, 1998:112).

Data mining is a multi-disciplinary field of research techniques, which include statistical analysis methodologies in accessing and analysing data from large databases through extraction of useful information, and then endeavours to determine relationships and patterns. DM is a process that focuses on extracting and transforming data, managing and storing data from multidimensional databases to provide data access to analysts.

Hermida *et al.* (2013:411) add a WEB dimension to the traditional BI business environment and call it BIWEB, but go further to suggest that because of this new source of information that complement BI, new data mining and visualisation tools should be developed for management to access knowledge from the Web. They recognise the modern importance of the Web to gain information, how to gather it and the processing thereof. WEB base DM will be part of the modern DM analysis environment, therefore data integrity must have a high priority when accessing data through the WEB. Yu and Shan (2014:1503) also recognise the modern role of Web based data and the importance of it in the modern DM process. Both authors point out

the importance of data integrity for effective data analysis, as well as the development of specific analytical techniques required for Web based data analysis.

The use of all data in data analysis was an earlier option with limited data, but with the current data explosion, it is nearly impossible to use all data, unless you have software to accommodate all data. With modern software, using all data for analysis is becoming popular. Brown and Kros (2003:614) echo this approach, which proposes to use complete data sets to avoid managing missing data. Missing data during data analysis should not be deleted with the assumption of it not having any influence on the analysis, but should be evaluated to establish why missing values occurred, the details for each of the types of missing data, reasons for missing data and the methods of solving missing data - all part of the analysis process.

A plethora DM techniques exist; Petre (2013:23) mentions only a few statistical methods, each with its own purpose in data analysis, but applying Data mining should be structured and methodical. Haphazard analysis will only add to the frustration of analysing data, and in most cases the results are irrational for managers to use. For this reason, the DM application process is more methodology based and less technique driven. Techniques help with analysing data where following a methodology that includes technique application ensures that the analyst stays focused and does not jump to conclusions.

Cao (2007:79) summarizes the difference between traditional data-driven DM and domain-driven DM in Figure 3.2. Traditional DM (data-driven) did not accommodate the real-life modern business problems and needs; the transformation from data-driven to domain-driven was developed through natural progression, focusing on the integrity of results for modern business enterprise. This transformation not only strengthens business intelligence in complex enterprise applications, it must also find ways to integrate human intelligence seamlessly in its processes.

Aspects	Traditional data-driven	Domain-driven
Object mined	Data tell the story	Data and domain tell the story
Aim	Develop innovative approaches	Generate business impacts
Objective	Algorithms are the focus	Solving business problems is the target
Data set	Mining abstract and refined data sets	Mining constrained real-life data
Extendability	Predefined models and methods	Ad hoc, runtime, and personalized model customization
Process	Data mining is an automated process	Humans are integral to the data mining process
Evaluation	Based on technical metrics	Based on actionable options
Accuracy	Results reflect solid theoretical computation	Results reflect complex context in a kind of artwork
Goal	Let data create and verify research innovation; demonstrate and push novel algorithms to discover knowledge of research interest	Let data and meta-synthetic knowledge tell the hidden business story; discover actionable knowledge to satisfy real user needs

Figure 3.2: Data-driven versus domain-driven data mining

Because traditional data-driven DM, in summary, focused mainly on data analysis, algorithm development, theoretical computations and technical advancement of analytical techniques, the contribution of the larger environment was neglected. The transformation to domain-driven DM was to accommodate “real life”, where the environment plays a critical role from data analysis to the implementation of results. Another way of looking at this transformation is to define it as from “hardware driven DM” to “human interaction DM”.

3.2.2 A data mining methodology with statistical analysis

In general, empirical research is a methodology of gaining knowledge through observations, measurements and from actual experiences, and not theories or beliefs. It is based on a repetitive process with generalized analytical steps:

Observation: Collection and organisation of empirical facts of the problem to be analysed

Induction: Formulation of hypothesis to evaluate comparative results

Deduction: Analysis and evaluation of results to test hypothesis

Testing: Testing the hypothesis with new empirical material

Evaluation: Evaluation of the outcome for process improvement or enhancement

A generic data mining methodology, depicted in Figure 3.3, used for data mining analysis, was introduced by Van Blerk (2006:16), which shows an overall framework for data mining analysis. This methodology is similar to generalised data analysis but does not incorporate design of experiments as a data analysis technique. For this research, design of experiments as a data analysis technique for process development will be introduced.

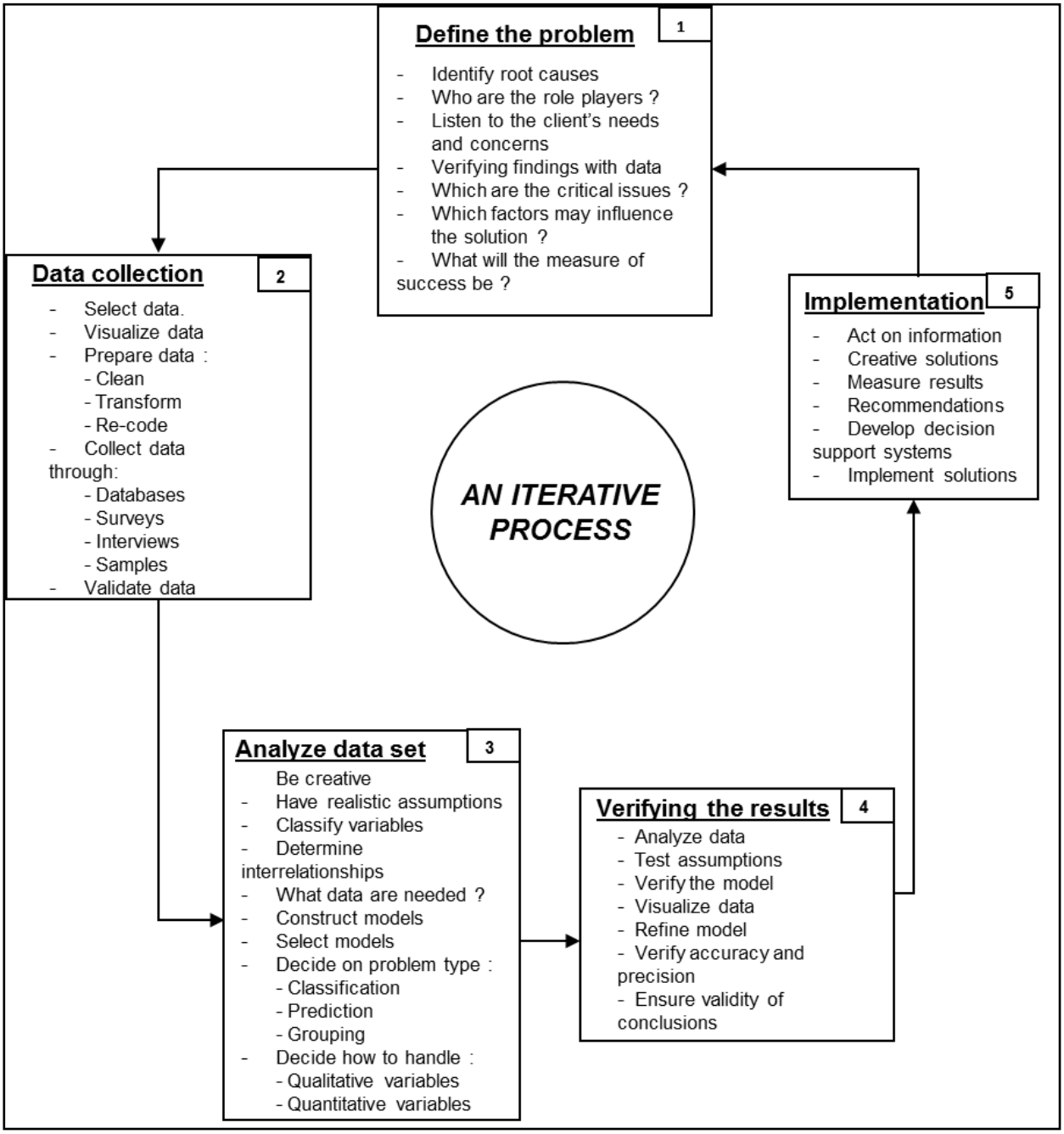


Figure 3.3: Data mining methodology

A brief description for each step of the generic data mining methodology as depicted in Figure 3.3 is:

Step 1: Define the problem

Define the scope of the study, specific goals, clearly defining the objectives and what it wants to achieve.

Identify the root causes for the problem through consultation with system users and key role players, and by studying reports on available data analysis concerning the problem. Ensure that the problem is real and in line with the company's objectives and goals, *inter alia* in terms of quality improvements, cost reduction efforts, competitive advantage programs, and reducing process variability.

Summarise the critical issues applicable to the problem. These issues are mainly from the customer's perspective, based on the reasons for justifying a formal study. All issues or concerns must be validated against the strategic direction of the company.

Try to visualise or identify which factors are critical for the research for analysing and influencing the measures of success once the research is finished. Brainstorming with the role players and affected parties will identify most, if not all, of the necessary factors. If the involvement of all affected parties is neglected, the risk is increased dramatically for selecting non-critical factors.

Step 2: Data collection

Data collection is not a simple process because quality of data will be influenced by the integrity of any results. Data can be collected through:

Samples, which could represent huge populations of data, surveys and interviews with people involved and responsible for that specific area of research and existing databases.

Once the required data are gathered, a tedious process must be followed to prepare the data in an appropriate format for analysis purposes. The process is:

Clean data of all possible human errors when recorded.

Transform data into a workable qualitative or quantitative format if required.

Re-code variables to split them into finer groups if necessary; this is necessary sometimes to search for finer associations amongst variables.

Validate data for integrity purposes.

Step 3: Analyse database

Analysing a database should not be prescriptive but should be left to the creativity of the analyst. Some guidelines may help:

The mixture of quantitative and qualitative variables should be well defined and understood. The type of variables to be used will guide the selection of techniques and methods for analysis.

How the data will be analysed lies with the analyst, but the selection of exploratory techniques to be used will be an extension on how the problem was defined.

Step 4: Verifying the results

Results should not be accepted as correct, but should be scrutinised for possible misinterpretation. Some guidelines are:

Visualise data through graphs or any graphical representation.

Test any assumptions and conclusions made against the historical data and goals set. This phase will quantify the integrity and validity of the analysis as a whole.

If models have been developed, refine them to such an extent that anyone can understand them. By doing this, an opportunity will be created for all managers to buy into it, irrespective of their technical preferences.

Step 5: Implementation

This stage should be a formality only if steps 1 to 4 have been done thoroughly. However, once a new system is implemented, a few things should not be disregarded:

Continuously measure the results obtained and rectify problems.

A support system has to be developed to maintain the new system.

Beware of falling back to the old system due to established comfort zones. When new problems (opportunities) arise, management has to stick with the new system.

Although the above methodology clearly focuses on a generic approach to data analysis as well as which exploratory statistical techniques to use, it does not accommodate the need for advanced statistical techniques like design of experiments to be utilised to explore historical data for process development. This methodology only focuses on data gathering in databases for different applications in business for transforming these data into information, summarised into extracts or reports, which are used by management in discussions and decision-making. Data mining uses various disciplines like statistics, data bases, machine learning techniques, and algorithms for extracting patterns and trends with the main objective of discovering knowledge.

The data mining methodology illustrated in Figure 3.3 does not include design of experiments (DOE) or statistical process control (SPC) as statistical techniques for DM. This DM methodology in Figure 3.3 is limited in its use for this study because it does not reflect two main components, DOE and SPC, that are critical for this study. The changes to the general methodology depicted in Figure 3.3 should include the following:

Include DOE in step 3 (Analysing), SPC in step 4 (Verifying the results) and in step 5 (Implementation) as statistical controlling techniques. By including DOE in step 3, analysing becomes more data driven, flexible, and adds additional analytical power to this methodology.

Include Six Sigma as a supportive data driven methodology that complements the current methodology. The Six Sigma methodology is similar to the above methodology **but** with the focus on data analysis. By having Six Sigma as a supportive methodology to the current methodology strengthens the DM methodology.

Table 3.1 provides a comparative representation between the current DM and the DMAIC methodology.

Current DM Methodology	Supportive Six Sigma Methodology
Define the problem	Define
Collect data	Measure
Analyse data	Analyse
Verify the results	Improve
Implement	Control

Table 3.1: Comparative methodologies

3.2.3 Statistical models

Refer to Table 3.2 below, showing a summary of statistical techniques used for this research. This table represents each statistical technique with its associated purpose of why the technique was used, the variable type the technique represents and the analytical dimensionality. For this study, data mining analysis includes Neural Networks as a supplementary analysis to compare results between the statistical analysis and data mining applications:

STATISTICAL TECHNIQUES	Visualisation	Classification	Relationship	Variable Type		Dimension		
				Continuous	Categorical	One	Two	More
STANDARD TECHNIQUES								
Histograms / Frequency Distributions	X	X			X	X	X	X
Scatter Plots	X	X	X	X			X	X
ADVANCED TECHNIQUES								
Multiple Regression	X	X	X	X				X
SPC (Statistical process Control)	X		X	X			X	
DOE (design of experiments)	X	X	X	X	X			X

Table 3.2: Summary of statistical techniques used for this research

3.2.4 Business intelligence and six sigma

Discussed below is the integration of BI with DMAIC. This integration shows that integration is also within methodologies and not with DM only. The DMAIC methodology compliments the BI methodology as a structured process for data analytics for process improvement. It shows that although the proposed framework shows these two methodologies as sequential steps, they do integrate well to complement each other.

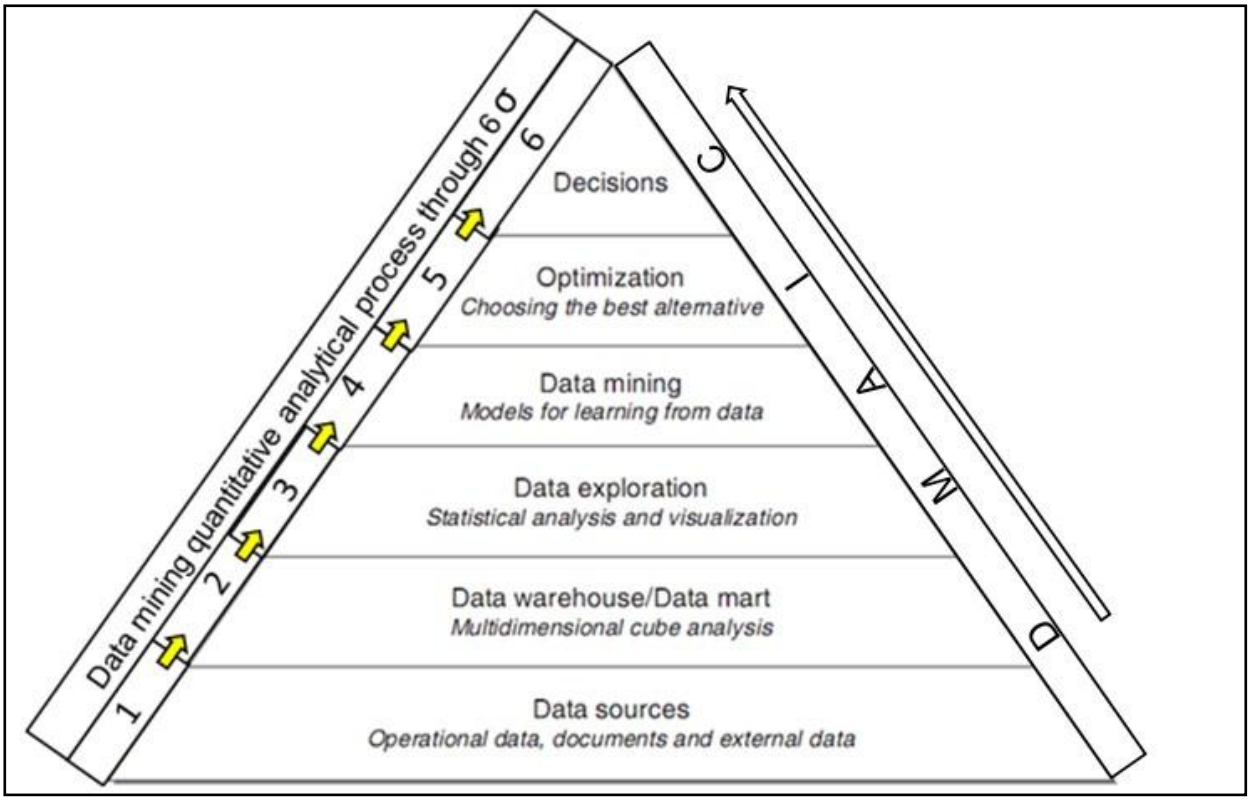


Figure 3.4: Main components of BI and DMAIC

Figure 3.4 by Vercellis (2009) illustrates a typical business architecture, with the main components of BI and methodologies used for BI. This figure summarises BI for data usage, data analysis and possible methodologies to use for exploratory data analysis in comparison to the DMAIC process. The process steps are similar for both BI and DMAIC, which shows the link between BI and DMAIC as well as DMAIC as a supportive data analytical methodology.

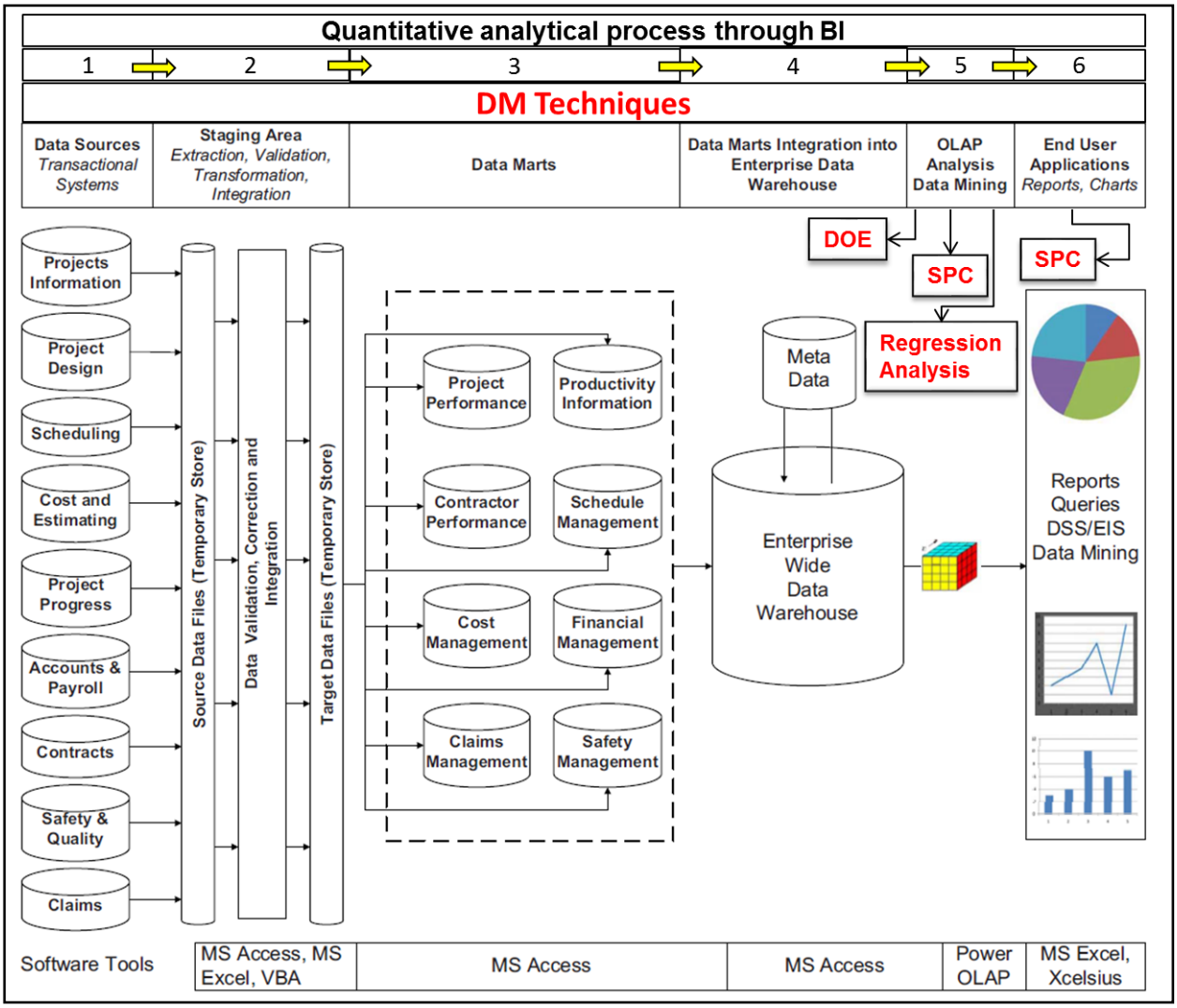


Figure 3.5: Typical business intelligence architecture and statistical methods

In reference to Figure 3.5, Azhar *et al.* (2010:92) illustrate how BI operates within a data base architecture. It shows the quantitative analytical process through BI (refer to processes 1 – 5 in Figure 3.5), a process prior to producing any reports or end user documentation that is assisted by OLAP. The figure was adapted to show where SPC, DOE, regression analysis and DM techniques fit into BI. The BI processes were numbered from 1 to 6, where SPC, regression and DOE fit with BI analysis and DM section. SPC fits both analysis and DM section as well as end user applications, because SPC is also very effective in controlling processes, old or new. DM techniques are applicable to any of the phases using OLAP. For typical BI, the use of OLAP for data analytics when using statistical techniques like SPC, DOE, MR, is not clearly visible. The adapted figure shows the links between these statistical techniques and OLAP. For this study off-line analytical processing was the basis for all analysis done. No on-line data processing for analytical purposes was done.

The importance of Figure 3.5 shows that all multi sourced data required for analytical data mining within an enterprise must be structured to ease the quantitative analytical process. When comparing the main components of BI illustrated in Figure 3.4, the similarities are very visible. This confirms that DOE links to OLAP through BI but they are not dependent on each other.

3.2.5 Knowledge discovery through data

DM does not replace nor is it identical to Knowledge discovery through data (KDD). Only the methodology for DM is similar to KDD. KDD is an extension to DM that concentrates on finite detail for knowledge discovery. DM is more a data analysis application phase to find overall results without refining results. Alazmi and Alazmi (2012:295) describe DM as a knowledge discovery process, it is the analysis step of KDD. These authors also distinguish between DM and KDD in the sense that DM is part of the KDD analytical methodology process. This view supports the evolution from DM to KDD, seeing that KDD became a recognized framework at a later stage than DM.

3.2.6 Big data science

Swan (2015:473) summarises conceptual issues in big data science in Table 3.3. This summary shows that even with advanced technology and the best data analysts, big data must be handled carefully. For this purpose he calls it big data science that refers, not to how, but the way big data are analysed.

Concept	Philosophical Questions
Causality	How should we find causes in the era of 'data-driven science'? Do we need a new conception of causality to fit with new practices?
Quality	How should we ensure that data are good enough quality for the purposes for which we use them? What should we make of the open access movement? What kind of new technologies are needed?
Security	How can we adequately secure data, while making them accessible to those who need it?
Big Data	What defines big data as a new scientific method? What is it (BD) and what are the challenges?
Uncertainty	Can big data help with uncertainty, or does it merely generate new uncertainties? What technologies are essential to reduce uncertainty elements in data-driven sciences?

Table 3.3: Conceptual issues in big data science

Big data, also defined as big data science, is an overall philosophy for managing large data sets in the modern manufacturing and services environment as well as the way big data is analysed, see Table 3.3. These philosophical questions give direction and purpose for BD handled as a science. It is not a methodology of how to handle BD but focuses directly on the wider meaning of BD.

3.2.7 Conclusions

Data warehousing and Data mining complement each other in the sense that Data mining is a natural progression from data warehousing. Large amounts of data are accumulated in data warehouses that, if not analysed, will just remain data with no value to management.

Big data (BD) and Data mining (DM) form an integral part of the proposed framework. Although the holistic focus is on DM for this study, BD forms an integral part and complements DM in terms of analysing large databases.

A progression from the traditional data mining of large databases is the challenge of big data. In some circles, big data and large data sets are used interchangeably, but their goals are different.

Although Data mining seems to be the answer for large data analysis for business solutions, one of the issues encountered when analysing data, is missing data.

The emphasis shows that Data mining is an integrated process of various data analysis disciplines and is not a stand-alone analytical discipline.

Data mining not only concentrates on process industry, but is also relevant in other sectors, like the service industry. Operating conditions, environment, raw materials, process changes and traditional analytical methodologies will be challenged to investigate alternative operating conditions.

3.3 SUMMARY

From a research methodology perspective this research includes an interpretivist approach followed by a positivist approach and ends with a constructivism worldview due to the Experimental design approach design. For this reason, not a single research

approach is applicable to this research but positivism seems to be the main approach because of the large empirical portion of the research.

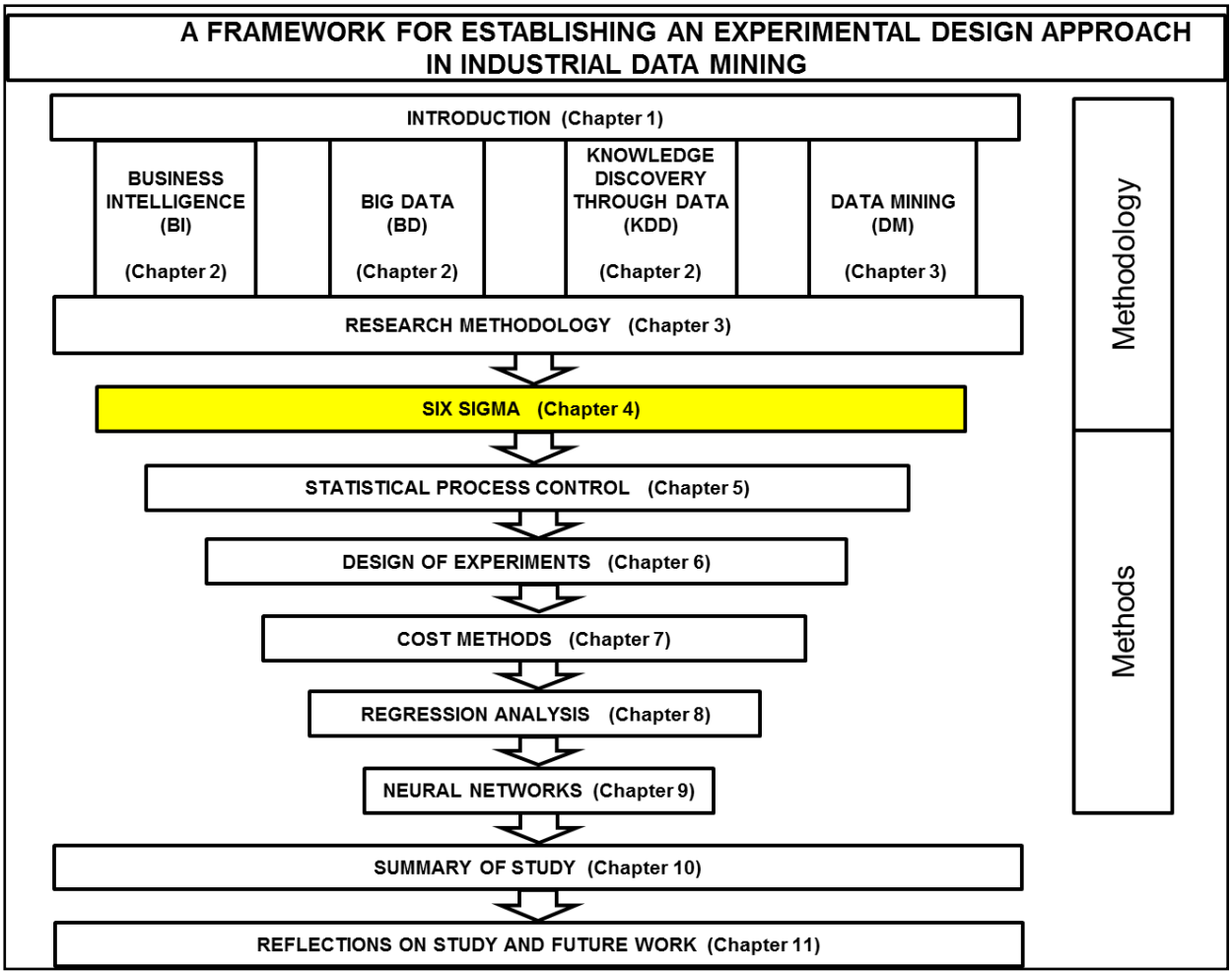
There is not only one data mining methodology to fit all analytical situations. For this study, the following goals should form part of a data mining methodology: Purpose driven, Data driven, Domain driven, Six Sigma driven and DOE driven. These goals will give focus and direction to the analytical process as well as strengthen the proposed framework.

Big data, also defined as big data science, is an overall philosophy for managing large data sets in the modern manufacturing and services environment, as well as the way big data are analysed.

The research approach provides the core content, research parameters and roadmap for this research. All subsequent chapters flow from the research methodology. The statistical methods, SPC, Regression Analysis and specifically DOE, which is a cornerstone for this research, are discussed in the following chapters. The DM technique NN is also part of the framework for this study, and is discussed in a later chapter.

CHAPTER 4

SIX SIGMA



4.1 INTRODUCTION

In this chapter, Six Sigma (SS) as a methodology will be discussed and how it fit within the framework. The company on which the case study is based is independently discussed from SS within this chapter. The reason is that subsequent chapters will focus on the application of the methods based on a case study within this company shown in the framework. It gives the reader a perspective of the company before data analytics start. Whether the company profile is discussed as a separate chapter will make no difference.

This study is an example of a DMAIC project and how it integrates with the framework, with:

D - Chapters 3 and 4

M - Chapter 5

A - Chapter 6

I - Chapter 7

C - Chapter 8

Six Sigma (SS) fits within DOE and covers all stages of a DM methodology. SS methodology embraces the DOE and fits all five stages of a DM methodology. Six Sigma has been utilised for years through a set **DMAIC (define, measure, analyse, improve and control)** methodology as a process improvement tool. This methodology is very similar to the methodology used for data mining in that both want to find patterns and associations in data not normally detected, for process improvement. The similarities and differences in the DMAIC approach are relevant to this research because Six Sigma also focuses on process development and streamlining of processes. Because DOE is an integral part of this study, it is also an integral part of Six Sigma. Design of experiments is a valuable tool part of Six Sigma because it finds relationships that may be hidden within the large amount of data being analysed.

"From all the definitions and discussions concerning Six Sigma and design of experiments, we conclude that Six Sigma is all about understanding and controlling the inputs of processes in order to obtain improvements and design of experiments is used to validate and discover the relationships between inputs and outputs of processes (Peterka, 2008b)."

The two concepts, DOE and SS, are integrated within the SS methodology (DMAIC) process where DOE features as a vital statistical tool for data analysis.

Following DMAIC is a project methodology, the project database upon which the research is based is also discussed because the database integrates with the DMAIC methodology. All subsequent statistical analysis in chapters to follow use the same database.

4.2 SIX SIGMA

Six Sigma is a **management strategy,** according to de Carvalho *et al,* (2014:328) which focuses on improving product quality and streamlining production processes. According to EI-Haik *et al.* (2011:150) and Enoch *et al*. (2015:1189) it is a philosophy, with the business goal of increasing process capability, decreasing process variability, and with the main goal of removing defects from business processes. It is a **framework**

for process quality improvements according to Desale and Deohare (2013:134); a **management methodology** according to Rajkumar and Ramesh (2014:66), with the goal increase process predictability by eliminating defects in order to improve and sustain quality, eliminate waste and achieve sustainable profits.

Management and analysts experience Six Sigma as a strategy, a philosophy, a methodology and a framework and not as single quality improvement process. Irrespective of how users experience Six Sigma, the objectives are the same. The flexibility of the Six Sigma approach for process improvement is evident.

The DMAIC (Define, Measure, Analyse, Improve and Control) process according to Boslaugh (2013:45) is a **graphically driven** process to present visual interpretations for analysis to management. Atkinson (2014:15) suggests using histograms, run charts, box plots and Pareto charts for graphical representation of collected data for analysis. DMAIC is a sequential analytical process used by analysts for process improvement.

Pande and Holpp (2002:2) describe Six Sigma as an analytical approach, not to the benefit of process analysis only, but also to the benefit of customers. This clearly shows that Six Sigma is customer focused through process analytics. Without customer product satisfaction, no business will survive.

Generally, the benefits of using Six Sigma are process improvement related, but could also lead to a cultural change and employee performance improvement in an organization, according to Tyagi *et al.* (2014:139). The Six-sigma process forces management to achieve a deep understanding of production processes, customer needs, data, and statistical analytical techniques, managing, and improving the business. This profound understanding of the business as a whole may lead to a creative company and a progressive company culture. Improving processes is the easy part of quality improvement; employee enhancement and the creation of a motivating climate are a separate but integral challenge.

Both Kai and Basem (2009:741) and Enoch *et al*. (2015:1189) describe a new extension of Six Sigma that is Design for Six Sigma (DFSS), which is a strategy with the ultimate goal of design or re-design a product from the beginning of the process life cycle to develop optimized designs. DSS has four phases, described by the acronym ICOV which are: identify requirements (I), characterize the design (C), optimize the design (O), and verify the design (V). DFSS is different from DMAIC in the sense that DFSS

focuses on the design stage of a process where DMIAC focuses on improving the existing process. Both add value in the optimizing quality improvement process but in different stages of the product or process life cycle.

Business made transformations in shifting from pure Six Sigma that focuses on process improvement using statistical analytical techniques to a holistic approach focusing on the organization as a whole, using business improvement approaches. The goal is to firstly get the organization lean by reducing variability and waste, and then perform Six Sigma. For this reason, Gygi *et al*. (2010:9) introduce a Six Sigma toolset, consisting of defined methods and statistical tools for process improvement and a Six Sigma methodology for improvements in businesses.

Five key concepts of Six Sigma described by Gygi *et al.* (2010:9) are as follows:

- **Six Sigma as a problem solving methodology** − Six Sigma is considered an effective methodology for improving business and organisational performance.
- **Six Sigma performance** – success is measured statistically when a process produces fewer than 3.4 defects per million opportunities for defects.
- **Six Sigma improvements** – drastic improvement is the goal; improvements to key business outcomes and work processes often measure more than seventy (70) percent.
- **Six Sigma deployment** – this is the implementation of Six Sigma methodology. A detailed implementation plan is critical, clearly defining roles of all stakeholders, procedures and an implementation roadmap across an organisation. All employees must buy into this methodology; if not it will be seen as just another “management flavour of the day” initiative.
- El-Haik *et al*. (2011:150) explain that Six Sigma can be used as a **measure of progress** or a benchmark against other companies, processes or products. As a comparative measure of business success Six Sigma is applicable for all types of businesses, whether they represent the services or processing industry.

Pande and Holpp (2002:30) provide seven advantages for management by implementing Six Sigma following DMAIC: 1) Measuring the business problem must be fact based in validating the understanding of the business problem. 2) Focusing on the customer may assist in verifying the problem of origin. 3) Providing factual data will

assist in proving the origin of the problem. 4) Old working methods and routines are challenged and up to date effective new routines will be proposed. 5) Testing viability of proposed solutions reduces the risk of failure when before implementation. 6) Controlling outputs of implemented proposed solutions through factual data. 7) Maintaining process changes by ensuring management support for the new solution.

Surange (2015:283) provides a plethora of benefits of Six Sigma. They are, amongst others, focusing on customers; improved customer loyalty; reduced cycle time; less waste; improved data based decisions; better time management; sustained gains and improvements; systematic problem solving; enhancing employee motivation; **better data analysis before decision making;** team building; improved customer relations; assuring strategic planning; reduction of incident numbers; measuring value according to the customer; increased safety performance; understanding of processes; effective supply chain management; designing and redesigning products/services; deeper knowledge of competitors; developing leadership skills; breaking down barriers between departments and functions; management training; improving presentation skills; integration of products services and distribution; using standard operating procedures; better overall decision making; improving project management skills; sustained improvements; alignment with strategic vision and values; increased margins; greater market share; supervisor training; lower costs to provide goods and services, and fewer customer complaints.

Figure 4.1 describes the Six-Sigma DMAIC methodology as depicted by Gygi *et al.* (2005:44) as:

Define – Describe the process background, key facts and set the project scope. This is the critical stage of the methodology because the parameters of the study are set here.

Measure – Calculate the current process performance and processing capabilities of the process to be analysed and identify the potential contributing problem factors.

Analyse – Reduce potential factors influencing the problem to critical factors of influence. This process uses statistical techniques to gather data for analysis to understand the true influences of process variables to the problems identified.

Improve – Implement proposed new process improvements gained through the analysis phase.

Control – Implement procedures, policies and statistical methods to ensure that improvements are maintained, accepted, and supported by all stakeholders.

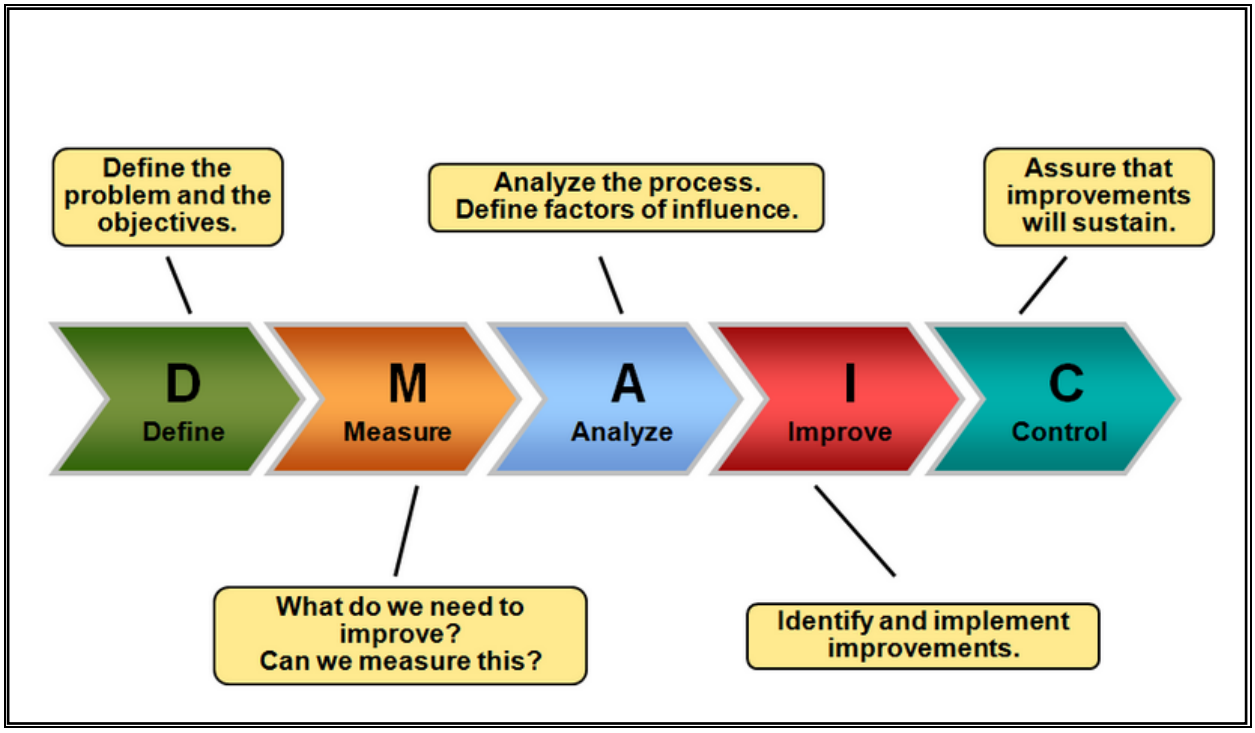


Figure 4.1: DMAIC methodology.

Six Sigma is not only restricted to a production process environment as Desale and Deohare (2013:134) proclaim. They discovered that for the construction project environment no definite goals for performance improvement were evident and therefore feel that Six Sigma may be used by management for a performance indicator. There are many reasons for not using Six Sigma, but the main reason is a lack of knowledge about the Six Sigma methodology and the benefits for implementing it.

The resistance towards using Six Sigma is not new and is expected, seeing that Six Sigma is a relatively new quality improvement program; therefore resistance to change is expected. Karout (2015:7) recognises the resistance to change for business to have Six Sigma as a strategic quality programme and proposes a few concepts to reduce resistance to change in implementing the Six Sigma (D,M,A,I,C) methodology in any company. In summary they are:

Align key business indicators, customer requirements and overall management objectives; involve corporate project sponsors when championing improvement projects; openly support team process improvement activities; reduce barriers for overcoming

resistance to change in the company, and provide needed resources when needed; introduce quantifiable performance measurements to all parts of an organization; identify appropriate metrics early in the process and ensure that these metrics focus on business results and provide incentives and accountability.

Provide extensive training on business improvement, focusing on reducing waste, improving profitability and ensure a sustainable business. Employ highly qualified process improvement experts, internally or externally who can apply improvement tools and lead process improvement teams in setting stretched objectives for improvement.

These philosophical concepts are important for the implementation of the DMAIC methodology because they also focus on the soft issues around implementing a methodology.

Rajkumar and Ramesh (2014:66) believe that successful implementation of Six Sigma is achieved by 1) Reduction of variations in processes; 2) Measuring, analysis, improvement and control of processes; 3) Involvement and dedication from the whole organization including top-level management. The authors also emphasize that Six Sigma is a methodology that focuses on streamlining an organization through elimination of defects, this in turn will reduce total variation within the organization to assist management with strategic decisions.

Six Sigma defined through **DMAIC (Define, Measure, Analyse, Improve and Control)** methodology is a fast growing business management system in industry today.

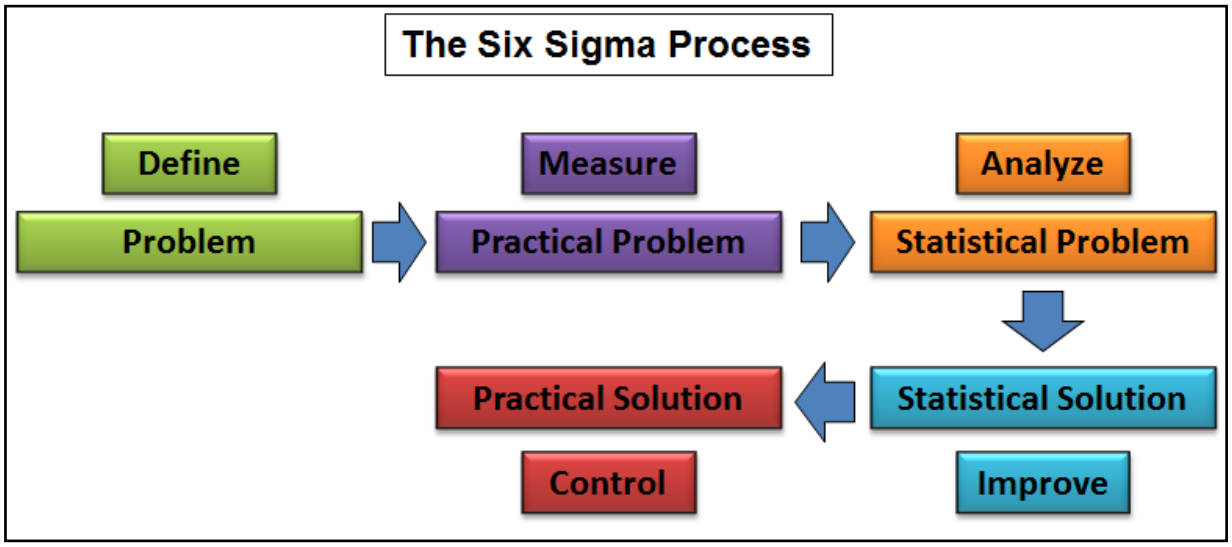


Diagram 4.1: The analytical Six Sigma process

Diagram 4.1 illustrates how the analytical Six Sigma process changes from working on problems to finding solutions. This transition in the analytical process shows how a structured problem solving process switches between practicality and the use of statistical methods.

Diagram 4.2 represents a broad Six Sigma framework that shows at which stage of the DMAIC process the identified problem is recognized as a statistical or practical solution. This is an important representation for the analysts to guide them, to structurally assist them with what they are busy with during which stage of the analytical process.

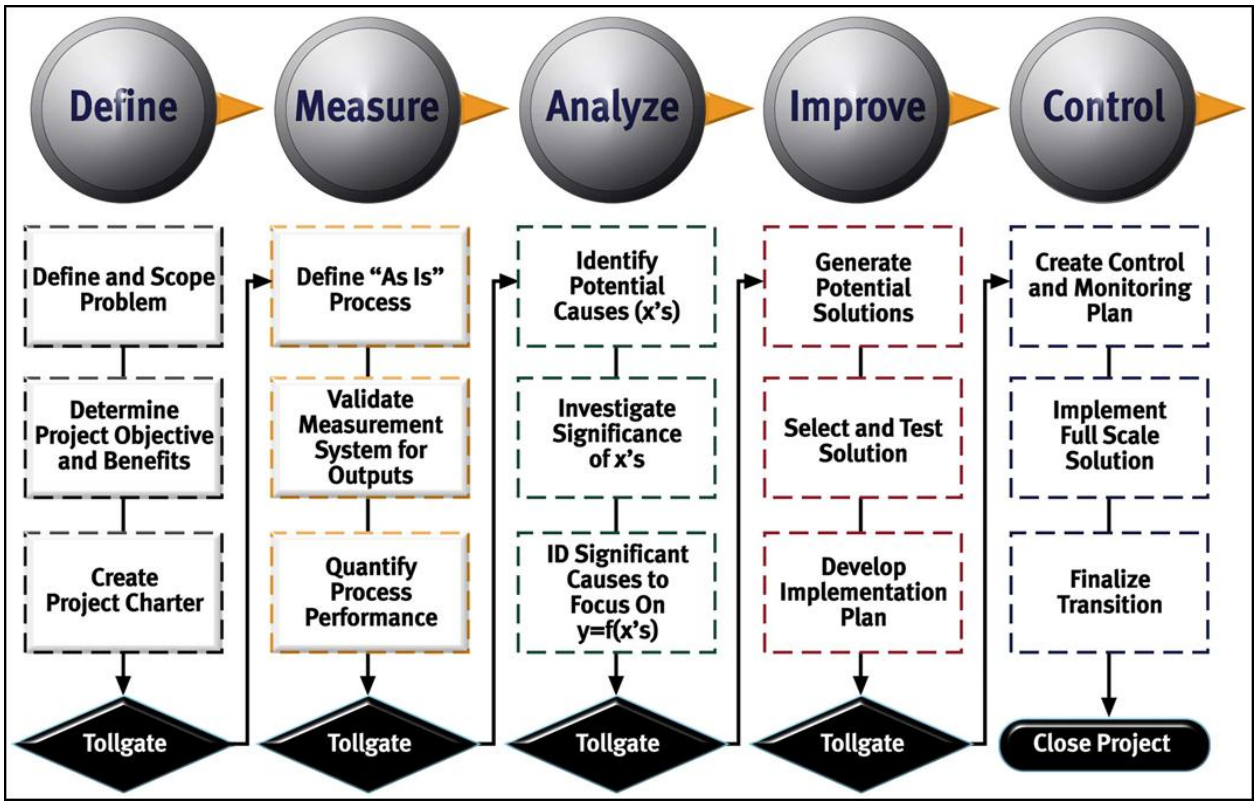


Diagram 4.2: The Six Sigma process

Diagram 4.2 shows the DMAIC process that refers to a data-driven improvement cycle used for designing, improving, optimizing and controlling business processes for continued improvement. The SS methodology is based on the combination of basic statistical quality control techniques, simple and advanced data analysis methods, as well as systematic training of all personnel at every level in the organisation involved in process or system development. This methodology fits in well for this research showing the importance of a data-driven methodology that, on a macro basis, is used for process development.

A new extension to Six Sigma is Design for Six Sigma (DFSS) which is the strategy with the ultimate goal of designing or re-designing a product from the beginning of the process life cycle to develop optimized designs. Figure 4.2 by Enoch *et al*. (2015) represents a P-diagram that illustrate DFSS. It is the process model for the Taguchi Robust Design method with the main goal to improve productivity. The P-diagram classifies variables associated with product design into inputs (M), noise factors (Z), design parameters and output (Y), and design parameters or control factors (X). During the development stage inputs (M) and output (Y) associated with the design concept are first identified, and then factors beyond the control of the designer, called the noise factors (Z). Design parameters or control factors (X) specified by the designer are then determined. The robust design method is a fundamental part of design of experiments (DOE) that will be addressed during the Six Sigma application phase in subsequent chapters.

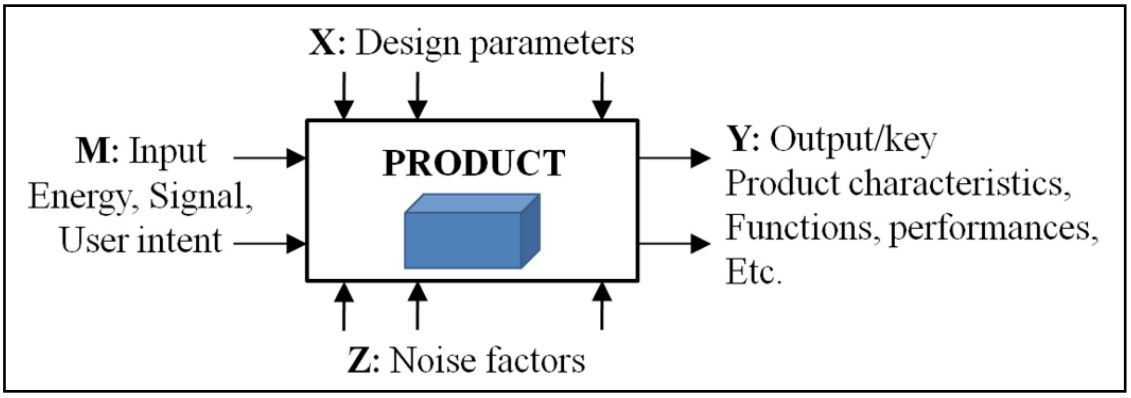


Figure 4.2: Taguchi P-diagram

In DOE an important shift also developed from concentrating in optimising the average performance parameters of dependent variables to designing for reducing variation which complement design for Six Sigma (DFSS).

In a study, to improve the efficiency of silicon solar cells to stay competitive in the crystalline silicon solar cell technology, Saravanana *et al*. (2012:143) describe the positive impact of two approaches for conducting Six Sigma projects in the Silicon industry, the first being DMAIC (Define, Measure, Analyse, Improve and Control) and the second being DFSS (Design for Six Sigma). During their efficiency improvement process, incorporating DFFS into the DMAIC methodology as a complimentary process assisting in improving the efficiency of the multi-crystalline silicon wafer, was a huge success. Using this combined process improvement approach gave them a quantum

improvement in S/N ratio. This study shows that DMAIC and DFSS methodologies could be very effective when combined and are not restricted to be used as independent process improvement methodologies.

4.3 CASE STUDY

4.3.1 The company

This case study is of a local company, which forms part of five international manufacturing plants worldwide. All of these plants are a processing blue print of one another, therefore any processing changes or product enhancements are globally implementable. This group of companies is listed as one company on the New York stock exchange. This group of companies is the single largest producer of the product worldwide and they maintain a dominant footprint in this market. The core business is to provide their product to the steel industry for electrical arc furnaces to smelt steel. There are many products that an electrical arc furnace can utilize to smelt steel, but this product is currently the only cost effective product for an electrical arc furnace to smelt steel.

4.3.2 The process

For this case study, a manufacturing process was identified that produces products for a niche market within the steel industry. This production process consists of five inline sequential processes for a product to follow from raw material to the final product. Each product must follow the sequential processing steps and cannot bypass a process; therefore are sequential processing dependent. Each of the five production processes adds unique value added characteristics to the product and has its own processing parameters. A determining factor for this research is the long processing time of three months to process one product. This manufacturing process does not allow the privilege of evaluating results quickly for every corresponding process parameter change. For this reason, the risk of producing non-conforming products when operational changes are made is high and very costly.

One product with its associated processing data representing the first of five-production processes was selected for this case study. All actual processing data for this product were transformed to represent fictitious data to protect technical product processing standards. The fictitious data have no correlation to actual processing data for this

process, but for the purposes of this case study provide data that represent a product produced at a different processing level. The database used represents a specific product that ensures variability within and between variables, is representative to this specific product and, therefore, not compromised by other similar products. Therefore, the selected database reflects the processing data from one product manufactured in the first process.

4.4 PROJECT DATABASE

For this research, the database used for the DMAIC project, is discussed in the following sections.

4.4.1 Database description

The database utilised for this research in complimenting the proposed framework is applicable to one product only. This one product of the business is approximately 97% of the business. The remaining 3% is speciality products. Data captured for this product which is used, are process data gathered during the production process, by the quality function to monitor for analysing of data to assist in process and product improvement. This database is specific to one production process and one product only. Although the reference database is a small portion of data within the company database, the focus here is to show how realistic the framework is by utilising a selected database based on product data for quality improvement. This database consists of Raw process database, Cleaned database, Transformed database and Working database.

These databases include only one product manufactured with all the processing parameters applicable for each product produced. The data are a composition of ***quantitative*** and ***qualitative*** data types.

According to Wagner (2016:10),:“Quantitative random variables generate numeric response data. These real numbers that can be manipulated using arithmetic operations (add, subtract, multiply and divide).”

These variables are numerical values that are of continuous nature, which can take any value in a specific range for example, temperature, product weight, processing speed, pressure, etc.

Wagner (2016:10) also defines qualitative variables as “Qualitative categorical (non-numeric) response data. These data are represented by categories only”. These variables are categorical of nature, which can be discrete or dichotomous. Discrete values are a countable number of distinct possible values. Only discrete values are applicable for this database, for example: mixer pair and cooler number.

4.4.2 Database transformation

The raw database was reduced through seven phases to a working database that was used for this study when applying the proposed framework. A description of the seven phases is as follows:

Each independent and dependent variable was tested for values that do not belong to the core database. During this data integrity-testing phase, variables (columns) and records (rows) were removed from the database in seven phases on the following basis:

Phase 1: Independent variables with fixed processing mechanical settings or set points, determined by global R&D. For this study these set points are not challenged.

Phase 2: Missing or zero values. These are a direct result of system recording failure and therefore cannot be included. If included they will only add to uncontrolled system variation that could lead to high prediction error.

Phase 3: For values outside process specification, which were found to be bad value recordings, the complete record was removed.

Phase 4: Independent variables with no variability. Values for these variables will not change, irrespective of model design, because it is raw material and product specific.

Phase 5: Variables not in use anymore for product processing. These variables were kept in the database that was part of the original structured query language (SQL) design within the process database. Since then processing parameters and critical variable selection have changed but have never been taken out.

Phase 6: Variables that are not process related.

Phase 7: Process experience and technical processing knowledge. This phase is commonly overlooked, but is maybe the most critical criterion because if people with

sound process knowledge are not part of this process, it will be very difficult to expect from them to buy into any solution found during the analysis phase.

Table 4.1 represents a summary of the seven phases for reducing independent and dependent variables. This table shows the reduction in records and variables for each phase resulting after each phase. In summary, there were 15414 records and 44 variables (dependent and independent) before data cleaning started. At the end of the seventh phase there were 9538 records and eight (8) variables (dependent and independent) left. Also refer to Appendix 2 for a sample of the reduced database ending with eight variables.

Case study detail will now be part of all subsequent chapters 5 – 9 that follow the framework.

Phase	Original variables (independent & dependent) Qty	Original records Qty	New variables (independent & dependent) Qty	new records Qty	Cumulative Variable reduction	Record reduction	Cumulative Record % loss
Raw DB	**44**	**15414**	**44**	**15414**	**0**	**0**	**0.00%**
1	44	15414	26	15414	18	0	0.00%
2	44	15414	24	15414	20	0	0.00%
3	44	15414	20	12754	24	2660	17%
4	44	15414	20	9538	24	5876	38%
5	44	15414	17	9538	27	5876	38%
6	44	15414	13	9538	31	5876	38%
7	44	15414	8	9538	36	5876	38%

Table 4.1: Database variable reduction summary

4.4.3 File identification

After completing phase 7, it becomes apparent how critical database design is because data integrity must be one of the prime design drivers. For records to reduce from 15400 to 9500, 38 percent record reduction is required and is evidence of weak database integrity.

For this exercise, it means that at approximately 40 percent of data recorded add no value for management and cannot be used for analysing data for strategic decision making, they only fill up space in database. For this reason, when designing a database, all users affected, analysts and management must be part of the final implementation solution.

The relevant files from which data had to be extracted for the creation of the raw database were identified as: Production throughput master file; Process specific processing file; Process data file, and Product master file.

Each of these files is an independent module from which combinations and portions of data were extracted to create the raw master file.

4.4.4 Extracting of data

The raw database was extracted with a structured query language (SQL) query tool accessing the relevant process data from the selected production process for this study.

4.4.5 Field identification

The final database comprising seven independent and one dependent variable from the original 44 variables was identified as:

Description	Data Type
Mix discharge temp	QUANTITATIVE
Cool begin temp	QUANTITATIVE
Actual cool time	QUANTITATIVE
Actual dump temp	QUANTITATIVE
Actual tamp pressure	QUANTITATIVE
Actual extrusion rate	QUANTITATIVE
Actual extrusion speed	QUANTITATIVE
Actual average extrusion pressure	QUANTITATIVE

These fields contain the final data necessary to illustrate how quantitative analysis supports the proposed framework. The reduction of 36 variables may cause analytical accuracy loss in the final results because some variables that were discarded may still have an impact, but for the purpose of this study where techniques are used for illustrations only, finite accuracy is not applicable.

4.4.6 Conversion of data

No conversions were necessary because the JD Edwards platform converts all process data into a standard format before making them available to all users to access for any data related needs.

4.4.7 Transferring of data

The raw database was transferred into a Personal computer using Microsoft Excel, through the client access utility. Because Statistica was used for the quantitative analysis, the raw data excel file was easily imported into the Statistica database. The data were then ready for processing by a Personal computer using Statistica by Dell Inc. (2015). (Dell Statistica (data analysis software system), version 13. software.dell.com.)

4.5 SUMMARY

The main drive for Six Sigma organisations using Six Sigma methods and tools is to drive down costs, grow revenue, improve customer satisfaction, increase production

capacity and capability, reduce process complexity, lower cycle time, minimise defects and process waste. In summary, **it is a process waste reduction and variability reduction initiative.**

A high degree of data integrity is usually taken for granted in strategic decision making, which is a dangerous assumption without validating data accuracy. By utilising the DMAIC methodology, validating the integrity of data will consequently lead to data credibility. Data analysis practitioners will benefit by having a profound understanding of their processes before attempting to optimise them.

Six Sigma is **an extension from the theories of Total Quality Management (TQM)** and Statistical Process Control (SPC) by incorporating elements of the Shewhart/Deming Plan, Do, Check and Act(PDCA) continuous improvement cycle.

Six Sigma is not only effective in manufacturing but in all spheres of business. Every person, irrespective of his or her role in the organization, will benefit when implementing the Six Sigma methodology to assist in improving strategic decision-making.

For this research, only selected techniques from the DMAIC process, for fitting, cleaning of data, design of experiments, predicting process behaviour and model building are used. The technique selection process was done to show the power of using basic statistical techniques in a complex analytical environment. The analytical process, however, shows a progression which started with univariate analysis (Histograms) then progressed to bi-variate analysis (Statistical Process Control) then to Multivariate analysis (Multi Regression, DOE).

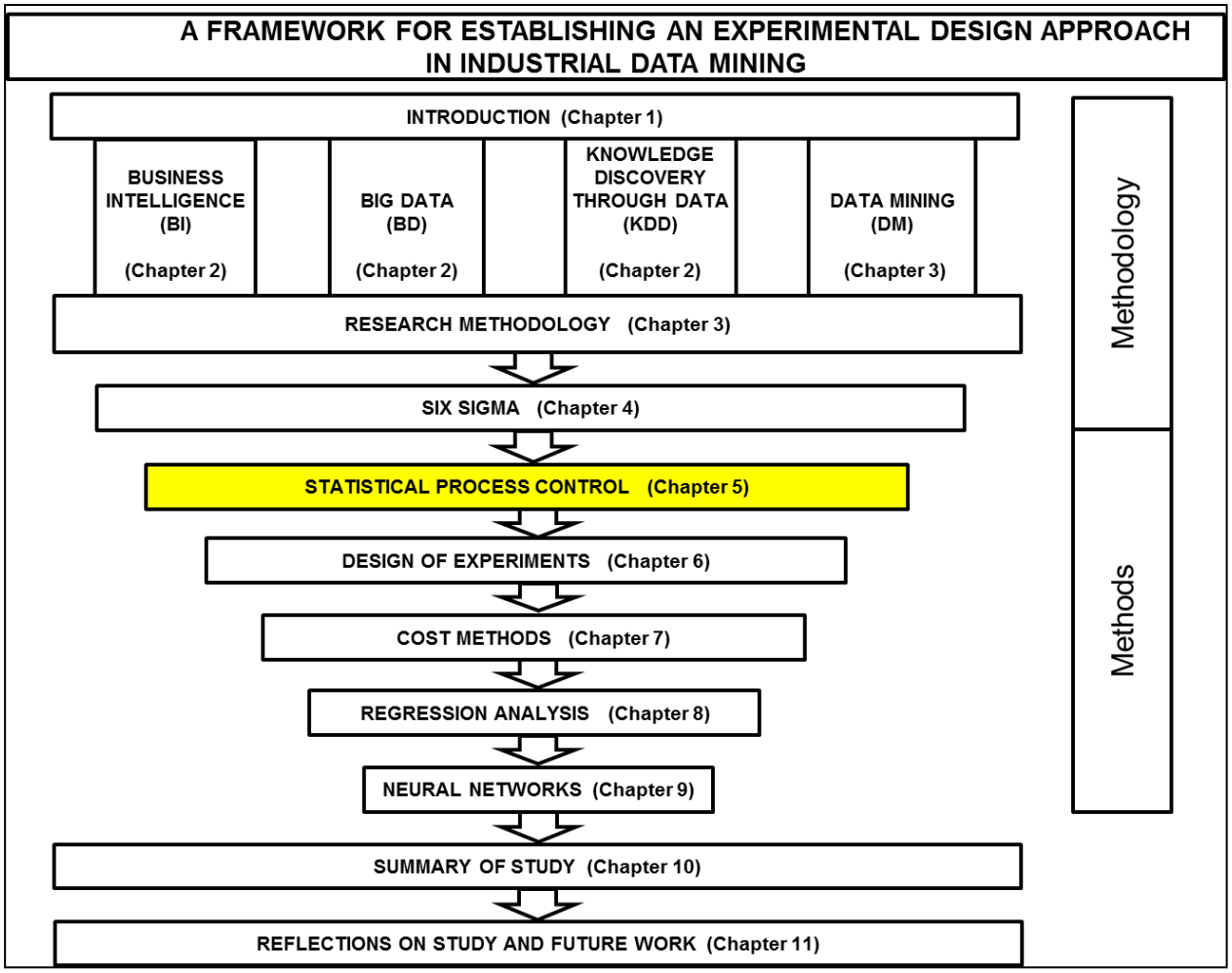
The DMAIC improvement cycle is a process used to drive Six Sigma projects in various fields. DMAIC is not exclusive to Six Sigma and is used as a framework for other improvement applications in quality management. It implements the idea of continuous process improvements where the constant measuring of processes for possible improvement opportunities exists. DMAIC contributes to the creation of a conceptual framework for consistent performance measurement, process improvement, and control of any process, quantitative or qualitative. A DMAIC project has a short duration compared to product development projects. It is a systematic, scientific and a fact-based approach.

This study is an example of a DMAIC project, with:

D (Define)	Chapter 3: Research Methodology Chapter 4: SS
M (Measure)	Chapter 5: SPC
A (Analize)	Chapter 6: DOE
I (Improve)	Chapter 7: Cost Methods
C (Control)	Chapter 5: SPC Chapter 8: MR Chapter 9: NN

CHAPTER 5

STATISTICAL PROCESS CONTROL



5.1 INTRODUCTION

The goal for this chapter is to identify if there is room for process improvement in this case study. This important because room for potential improvements is the catalyst for ongoing data analysis. For this research, a manufacturing process was identified that produces products for a niche market within the steel industry. This production process consists of five inline sequential processes for a product to follow from raw material to the final product. Each product must follow the sequential processing steps and cannot bypass a process, and therefore products are sequential processing dependent. Each of the five production processes adds unique value added characteristics to the product and has its own processing parameters.

One product, with its associated processing data representing the first of five production processes, was selected for analysis. All actual processing data for this product were transformed to represent fictitious data to protect technical product processing standards. The fictitious data have no relation to actual processing data for this process, but for the purposes of this study, provide data that represent a product produced at different processing level. The database used represents a specific product that ensures variability within and between variables, is representative to this specific product and, therefore, not compromised by other similar products. Therefore, the selected database reflects the processing data from one product manufactured in the first process.

Because one of the goals of this study is to analyse historical data to predict future processing behaviour through DOE and regression analysis, the database was divided into two parts. The first part consists of processing data from 2005 to 2009 and the second from 2010 to 2013. The first part will be analysed to predict the validation period. Comparative results will then be used to predict future processing data.

Statistical process control (SPC) as a statistical technique will be used as a variable selection technique on input data to determine which independent and dependent variables are significant for DOE and regression analysis. SPC will not feature in the control process in this study because proposals could not be implemented owing to the shutdown of the company.

5.2 STATISTICAL PROCESS CONTROL

Wheeler and Chambers (1986:21) illustrate and explain the importance of statistically controlling all processes in a company. They illustrate the balance between process improvement and entropy that causes a process to be either in the ideal state or in chaos.

Statistical process control (SPC) is a quality control method used to monitor and control a process. Oakland (2012:3) indicates that SPC is not about statistics or control, it is about gaining the competitive edge in the market. He states that SPC comprises three key elements: quality of product or service to the customer, all processes that produce products or provide a service and the control of those processes. Further, all organisations are in competition on quality of the product offered, promised delivery and competitive price strategies.

Xie and Kruger (2012:3) stipulate that SPC has been introduced into the general manufacturing industries for monitoring process performance and product quality and to monitor the general process variation that is caused by a few key process variables.

SPC is a method mainly used to monitor and control processes within a business. All industries in product manufacturing and service delivery sectors can use and apply it. Figure 5.1 shows a typical methodology behind SPC, starting with inputs like materials, labour and facilities, then utilising the production phase, which produces outputs. These outputs are then subjected to SPC, which evaluates them for statistical control, and if not in statistical process control, adjusts the entire process to the designed process parameters.

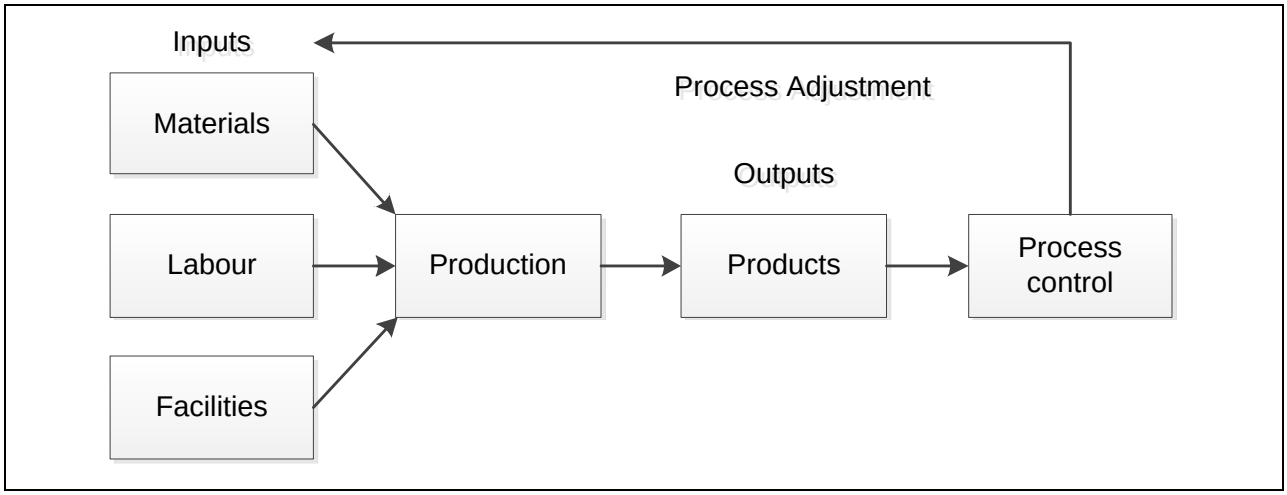


Figure 5.1: A SPC methodology

Process capability is one of the fundamentals of applying SPC, because in order to improve or enhance a process, the process must be capable of meeting customers' product specifications continuously and in a predictable manner.

According to Ryan (1989:172), process capability is when a process operates within specified processing limits, ensuring that the independent variables producing products lie within processing specifications to conform to customer expectations. The ratio between the variation of the independent variables compared to the customer specifications determines the process capability.

Liu *et al.* (2014:4000) confirm that SPC is one of the basic concepts of modern quality management. SPC is based on the theory of statistical variation of any process or outcome, measured over time, with the goal of improvement or maintaining a high level

of process capability. Various statistical process control charts are available for measuring processing variation. The appropriate charts depend on the process type being measured, and show the process variation for the period the process is observed or measured.

5.3 DISTRIBUTION AND CAPABILITY ANALYSIS

5.3.1 Distributions

Frequency distributions are used mainly to condense large sets of data. Histograms are the graphical representation of frequency distributions, usually displayed in one dimension.

Freund (1988:25) describes a histogram as the representation of measurement or observations grouped on a horizontal scale, the class frequencies on a vertical scale, and drawing rectangles whose bases equal the class intervals and whose heights are the corresponding class frequencies.

Hosking and Wallies (2005:1) state that frequency analysis is the estimation of how often a specific event takes place. Frequency distribution can be defined as the grouping of the values that one or more variables have taken in a sample to show the occurrences of values within a particular group and summarise the distribution of values in the sample. The frequency distribution technique can be utilised to determine the frequency of a specific event in a data set.

Brightman (1999:113) discuss sampling distributions, which is also applicable for SPC. For the x-bar chart, which is based on the central limit theorem, that sample means of the population should be expected to form a bell shape or a normal distribution as the sample size increases.

Hines and Montgomery (1990:317) state for the s-chart that is based on the sample standard deviation, the sample standard deviation distribution is normal for a large sample size.

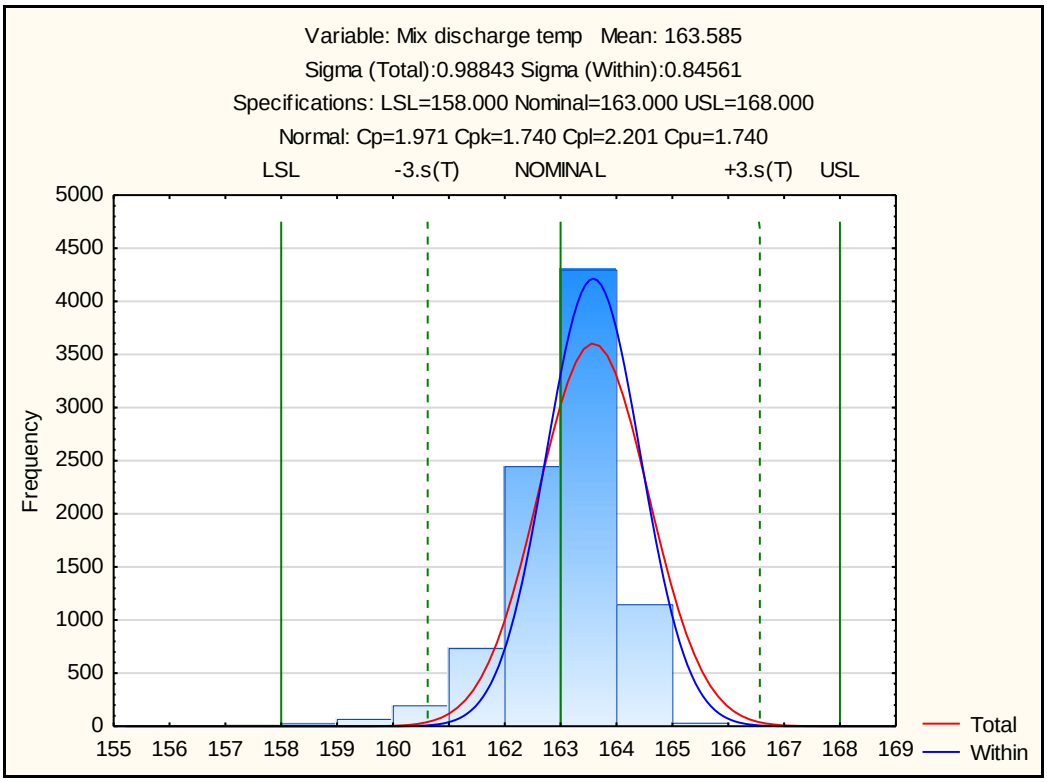
Using SPC as a process evaluation method, the presence of normal like shapes for both sample mean and sample standard deviation for an independent variable serves as an indicator for the analyst finding room for process improvement.

5.3.2 Capability analysis

The assumed shape for capability analysis for these variables is expected to be normal like or bell shaped. Capability calculations are based on the within sample standard deviation which is an estimate based on a sample size of 40, that represents approximately one week of production.

To illustrate distribution shapes, capability analyses for selected independent and dependent variables were done. Keeping in mind when comparing the shape for the capability analysis to the x-bar for SPC analysis, the spread will be narrower for the x-bar, seeing that the represented data are sample averages (x-bar) and not individual data.

Combining SPC and DOE results for each independent variable will provide a better understanding of proposed operating levels when searching for room to process improvement.



Graph 5.1: Mix discharge temperature: Process capability chart

Graph 5.1 shows a C_p index of 1.971 (the ratio of the specification range over the process range) which is greater than 1, indicating a very capable process. The C_{pk} index of 1.740 (demonstrated excellence) is greater than 1, indicating a capable process

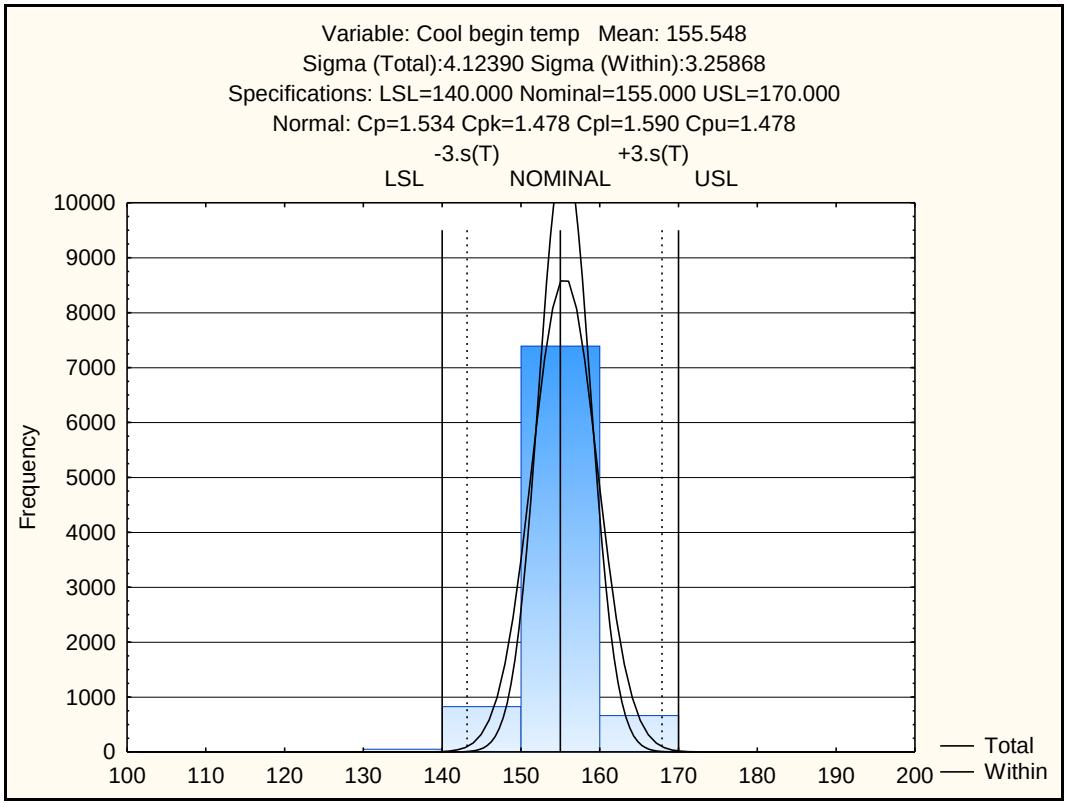
when the mix discharged temperature process is centred. C_{pl} index of 2.201 (lower capability index) and C_{pu} index of 1.740 (upper capability index) are both greater than 1, but because these two indices are different, it shows that the process is not centred.

All process capability indices are greater than 1, which provide a process improvement opportunity by reducing specifications or adjusting the process mean for mix discharge temperature towards a higher or lower operating level.

The distribution shape is not normal like but skewed to the left, which indicates low value fliers that have probably no influence on the process.

The DOE analysis should indicate on which operating level measured in cost of producing out of specification products, this variable should operate in order to optimize product quality.

At this stage, we only measure the overall operating level of 163.585 from Graph 5.1, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.2: Cool begin temperature: Process capability chart

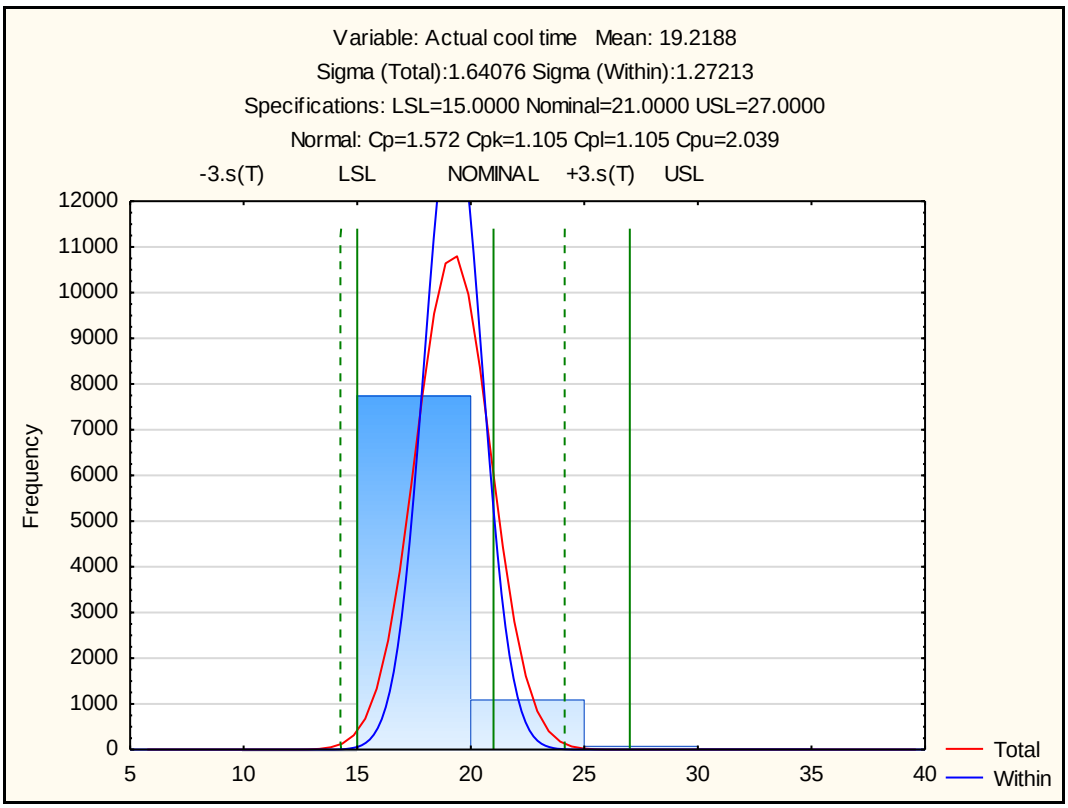
Graph 5.2 shows a C_p index of 1.534 (the ratio of the specification range over the process range) greater than 1, indicating a very capable process. The C_{pk} index of 1.478 (demonstrated excellence) is greater than 1, indicating a capable process when the cool beginning temperature process is centred. C_{pl} index of 1.590 (lower capability index) and C_{pu} index of 1.478 (upper capability index) are both greater than 1, but because these two indices are different, it shows that the process is not centred.

All process capability indices are greater than 1 which provide a process improvement opportunity by reducing specifications or adjusting the process mean for cool begin temperature towards a higher or lower operating level.

The distribution is normal like with a few low value fliers that have probably no influence on the process.

The DOE analysis should indicate on which operating level measured in cost of producing out of specification products, this variable should operate in order to optimize product quality.

At this stage, we only measure the overall operating level of 155.548 from Graph 5.2, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.3: Actual cooling time: Process capability chart

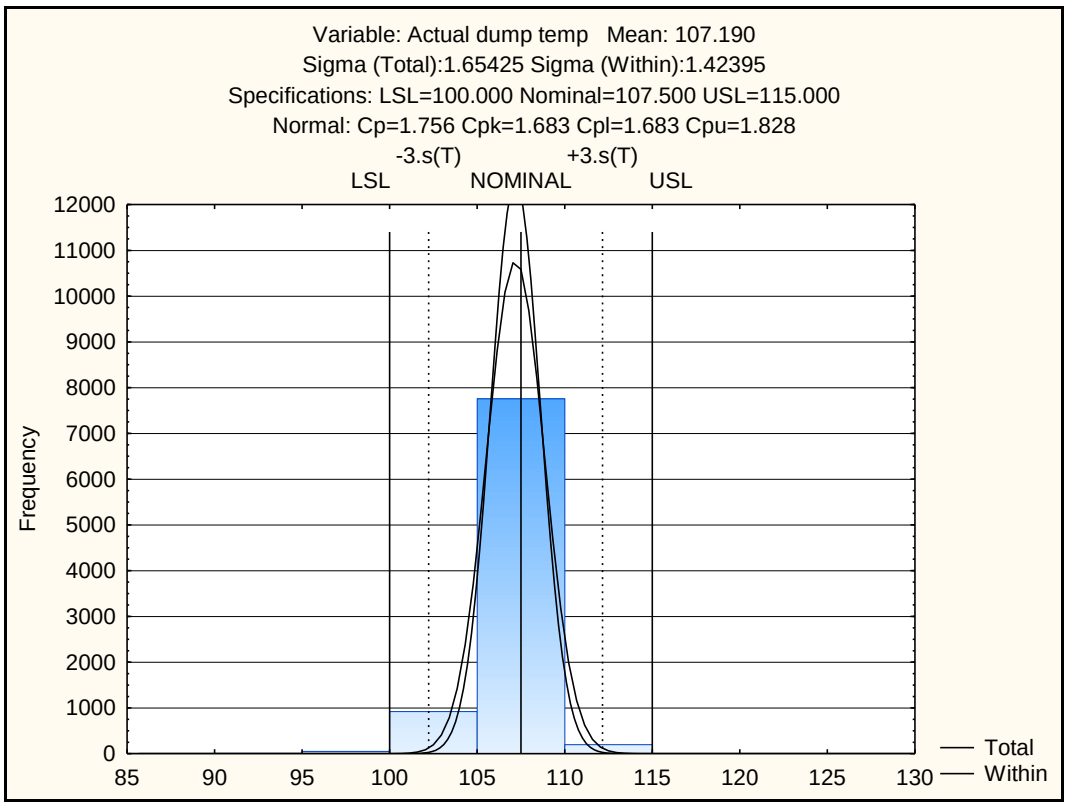
Graph 5.3 shows a C_p index of 1.572 (the ratio of the specification range over the process range) greater than 1, indicating a very capable process. C_{pk} index of 1.105 (demonstrated excellence) is greater than 1 indicating a capable process when the actual cooling time process is centred. C_{pl} index of 1.105 (lower capability index) and C_{pu} index of 2.039 (upper capability index) are both greater than 1, but because these two indices are different, it shows that **the process is not centred**.

All process capability indices are greater than 1 which provide a process improvement opportunity by reducing specifications or adjusting the process mean for actual cooling time towards a higher or lower operating level.

The distribution is not normal like and skewed to the right, showing a few high value fliers that have probably no influence on the process.

The DOE analysis should indicate which operating level measured in cost of producing out of specification products, this variable should operate on to optimize product quality.

At this stage, we only measure the overall operating level of 19.2188 form Graph 5.3, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.4: Actual dump temperature: Process capability chart

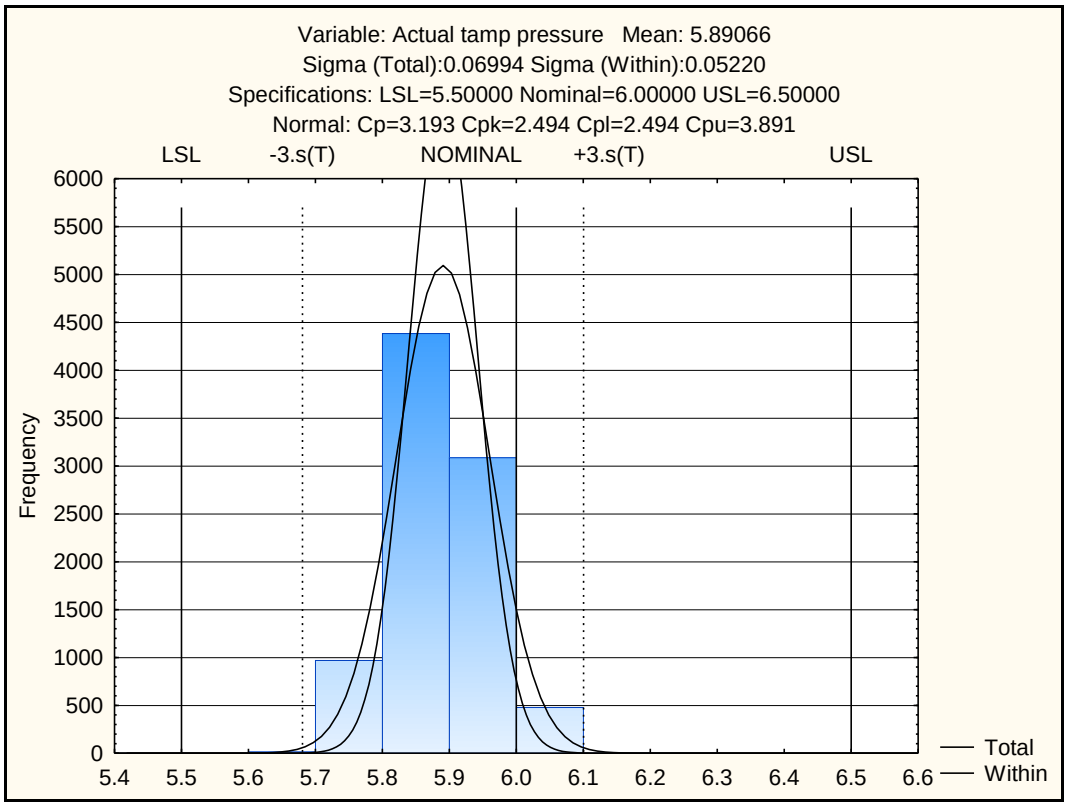
Graph 5.4 shows a C_p index of 1.756 (the ratio of the specification range over the process range) is greater than 1 indicating a very capable process. C_{pk} index of 1.683 (demonstrated excellence) is greater than 1 indicating a capable process when the actual cooling time process is centred. C_{pl} index of 1.683 (lower capability index) and C_{pu} index of 1.828 (upper capability index) are both greater than 1, but because these two indices are different, it shows that the process is not centred.

All process capability indices are greater than 1 which provide a process improvement opportunity by reducing specifications or adjusting the process mean for actual dump temperature towards a higher or lower operating level.

The distribution is not normal like and shows a few low value fliers that probably have no influence on the process.

The DOE analysis should indicate on which operating level, measured in cost of producing out of specification products, this variable should operate in order to optimize product quality.

At this stage, we only measure the overall operating level of 107.19 from Graph 5.4, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.5: Actual tamp pressure: Process capability chart

Graph 5.5 shows a C_p index of 3.193 (the ratio of the specification range over the process range) is greater than 1 indicating a very capable process. C_{pk} index of 2.494 (demonstrated excellence) is greater than 1 indicating a capable process when the actual tamping pressure process is centred. C_{pl} index of 2.494 (lower capability index) and C_{pu} index of 3.891 (upper capability index) are both greater than 1, but because these two indices are different it shows that the process is not centred.

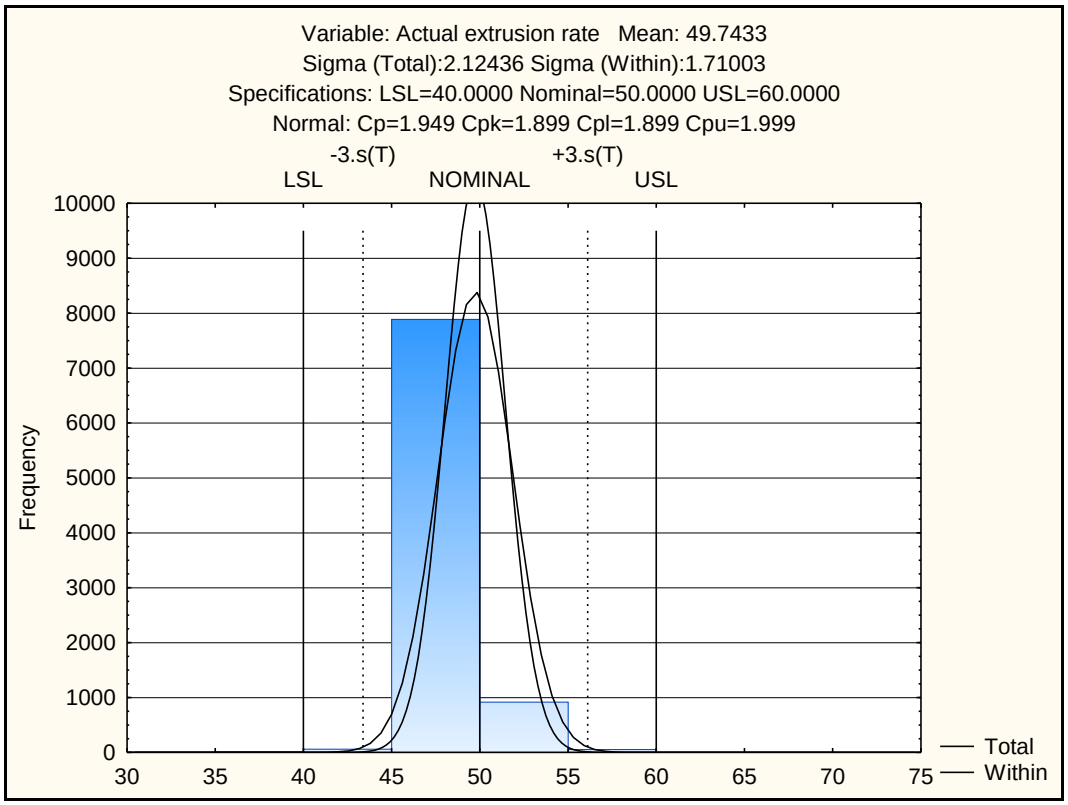
All process capability indices are greater than 1 which provides a process improvement opportunity by reducing specifications or adjusting the process mean for actual tamp pressure towards a higher or lower operating level. This

variable compared to all independent variable capability performance demonstrates the largest process improvement opportunity or room for process improvement.

The distribution is not normal like and skewed to the right, showing high value fliers that have probably no influence on the process seeing that the process operates well within the specification limits.

The DOE analysis should indicate on which operating level, measured in cost of producing out of specification products, this variable should operate in order to optimize product quality.

At this stage, we only measure the overall operating level of 5.891 from Graph 5.5, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.6: Actual extrusion rate: Process capability chart

Graph 5.6 shows a C_p index of 1.949 (the ratio of the specification range over the process range) is greater than 1 indicating a very capable process. C_{pk} index of 1.899 (demonstrated excellence) is greater than 1 indicating a capable process when the actual extrusion rate process is centred. C_{pl} index of 1.899 (lower capability index) and

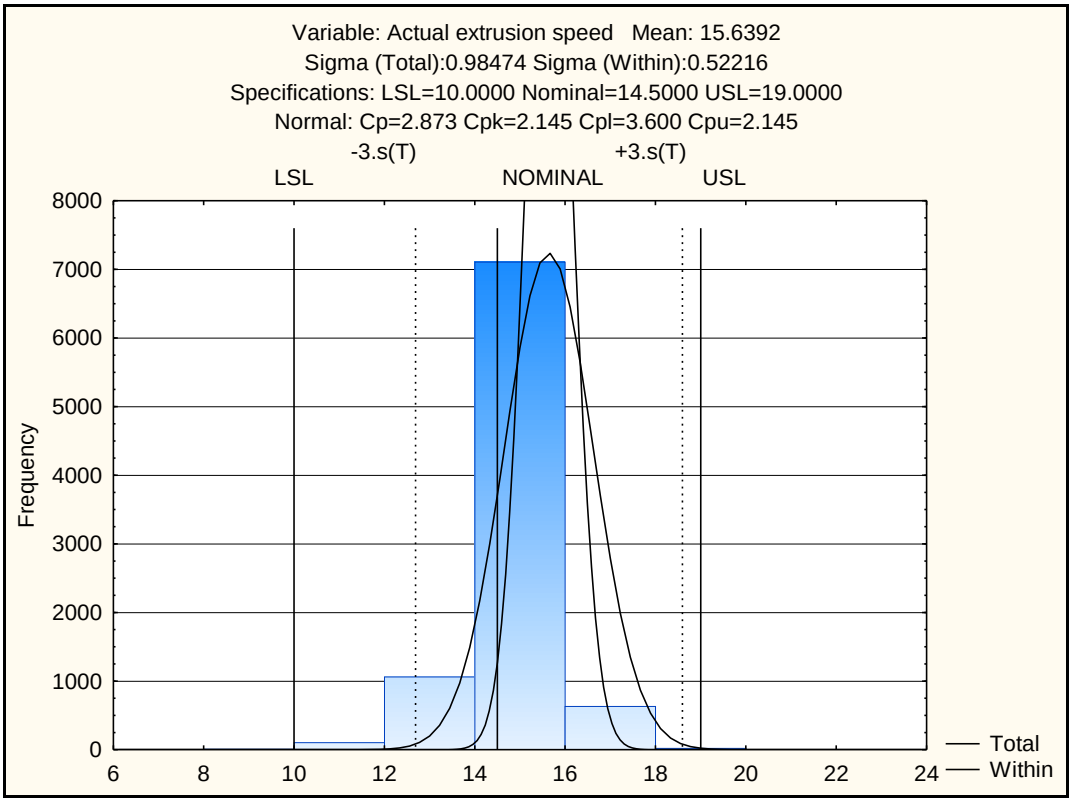
C_{pu} index of 1.999 (upper capability index) are both greater than 1, but because these two indices are different it shows that the process is not centred.

All process capability indices are greater than 1 which provides a process improvement opportunity by reducing specifications or adjusting the process mean for actual extrusion rate towards a higher or lower operating level.

The distribution is not normal like and skewed to the right showing a few high value fliers that probably have no influence on the process.

The DOE analysis should indicate which operating level measured in cost of producing out of specification products, this variable should operate on to optimize product quality.

At this stage, we only measure the overall operating level of 49.743 from Graph 5.6 and not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.7: Actual extrusion speed: Process capability chart

Graph 5.7 shows a C_p index of 2.873 (the ratio of the specification range over the process range) is greater than 1 indicating a very capable process. C_{pk} index of 2.145 (demonstrated excellence) is greater than 1 indicating a capable process when the

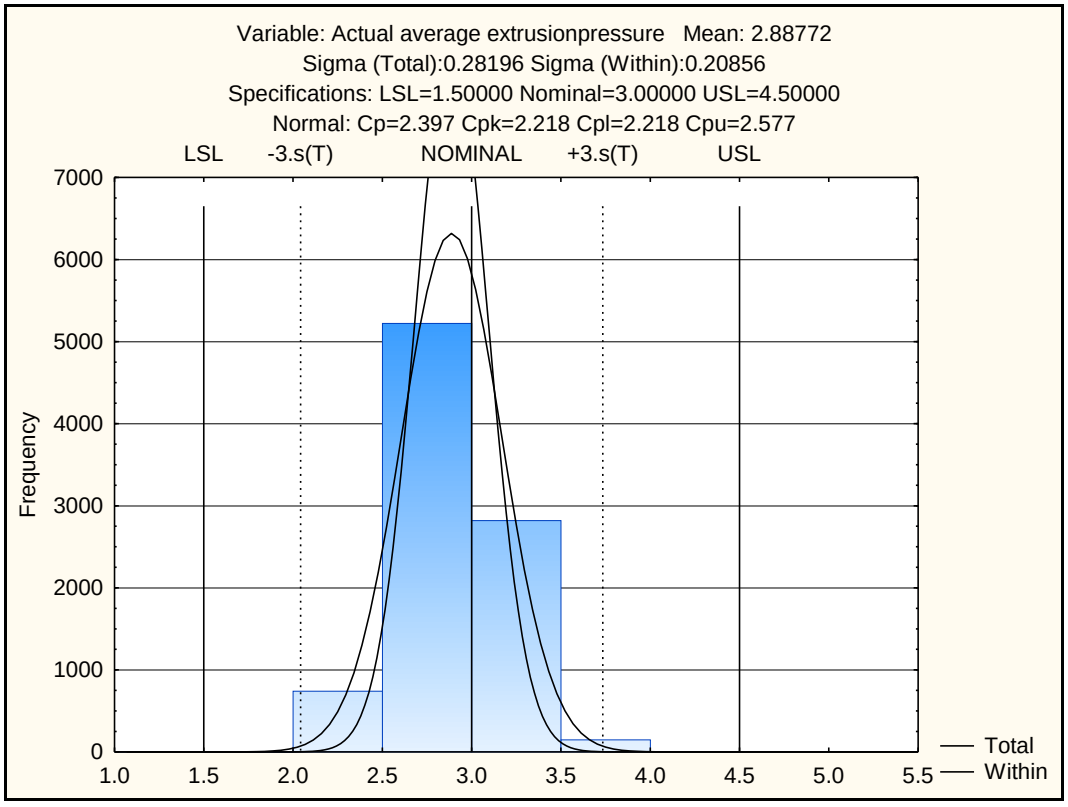
actual extrusion speed process is centred. C_{pl} index of 3.600 (lower capability index) and C_{pu} index of 2.145 (upper capability index) are both greater than 1, but because these two indices are different it shows that the process is not centred.

All process capability indices are greater than 1 which provides a process improvement opportunity by reducing specifications or adjusting the process mean for actual extrusion speed towards a higher or lower operating level.

The distribution is not normal like and skewed to the left, showing a few low value fliers that probably have no influence on the process.

The DOE analysis should indicate on which operating level, measured in cost of producing out of specification products, this variable should operate in order to optimize product quality.

At this stage, we only measure the overall operating level of 15.6392 from Graph 5.7, not validating the opportunity of adjusting the processing mean within the allowable experimental area representing the area between the specification parameters.



Graph 5.8: Actual average extrusion pressure: Process capability chart

Graph 5.8 shows a C_p index of 2.397 (the ratio of the specification range over the process range) is greater than 1 indicating a very capable process. C_{pk} index of 2.2218 (demonstrated excellence) is greaNter than 1 indicating a capable process when the actual average extrusion pressure process is centred. C_{pl} index of 2.2218 (lower capability index) and C_{pu} index of 2.577 (upper capability index) are both greater than 1, but because these two indices are different, it shows that the process is not centred.

A pragmatic independent variable selection criterion based on Cpk to identify which variable is most viable to provide room for process improvement is a variable with a Cpk > 2. Once a capability index reaches a Cpk > 2, the room for process improvement is fairly obvious but still needs a secondary analysis based on SPC to evaluate process variation over time. Combining Cpk with SPC gives a more objective method for variable selection for process improvement.

Although the process output is not centred, it operates well within specifications. For this reason, room for improvement is available for all independent variables, seeing that even with extreme variation of independent variables the dependent output variable stays well within specification. Only two variables, number 5 and 7, because of their high capability index (Cpk > 2.0), seem to provide the biggest opportunity for process improvement.

Number	Independent variable	Cpk	Selection (Y/N) Cpk > 2
1	Mix discharge temperature	1.740	N
2	Cool beginning temperature	1.478	N
3	Actual cooling time	1.105	N
4	Actual dump temperature	1.683	N
5	**Actual tamp pressure**	**2.494**	**Y**
6	Actual extrusion rate	1.899	N
7	**Actual extrusion speed**	**2.145**	**Y**

Table 5.1: Independent variable selection for process improvement

Because all of the independent process capability indices are greater than 1, the SPC analysis should confirm that no effect to product quality is evident. The process capability for the dependent variable above confirms that overall, process improvement

opportunity is possible by reducing the specifications **or** adjusting the process means for independent variables towards a higher or lower operating level, which may reduce the cost impact for deviating from the target.

5.4 SPC ANALYSIS

When analysing individual time-based data through distribution analysis alone, the process variation over time is not realised. The process may have normal like distribution but does not show cycles, shifts and trends within the process. For this reason, when combining distribution analysis with SPC a deeper understanding of the process is gained. SPC analysis for this section graphically represents the process variability on a time scale compared to the sample distribution for selected variables.

When analysing the x-bar and the associated s-chart we expect a normal like distribution for the x-bar chart and for the sample standard deviations.

The remaining seventeen (17) independent and dependent variables in phase 5 shown in chapter 4, Table 4.1, were a combination of continuous, qualitative and categorical variables. They were also evaluated for possible further reduction through distribution and SPC plots. After the evaluation process, seven (7) independent variables and one (1) dependent variable were the final selection.

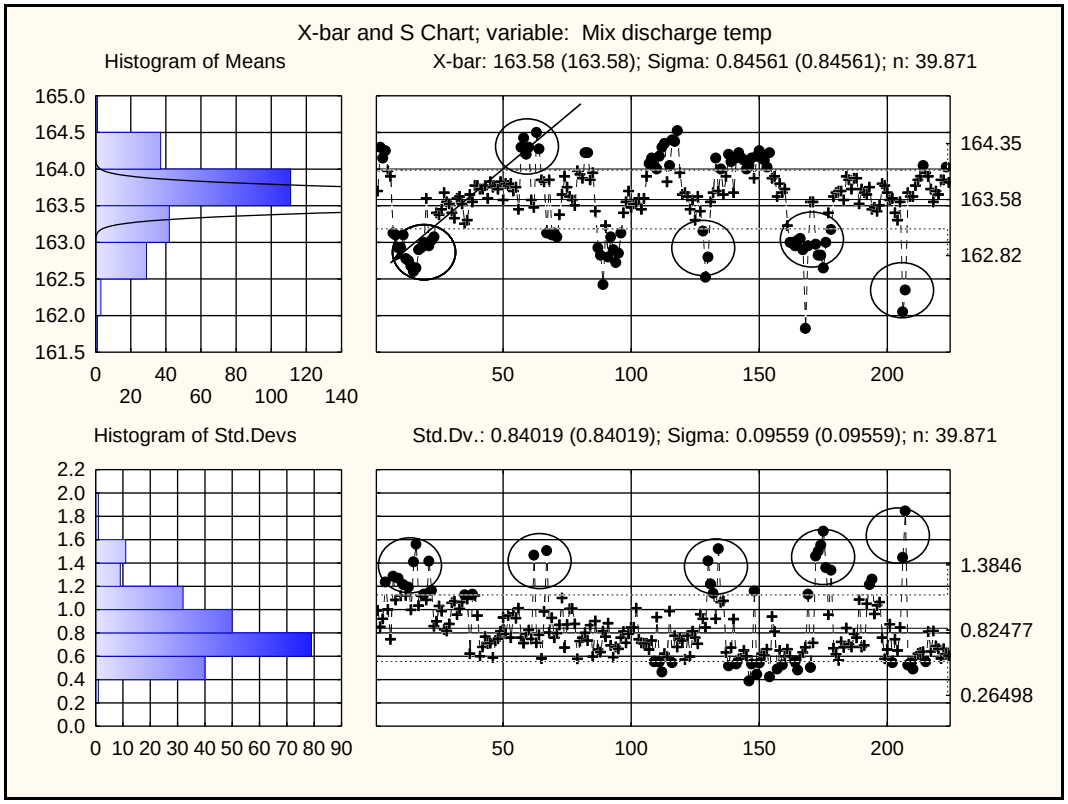
seven (7) independent variables were eliminated (mechanical set points with no bearing to the experiment)

seven (7) independent variables selected for analysis

three (3) dependent variables of which only one (1) was selected.

Individual distribution fitting for the seventeen (three dependent and fourteen independent) variables was done based on SPC, selecting the x-bar and s chart option. The reason for selecting these charts is that it shows the distribution variation between samples and within the samples. An advantage of this chart is that it shows the sample distribution in the form of a histogram in combination with the associated sample variation on a time scale. It is a simple way to understand variation within a histogram that is not always obvious and therefore may lead to incorrect inferences about the measured variable.

X-bar and s chart graphs were drawn for each of the seventeen variables to evaluate which of these variables should remain as the final variables to represent this study going forward. In addition to the SPC charts, scatter plots of the sample average and sample standard deviation from the SPC charts were drawn for each of the seven selected independent variables to evaluate if relationships exist between sample average and sample standard deviation. The SPC graph and scatter plots are as follows:

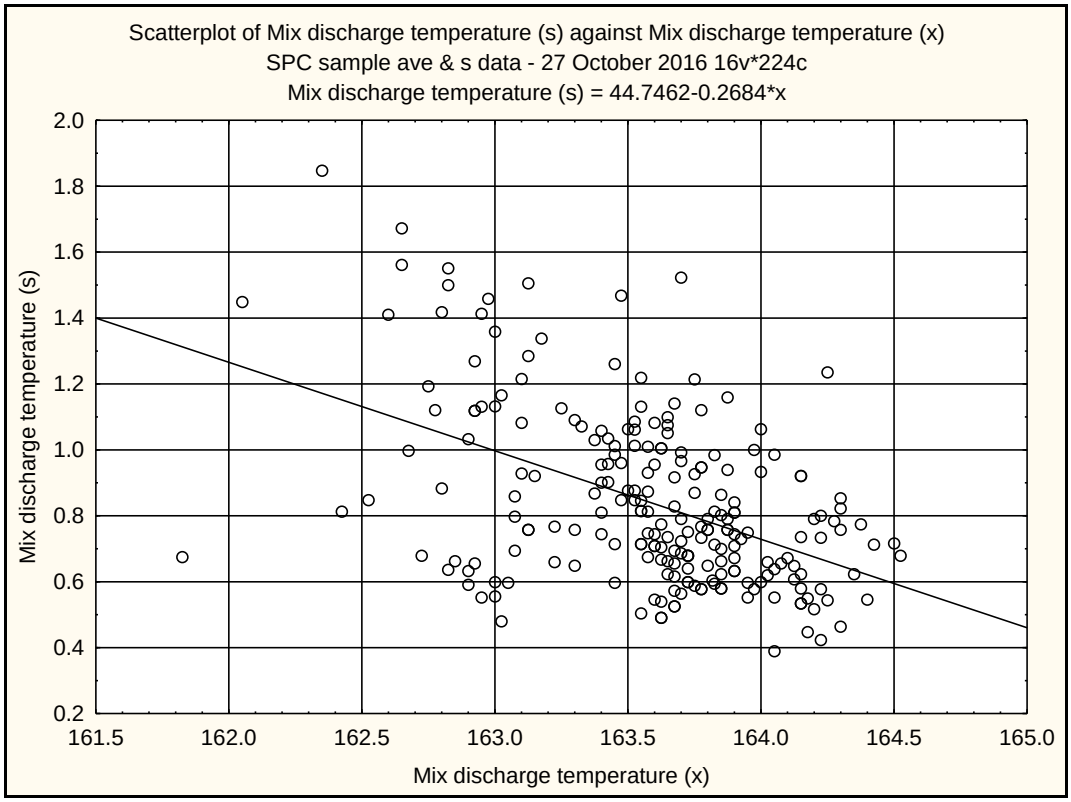


Graph 5.9: Mix discharge temperature: x-bar & s chart

The first independent variable is a critical independent variable for keeping the product from collapsing during the extrusion process. If the product is too hot, it will collapse, and if too cold, will affect the product density. Both conditions will cause scrap.

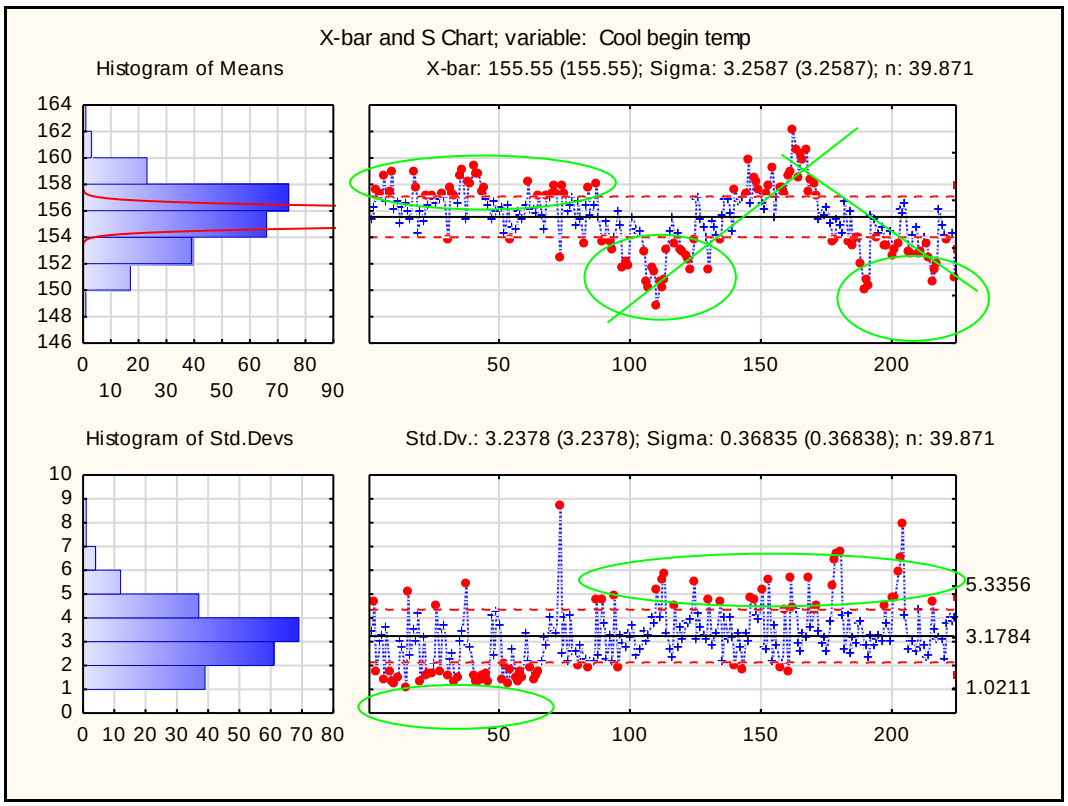
The x-bar sample variation shows a noticeable cyclical pattern that is expected because of the relationship with outside temperature variations as well as out of control points for both x-bar and s charts. These patterns are not desirable but because the temperature variation is well within the temperature specifications of 158 – 168 ºC, no negative effect on product performance will be detected.

However, if only the distribution of the x-bar is observed, the process variation is normal like apart from a few low fliers that still vary within process specification. The standard deviation within samples is skewed to the right which is negative, seeing that the distribution should also be normal like.



Graph 5.10: Mix discharge temperature: Sample (x-bar Vs s)

Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a noticeable negative relationship exists between these two parameters, see Graph 5.10. It shows that, as the average discharge temperature increases, the within sample standard deviation reduces. From a process capability perspective, a Cpk of 1.74 for this variable indicates that room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the upper specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.

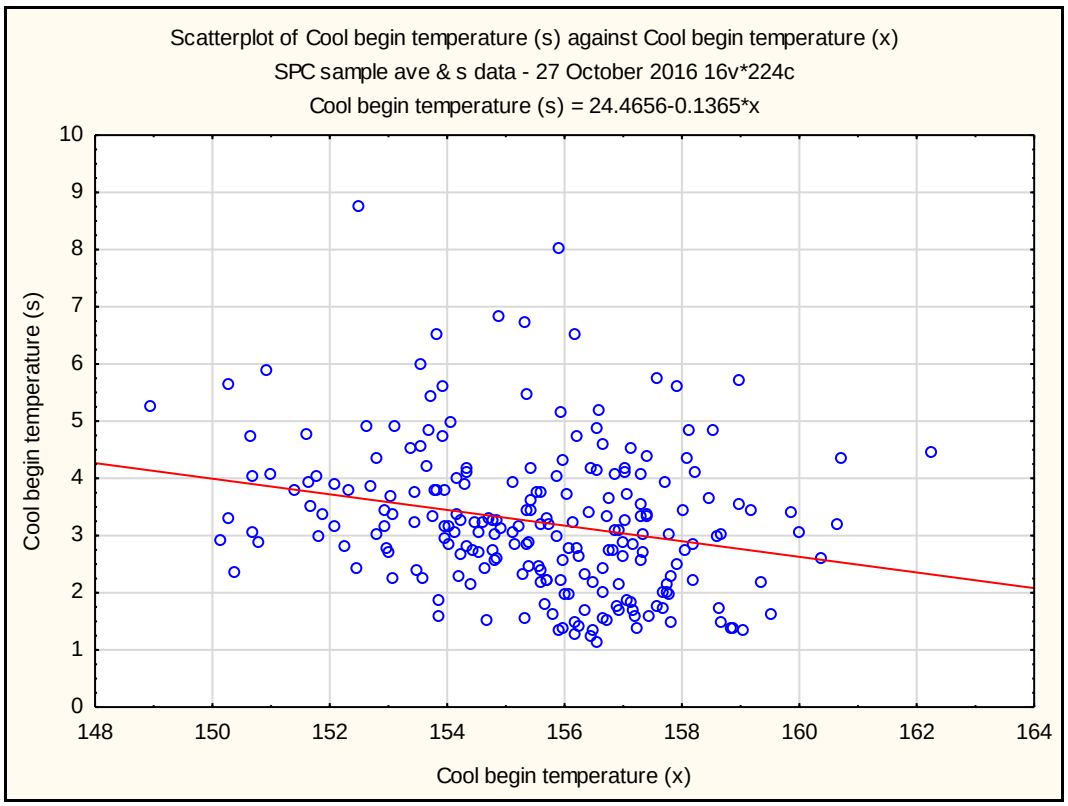


Graph 5.11: Cool begin temperature: x-bar & s chart

The second independent variable is critical for ensuring that the cooling cycle starts correctly at the correct temperature. If the raw material is too hot, the cooling cycle will be long, and when too cold, the cooling cycle will be short. Cooling cycle variation has a direct influence on further product process ability.

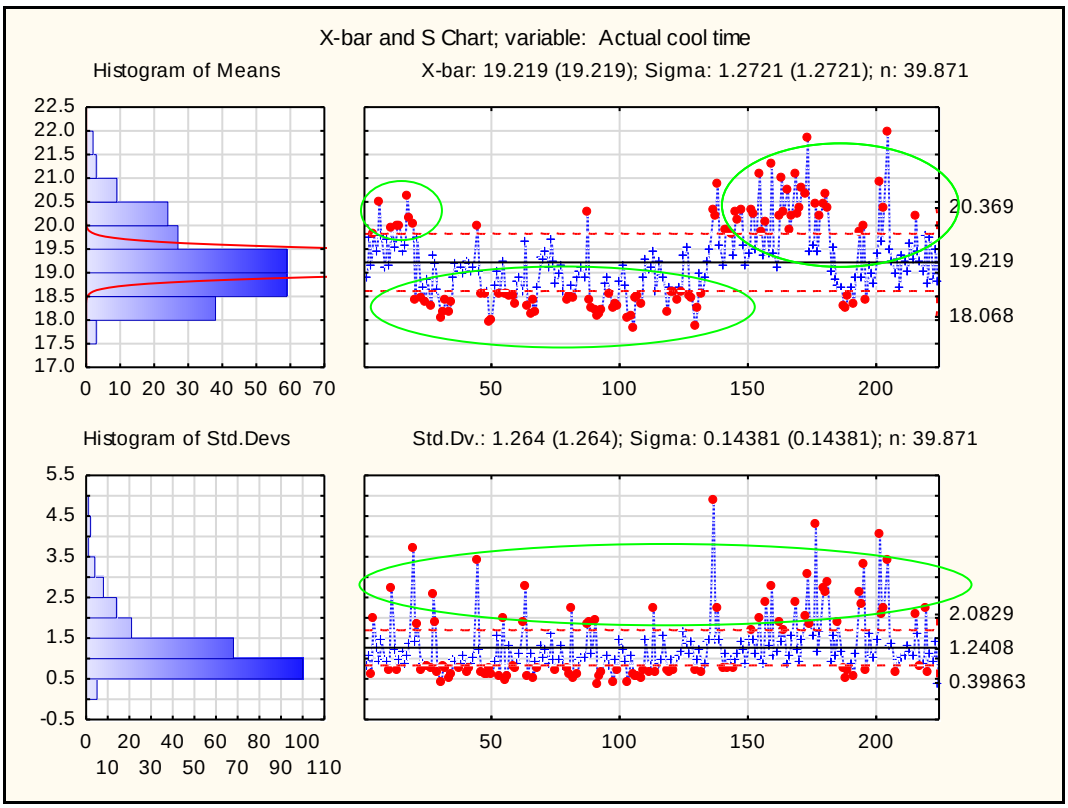
The distribution of the x-bar also seems to be normal like, apart from a few low and high fliers as well as a noticeable cyclical pattern in the middle portion of the graph, which was contributed to equipment issues, but was fixed. There was no negative effect on the final product during this period. These patterns are not desirable, but occur because the temperature variation is well within the temperature specifications of 140 – 170 ºC. No negative effect on product performance will be detected.

The sample standard deviation distribution within samples is skewed to the right, which is negative and should also be normal like.



Graph 5.12: Cool begin temperature: Sample (x-bar Vs s)

Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a noticeable negative relationship exists between these two parameters, see Graph 5.12. It shows that, as the average cool begin temperature increases, the within sample standard deviation reduces. From a process capability perspective, a Cpk of 1.478 for this variable indicates that some room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the upper specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.

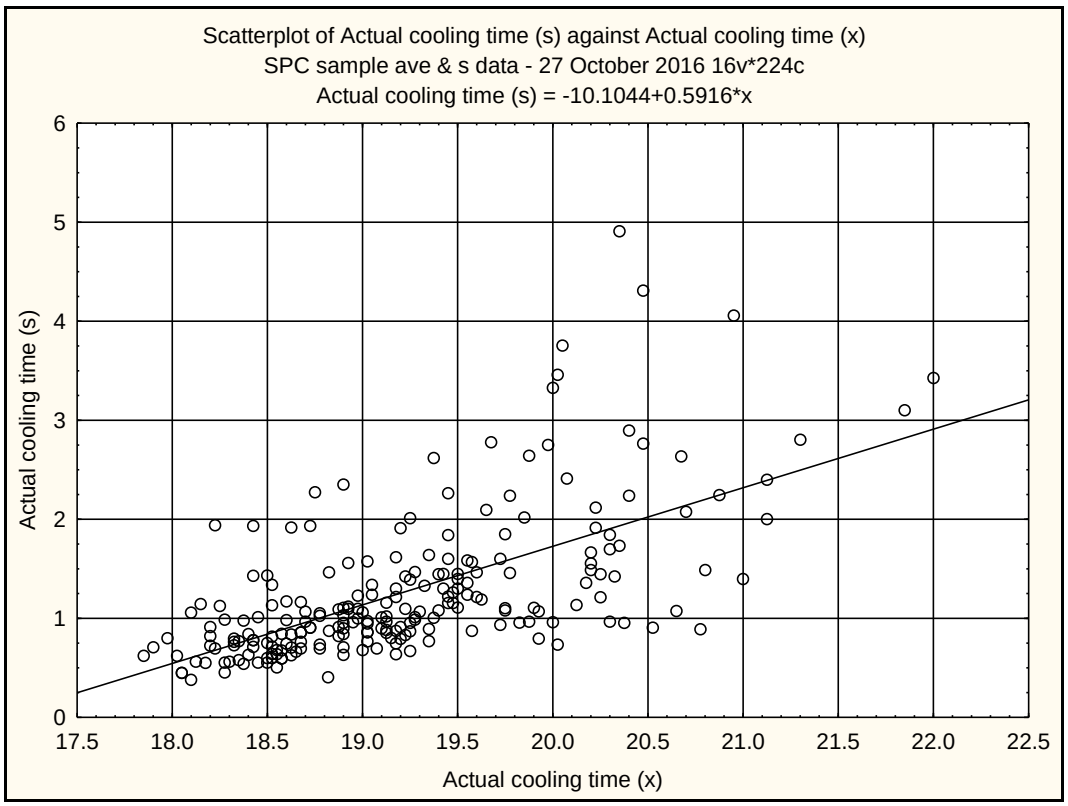


Graph 5.13: Actual cooling time: x-bar & s chart

The third independent variable is critical for ensuring a consistent cooling cycle while mixing raw materials. The cooling cycle variation has a direct influence on product internal structural integrity.

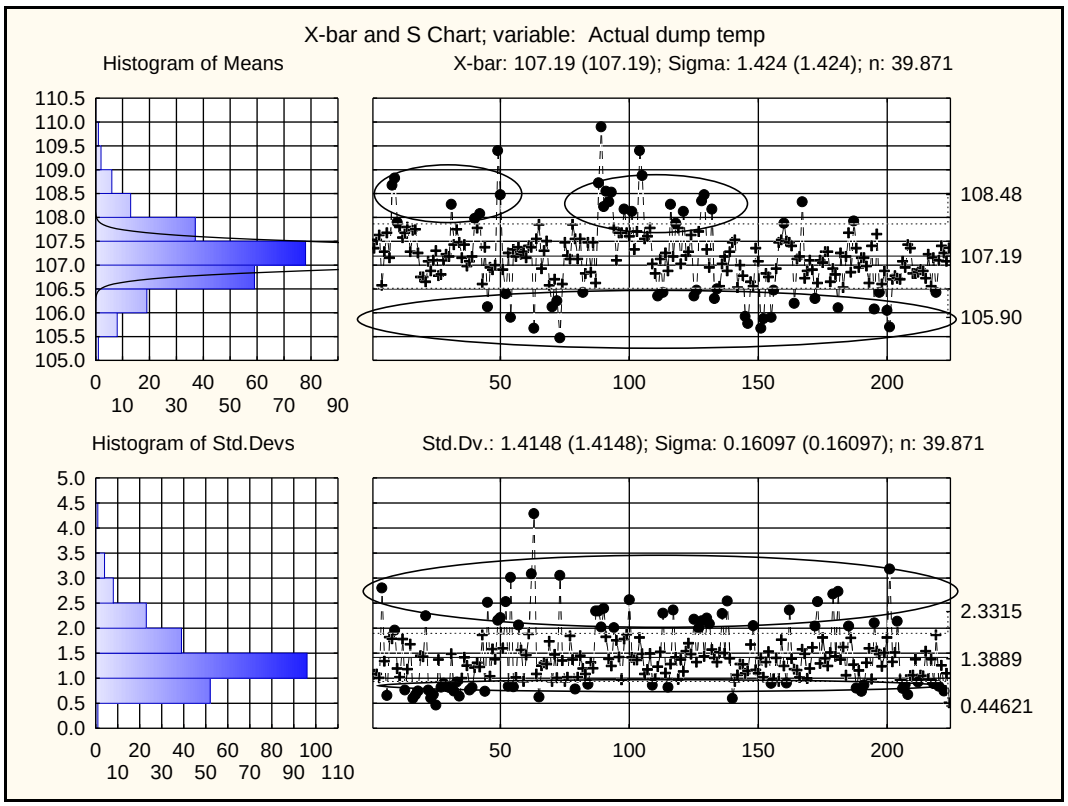
The distribution of the x-bar also seems to be a normal distribution, apart from a few low and high fliers as well as out of control points. A noticeable shift within and towards the end of the time series also occurred. For this particular variable, these shifts coincide with planned maintenance on pyrometers measuring product temperature. However, if only the distribution of the x-bar is observed, an acceptable process spread is evident. There was no negative effect on the final product during this period. This variation is acceptable because the time variation is well within the time specifications of 15 – 27 minutes. No negative effect on product performance will be detected.

The sample standard deviation within samples is skewed to the right, which is negative and should also be normal like.



Graph 5.14: Actual cooling time: Sample (x-bar Vs s)

Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a noticeable positive relationship exists between these two parameters, see Graph 5.14. It shows that, as the cooling time increases, the within sample standard deviation increases. From a process capability perspective, a Cpk of 1.105 for this variable indicates that little room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the lower specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.

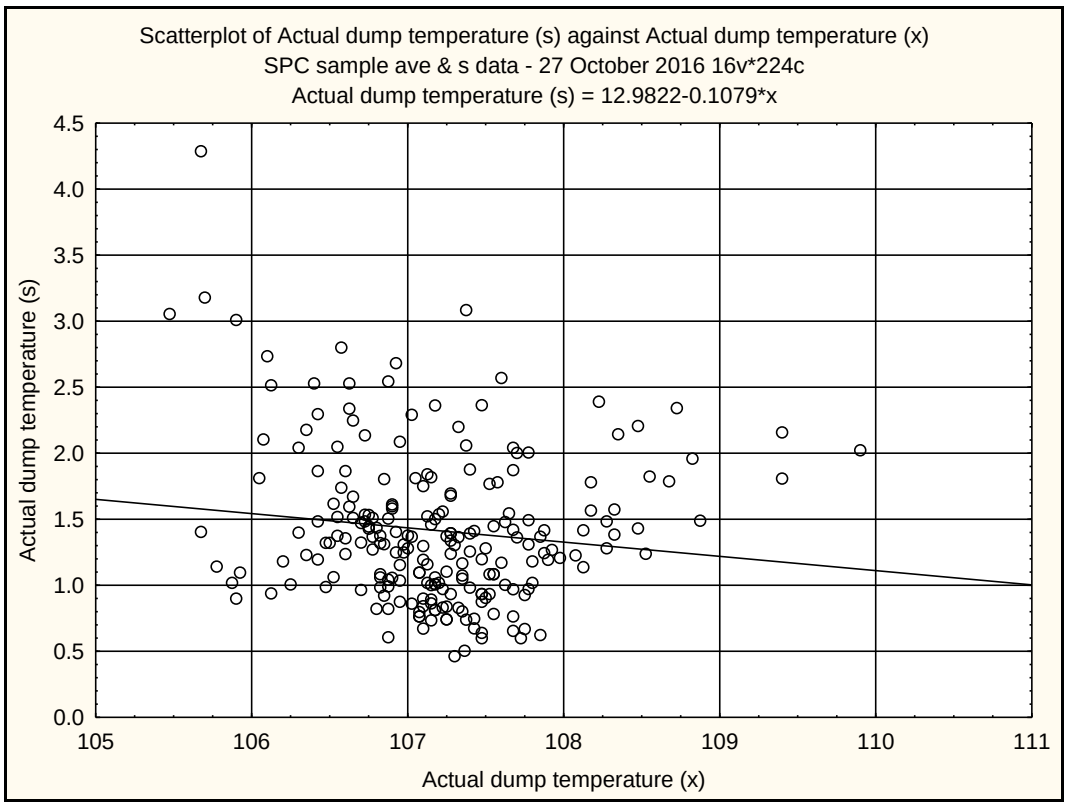


Graph 5.15: Actual dump temperature: x-bar & s chart

The fourth independent variable is critical for ensuring a consistent product dump temperature. The actual mix dump temperature has a direct influence on product internal structural integrity and process ability through the next production process.

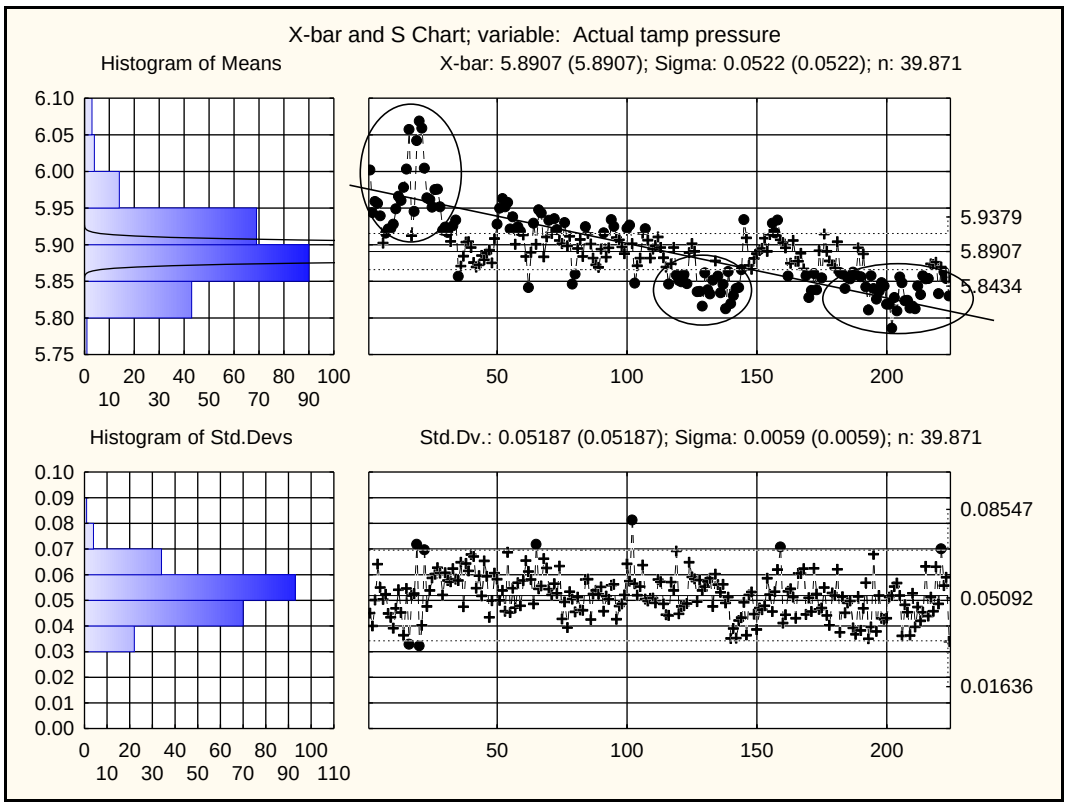
The distribution of the x-bar seems to be normal like or practically normal, apart from a few low and high fliers. The fourth independent variable shows no obvious patterns, but varies normally around the x-bar for the whole period. Any pattern is not desirable but because the temperature variation is well within the temperature specifications of 100 – 110 ºC. No negative effect on product performance is detected.

The standard deviation within samples is skewed to the right, which is negative; it should also be normal like.



Graph 5.16: Actual dump temperature: Sample (x-bar Vs s)

Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a weak negative relationship exists between these two parameters, see Graph 5.16. It shows that as the average dump temperature increases, the within sample standard deviation reduces. From a process capability perspective, a Cpk of 1.683 for this variable indicates that room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to 108 ºC, lower within sample standard deviation occurs that translates into lower process variability. For this variable the within sample standard deviation starts to increase beyond an average dump temperature of 108 ºC. Lower process variability assists in higher process output predictability, which is essential for product consistency.

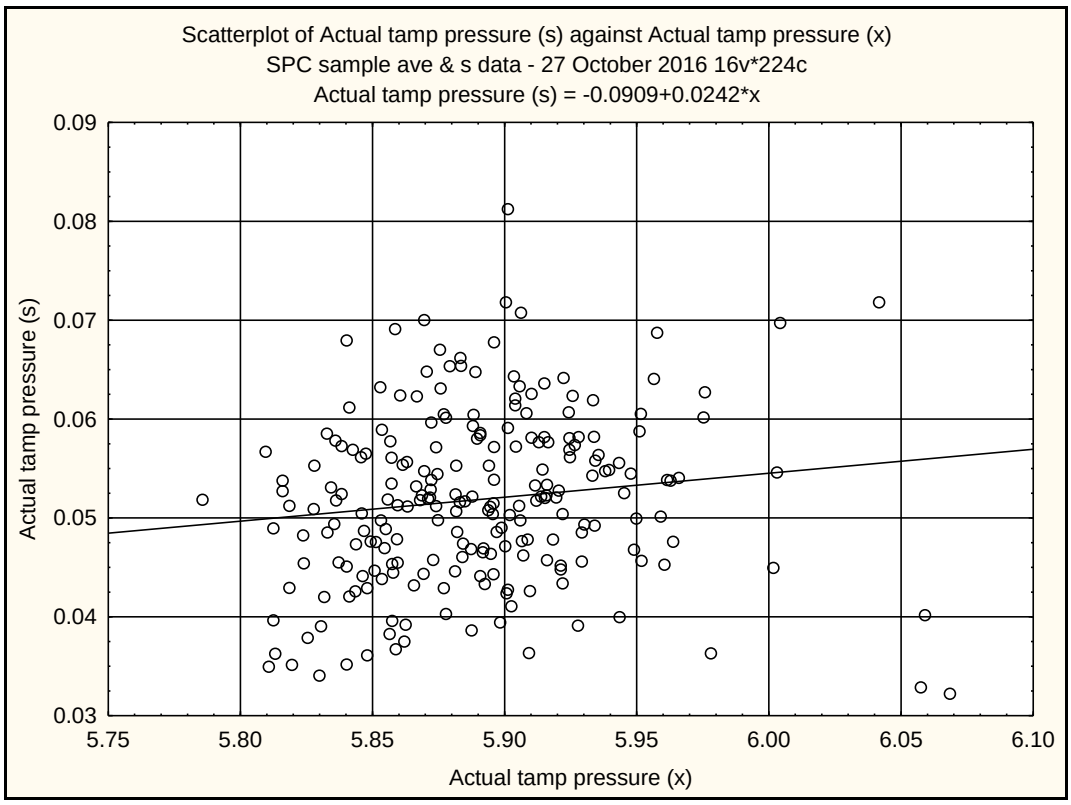


Graph 5.17: Actual tamp pressure: x-bar & s chart

The fifth independent variable is critical for ensuring product internal structural integrity and process ability through the next production process. If not at the correct level, internal cracking may appear during the next process.

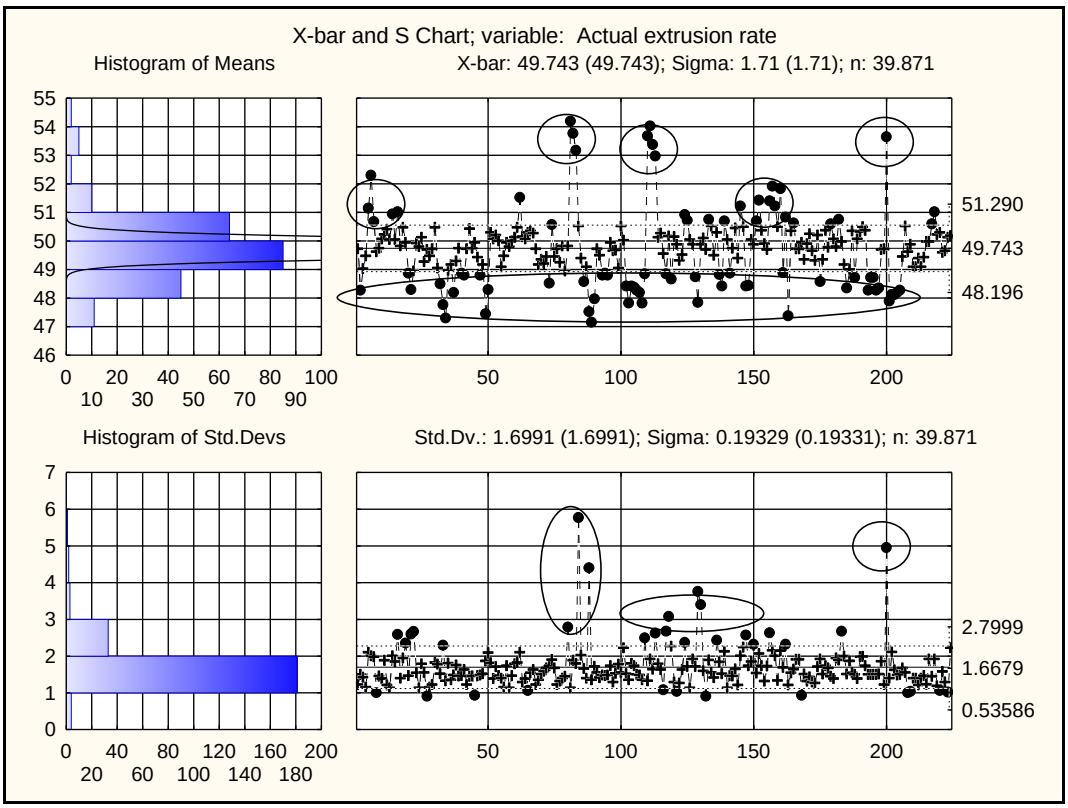
The distribution of the x-bar seems to be normal like and practically normal. The time trend shows a huge negative trend from start to end with high and low fliers, which is concerning but natural for the equipment measuring tamping pressure. Even with the measured trend, these patterns are not desirable but because the actual pressure is well within specification of 5.5 – 6.5 Mpa. No effect on product performance will be detected. Comparing the x-bar and s chart for this example is a typical illustration showing the difference between a distribution and the real data series plotted over time. The one without the other is meaningless.

The standard deviation distribution within samples is normal like as it should be, which is the only independent variable that that satisfies both distribution criteria. This variable was chosen because of its importance to the manufacturing process and not because of its statistical characteristics. Based on Cpk and SPC distribution evaluation, this variable offers the biggest room for process improvement.



Graph 5.18: Actual tamp pressure: Sample (x-bar Vs s)

This variable does have the expected distribution shapes for sample x-bar and standard deviation, and a weak positive relationship exists between these two parameters, see Graph 5.18. It shows that as the average tamp pressure increases, the within sample standard deviation increases. From a process capability perspective, a Cpk of 2.494 for this variable indicates that much room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the lower specification limit, lower within sample standard deviation occurs that translates into lower process variability. This variable is the only variable that complies with the expected distribution shapes, high Cpk, weak relationship and maintaining a fair process variability over time. Lower process variability assists in higher process output predictability, which is essential for product consistency.

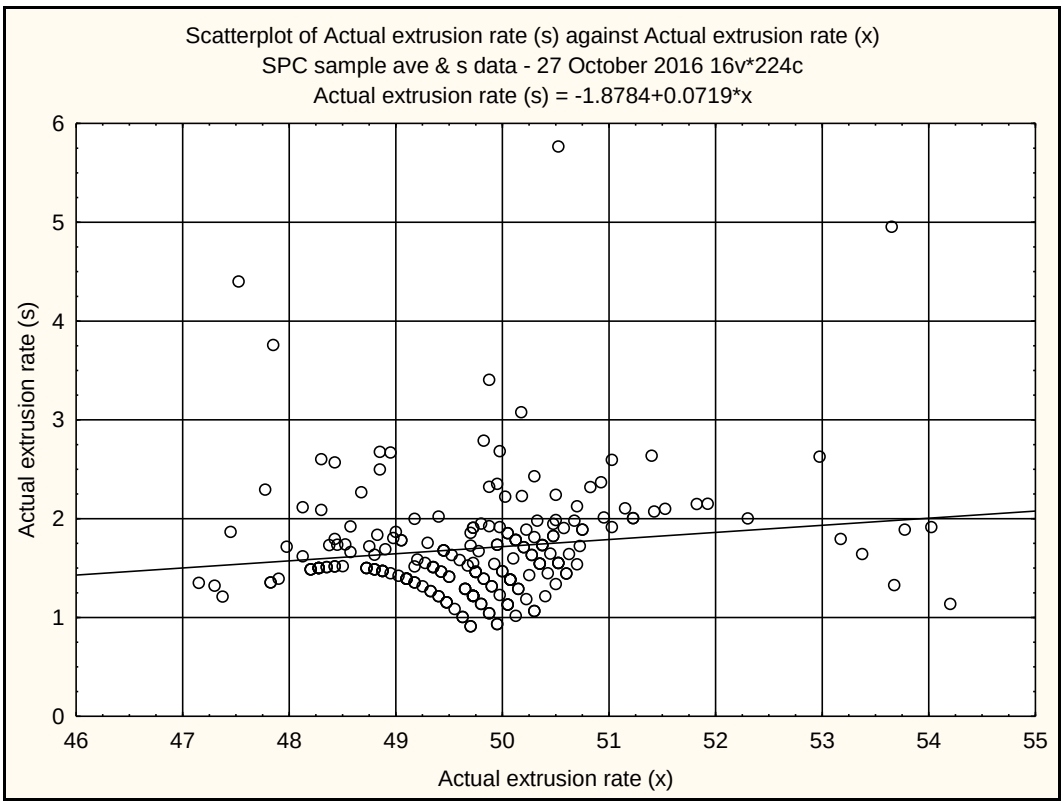


Graph 5.19: Actual extrusion rate: x-bar & s chart

The sixth independent variable represents the extruding pusher rate that ensures a consistent swelling of the product that influences the actual diameter of the product. In later processes, excessive variation in diameter causes capacity problems.

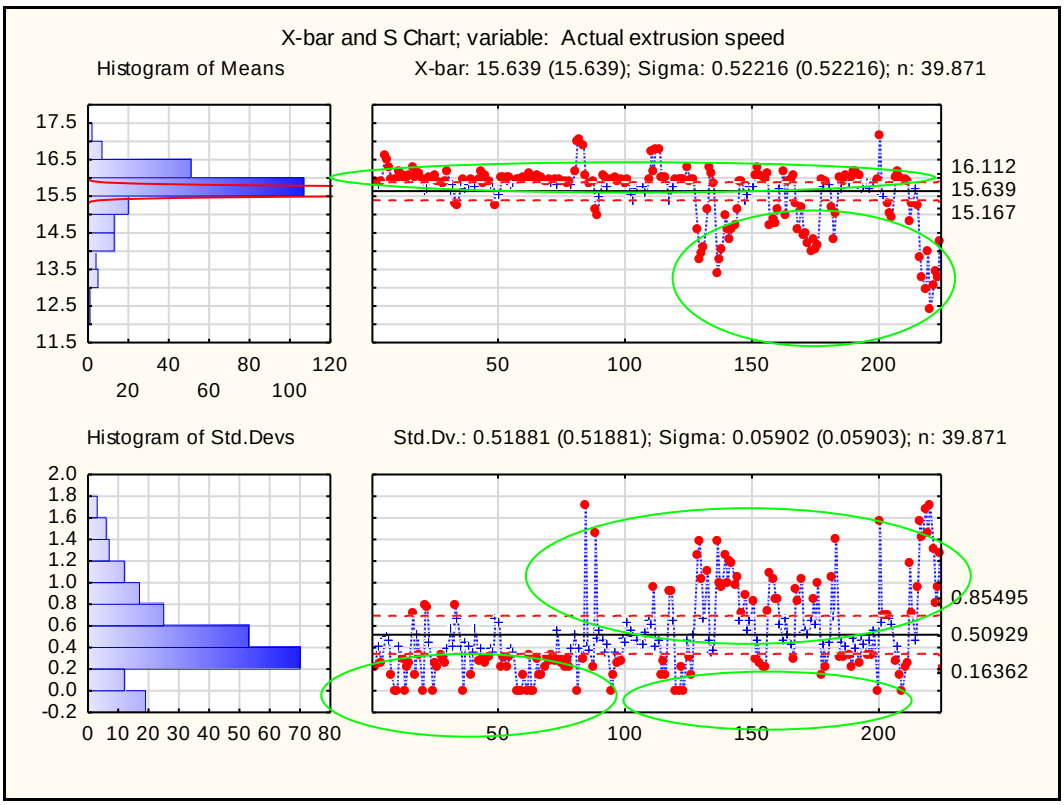
The distribution of the x-bar seems not to be normal like, but skewed to the right. The time trend shows no obvious patterns, but varies normally around the x-bar for the whole period with a few spikes. These patterns are not desirable but because the actual extrusion rate is well within specification of 40 – 60 cm/hr. No negative effect on product performance is detected.

The standard deviation distribution within samples is also skewed to the right, which is negative; it should also be normal like.



Graph 5.20: Actual extrusion rate: Sample (x-bar Vs s)

Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a weak positive relationship exists between these two parameters, see Graph 5.20. It shows that as the average discharge temperature increases the within sample standard deviation increase. From a process capability perspective, a Cpk of 1.899 for this variable indicates that room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the lower specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.

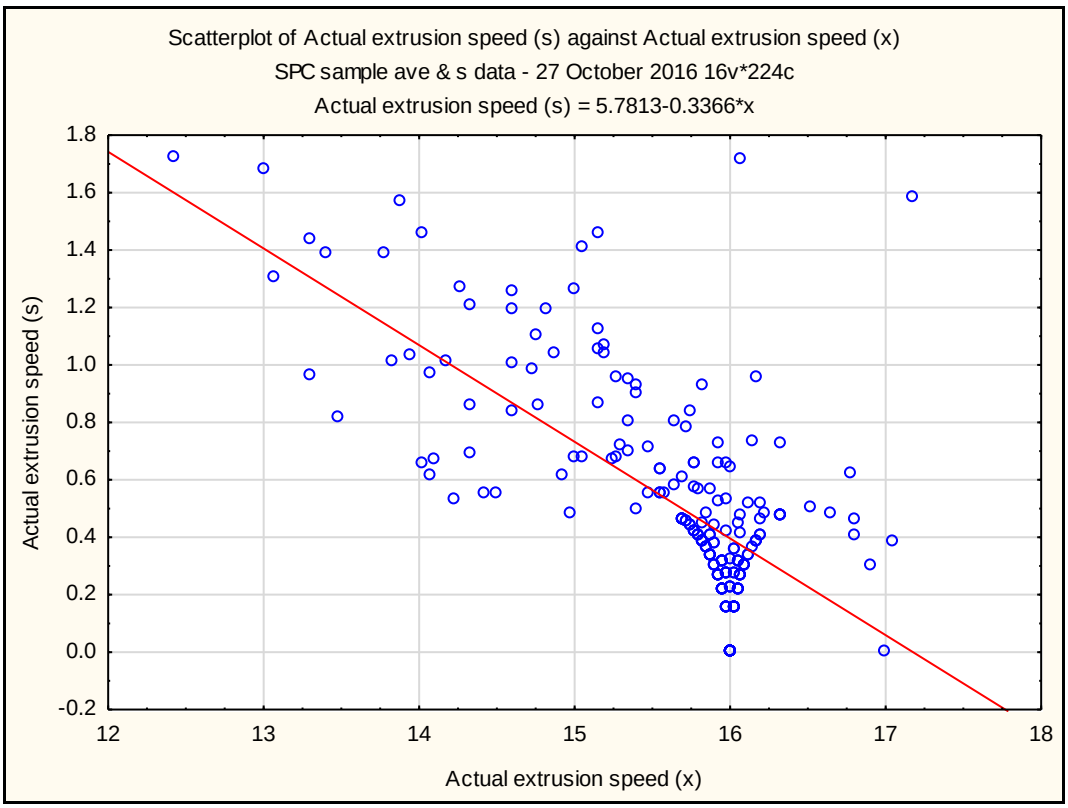


Graph 5.21: Actual extrusion speed: x-bar & s chart

The seventh independent variable measures the measured extruding speed by which the product is extruded, related to the extruding rate.

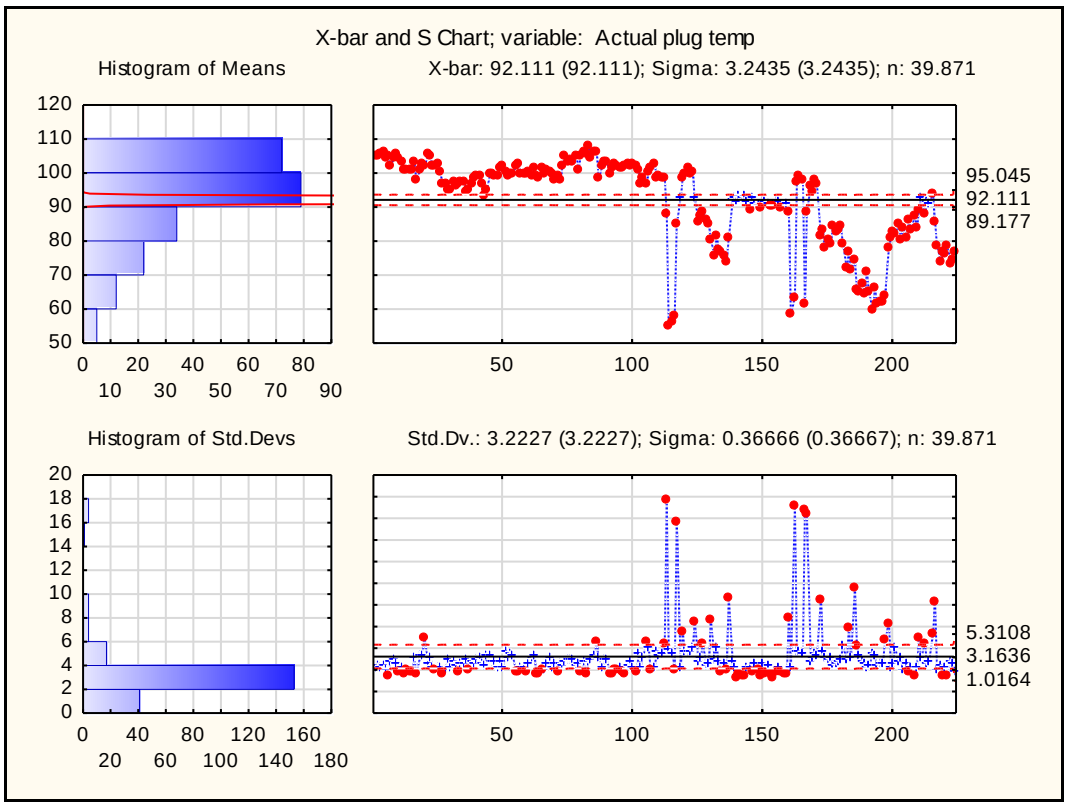
The x- bar chart shows a not normal like variation, but skewed to the left. Data start smooth at the start, then follow an erratic pattern with low fliers towards the end. This erratic pattern towards the end was due to excessive variability within a new raw material. Even for the measured erratic pattern, the actual extrusion speed is well within specification of 10 – 19 metres per hour. No negative effect on product performance is detected.

The standard deviation distribution within samples is skewed to the right, which is negative; it should be normal like as well.

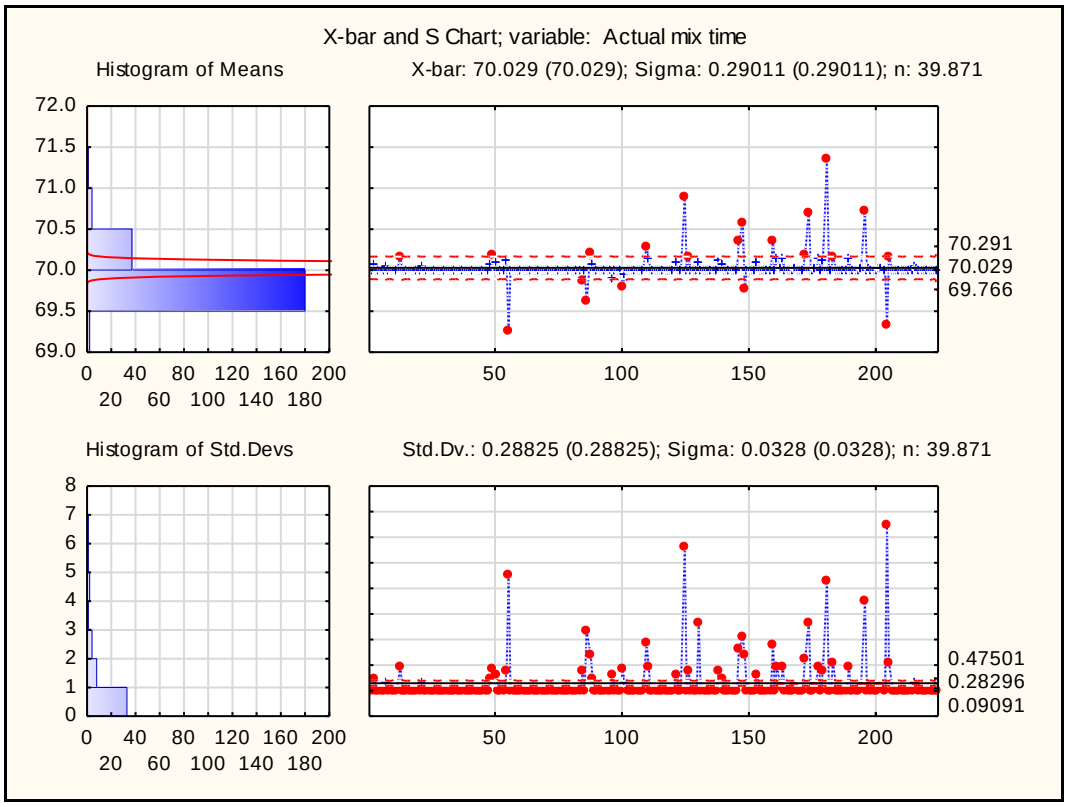


Graph 5.22: Actual extrusion speed: Sample (x-bar Vs s)

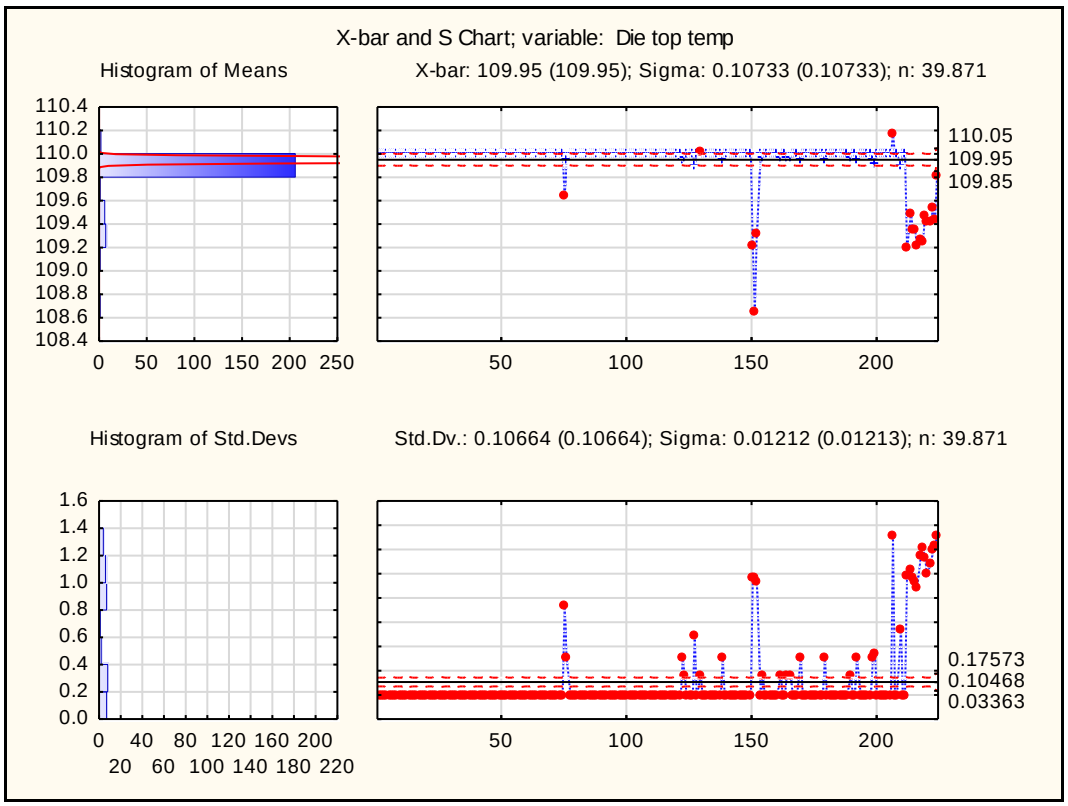
Although this variable does not have the expected distribution shapes for sample x-bar and standard deviation, a strong negative relationship exists between these two parameters, see Graph 5.22. It shows that as the average discharge temperature increases, the within sample standard deviation reduces. From a process capability perspective, a Cpk of 2.145 for this variable indicates that room for process improvement exists in reducing process variability. By shifting the average discharge temperature closer to the upper specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.



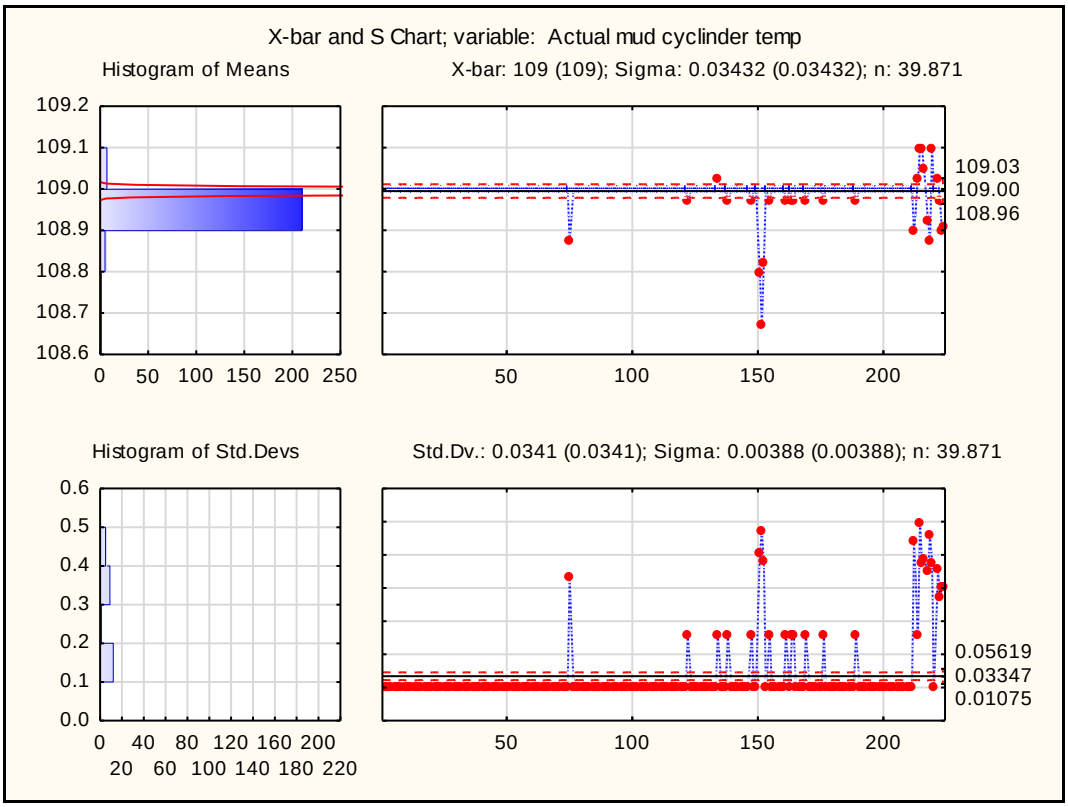
Graph 5.23: Actual plug temperature: x-bar & s chart



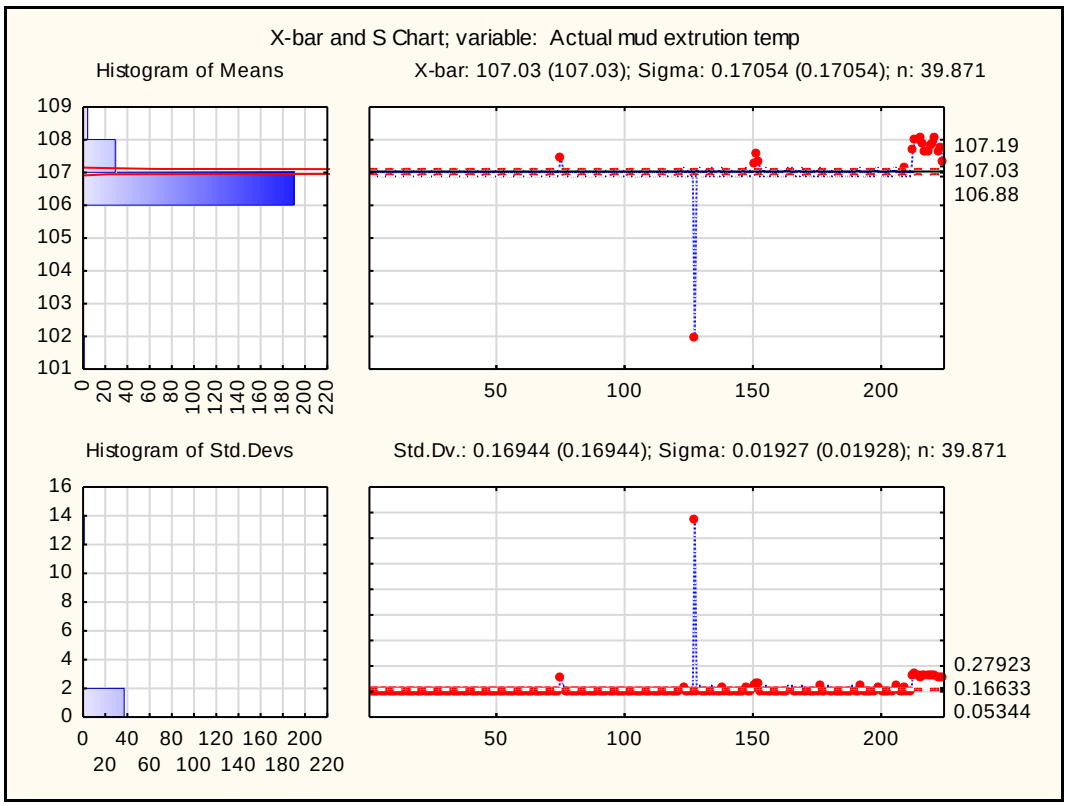
Graph 5.24: Actual mix time: x-bar & s chart



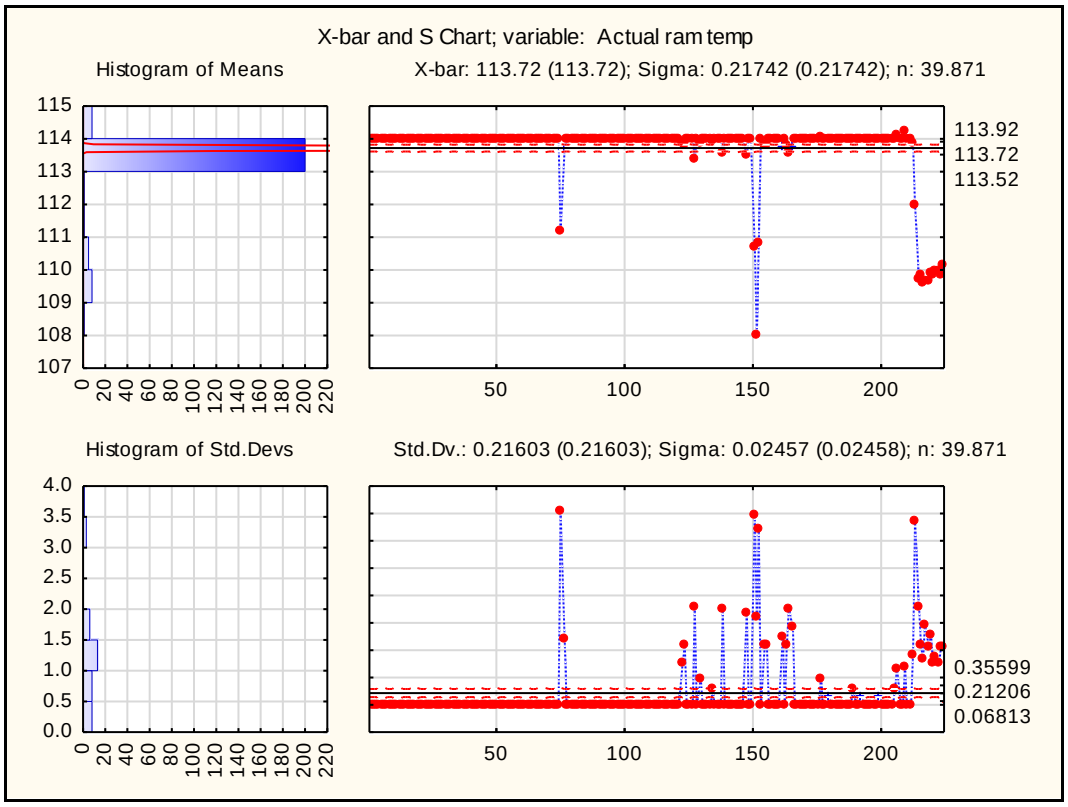
Graph 5.25: Die top temperature: x-bar & s chart



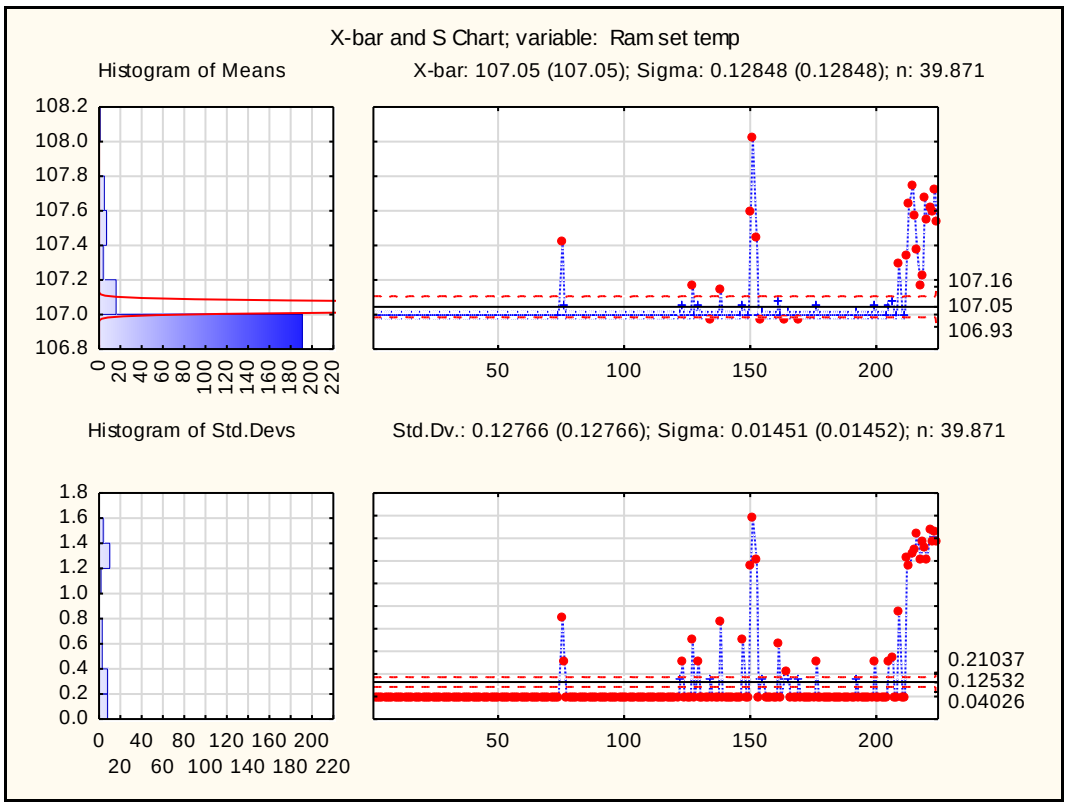
Graph 5.26: Actual mud cylinder temperature: x-bar & s chart



Graph 5.27: Actual mud cylinder extrusion temperature: x-bar & s chart



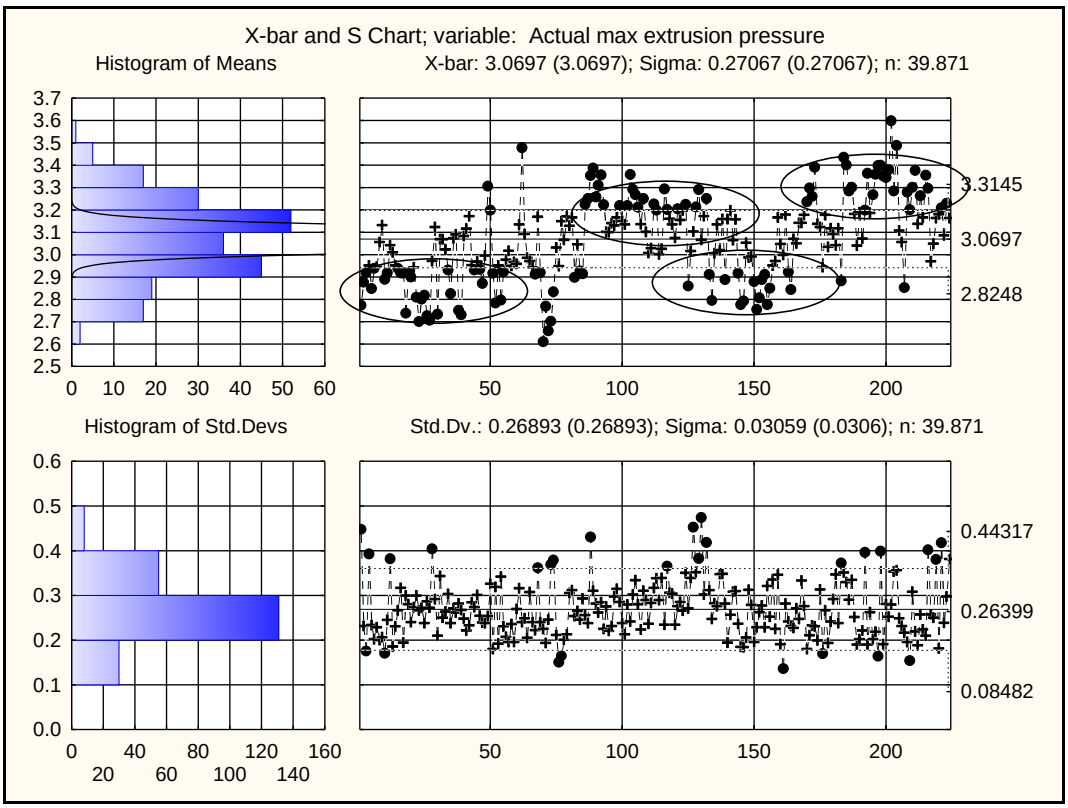
Graph 5.28: Actual ram temperature: x-bar & s chart



Graph 5.29: Ram set temperature: x-bar & s chart

Graphs for Independent variables 5.25 to 5.31 are mechanical set-point variables that cause no or little variation and cannot be changed. This is evident in their respective distributions, which also clearly show very little variation. That was expected.

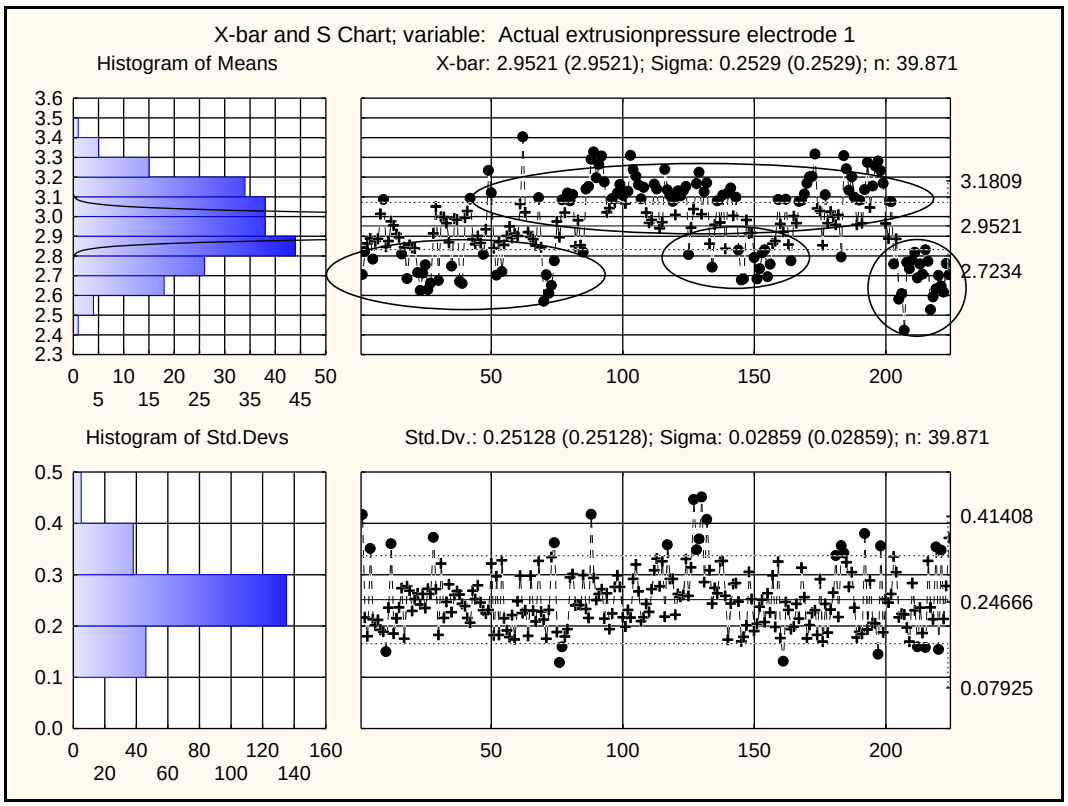
For this study, these independent variables were removed. They have little or no influence on the dependent variables because the respective set points will remain as is and will be changed only when a major impact study is done. A follow-up study is proposed to evaluate the impact of these variables on process development and product quality.



Graph 5.30: Actual max pressure: x-bar & s chart

This is the first of the three dependent variables that evaluate product quality. It represents the maximum pressure of an individual product recorded in a processed batch. This dependent variable shows shifts, cycles, as well as low and high fliers, but the x-bar still varies normal like for the whole period and well within the processing specification of 1.5 – 4.5 Kpa. No defects are detected.

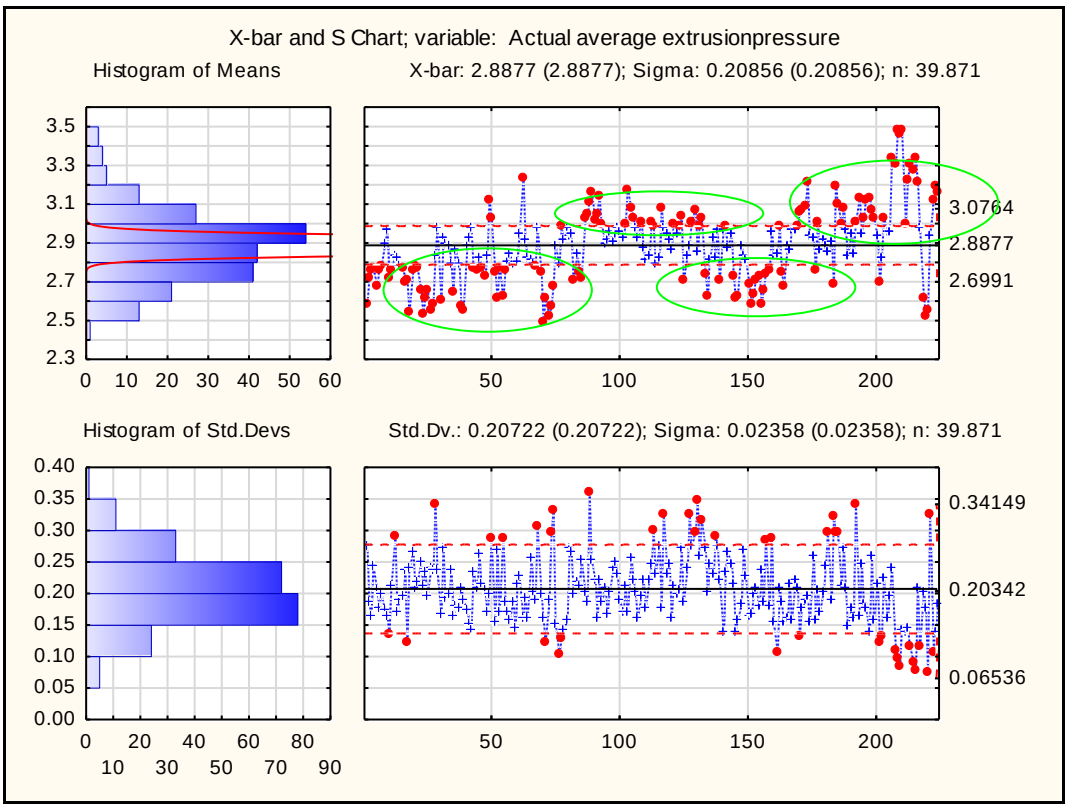
The standard deviation distribution within samples is skewed to the right, which is negative because the distribution should be normal like as well. The selection of this dependent variable is solely because it is a key performance indicator to product quality, but also shows accepted variability given the overall processing environment.



Graph 5.31: Actual extrusion pressure first product: x-bar & s chart

This variable is the second of the three dependent variables that evaluate product quality. It represents the pressure of the first individual product recorded in a processed batch. This dependent variable shows shifts, cycles, as well as low and high fliers, but the x-bar still varies normal like for the whole period and well within the processing specification of 1.5 – 4.5 Kpa. No defect will be detected.

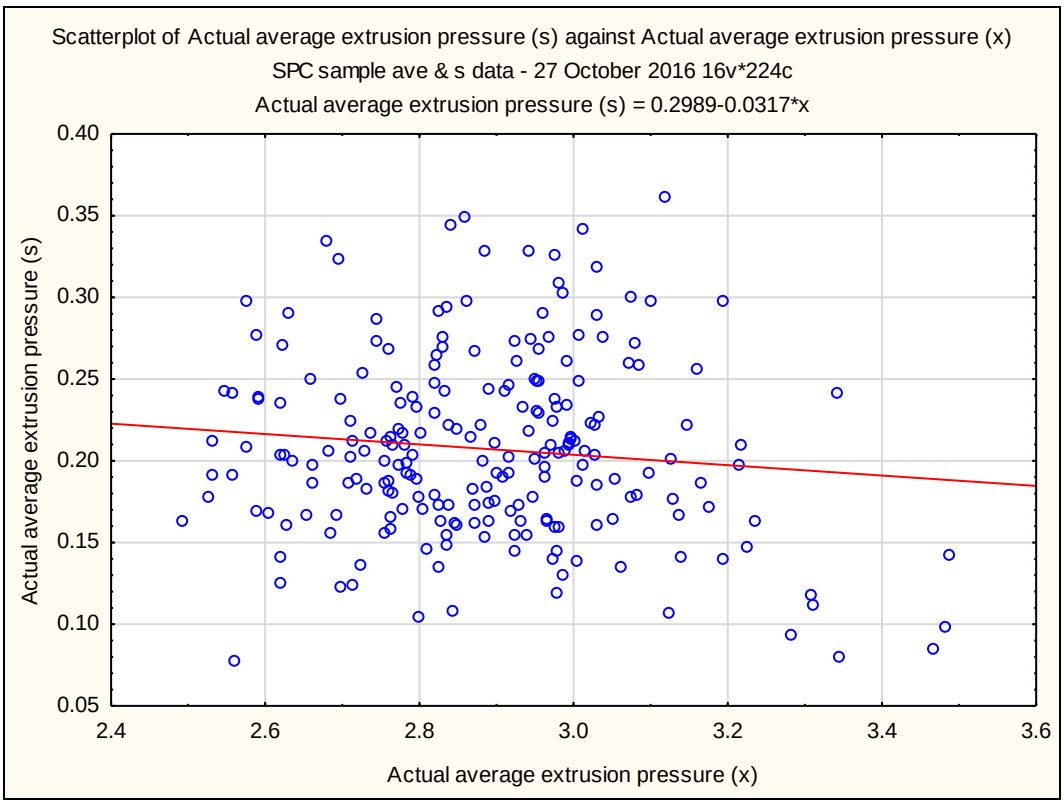
It follows similar patterns to the maximum pressure dependent variable. This could be batch related, irrespective of the outcome variable measured. The standard deviation within sample distribution is also skewed to the right, which is negative because both distributions should be normal like. The selection of this dependent variable is also solely because it is a key performance indicator to product quality.



Graph 5.32: Actual average extrusion pressure: x-bar & s chart

This variable is the third of the three dependent variables that evaluate product quality. It represents the average pressure of the products recorded in a processed batch. This dependent variable shows shifts, cycles, as well as low and high fliers, but the x-bar still varies normal like for the whole period and well within the processing specification of 1.5 – 4.5 Kpa. No defect is detected.

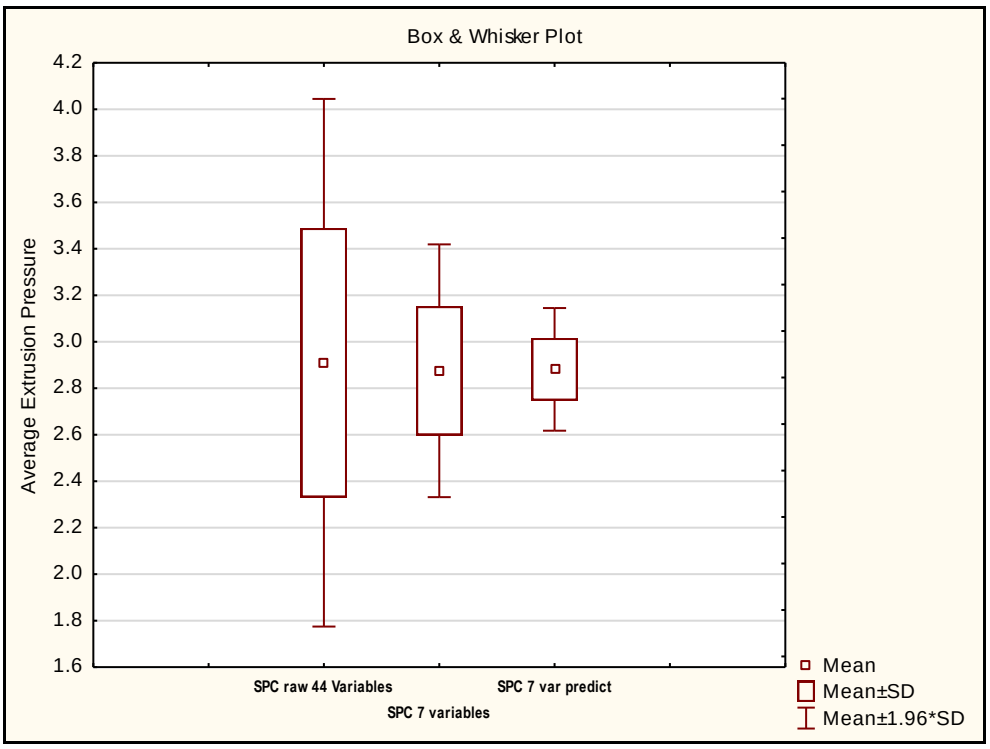
The x-bar distribution variance for this dependent variable is smaller than the previous two dependent variables. The standard deviation within samples is also normal like which is positive. The selection of this dependent variable is also solely because it is a key performance indicator to product quality.



Graph 5.33: Average extrusion pressure: Sample (x-bar Vs s)

This variable is an output variable and does have the expected distribution shapes for sample x-bar and standard deviation; a weak negative relationship exists between these two parameters, see Graph 5.33 It shows that as the average extrusion pressure increases the within sample standard deviation decrease slightly. From a process capability perspective, a Cpk of 2.2218 for this variable indicates that room for process improvement exists in reducing process output variability. By shifting the average extrusion pressure closer to the upper specification limit, lower within sample standard deviation occurs that translates into lower process variability. Lower process variability assists in higher process output predictability, which is essential for product consistency.

Because of the low variability, close relationship to the actual process variation, the normal like distributions for both x-bar and s charts this dependent variable will be used for all analysis going forward. There will be no scientific benefit to include the remaining two dependent variables for now. The purpose is not to compare the effects on dependent variables but only to show an analytical process.



Graph 5.34: Actual average extrusion pressure: Box & Whisker plot

Graph 5.34 represents three periods of variable reduction namely; Period 1 (Original raw database of 44 variables and 15400 records), repeatable period (Final 7 DOE selected variables and 9500 records) and validation period (DOE regression predicted values using the final 7 variables and 9500 records). The effect of output variability reduction as the quantity of variables and records reduces through the variable reduction process by SPC is evident. The average value is fairly constant at 2.9 amongst the three periods, but the unit variability spread around the mean reduces from 2.2 for period 1, to 1.0 for validation period and 0.4 for period 3, which is a significant reduction.

From a predictability point of view, this variability reduction is significant for prediction accuracy, but from an experimental region angle, it shows exploration opportunities represented by the eliminated variables not being analysed yet. As discussed in chapter 6 DOE exploration opportunities outside the experimental region should be part of future work.

5.5 SUMMARY

From a SPC perspective, none of the independent or dependent variables in statistical control showing many cyclical patterns, shits, trends, high and low fliers. Even though

these patterns exist, specification limits for each of these variables are wider than the three-sigma variation limits used for SPC. For this reason the opportunity for process improvement does exist for all variables, with independent variable 5 showing the most room for improvement. From a process capability perspective, see Table 5.1, all capability indices were greater than 1 but only two variables, actual tamp pressure and actual extrusion speed, show a Cpk > 2, which provides the best process improvement opportunities by reducing specifications or adjusting the process mean towards a higher or lower operating level.

In addition to SPC and capability analysis, the normality assumption that sample averages and sample standard deviations should be approximately normally distributed were tested, scatter plots were constructed for each independent variable, showing the relationship between sample average and sample standard deviation. The expectation was that no trends should be evident, only random scattered data points. Most variables have noticeable trends, with only two fairly randomly distributed.

Only actual tamp pressure shows a normal like distribution for both sample average and sample standard deviation distributions, a high CPk and close to random scattered data points for scatter plots representing sample average compared to sample standard deviation. This variable may be the “red X” at this stage but should be validated in the regression analysis chapter.

Automatically disqualifying outliers as bad data points without validating them against the impact on processing output could be a mistake and could be the unexpected discovery of valuable information. Always treat non-normal data points as part of natural variation, and they should be analysed accordingly. For this study, because outliers were still within process specification, they were not investigated. **For process variability reduction, they should be and may even cause specifications to tighten, which in turn may lead to a competitive edge. This is not part of this study.**

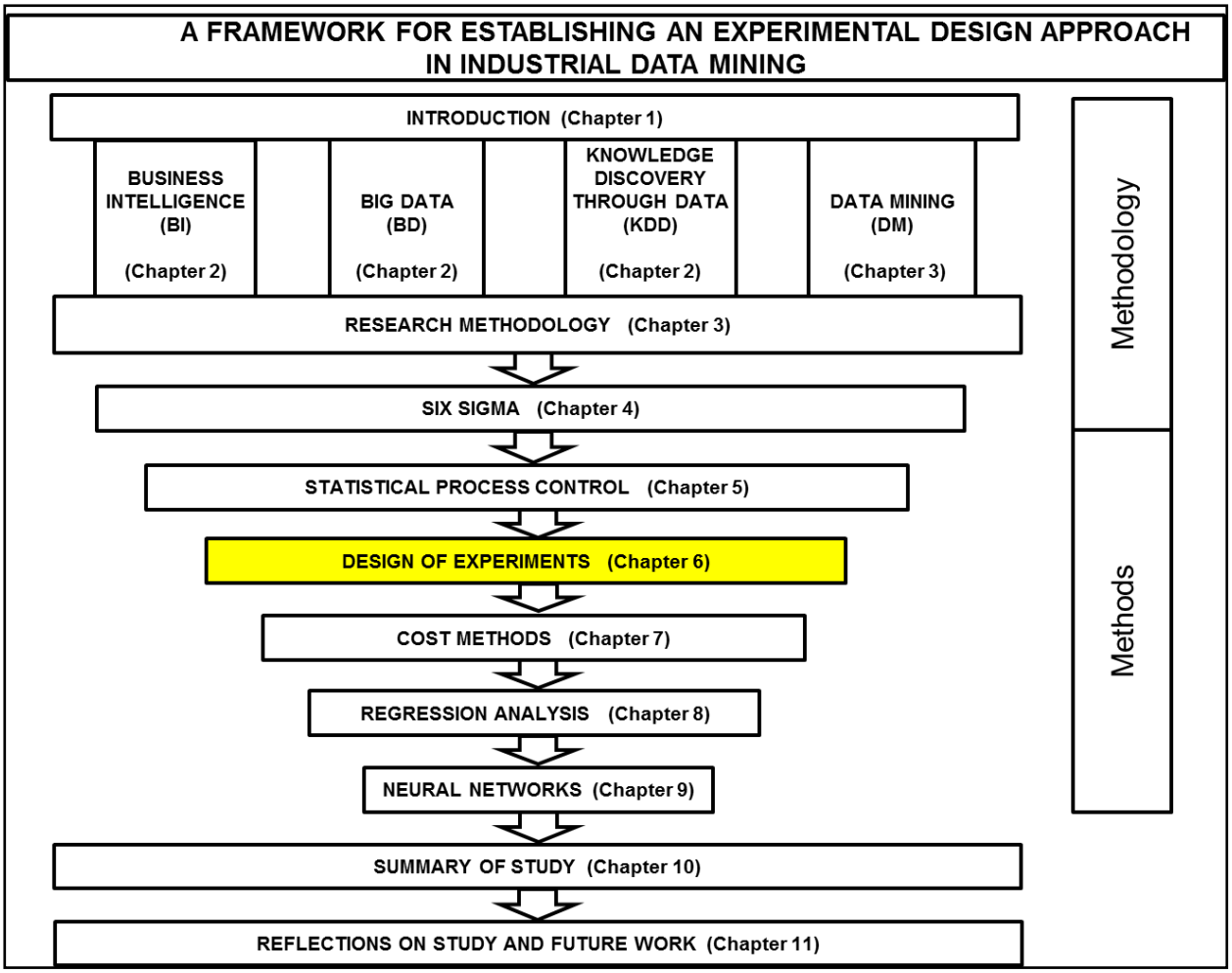
The process capability analysis for all selected independent and dependent variables shows that all are very capable and therefore should be strong predictors for DOE. For this reason, an opportunity exists to embark on a process improvement strategy to reduce specification limits and process variation to reduce the cost impact for deviating from the target value for each variable.

The effect of output variability reduction as the quantity of variables reduces through the variable reduction process by SPC is evident. Variability reduction is significant amongst the three periods, being the largest for period 1 (raw data) and the smallest for the validation (predicted data).

From a predictability point of view, this variability reduction is **significant for prediction accuracy**, but from an **experimental region angle, it shows exploration opportunities** represented by the eliminated variables not analysed yet. As discussed in chapter 6 (DOE), exploration opportunities outside the experimental region should be part of future work.

CHAPTER 6

DESIGN OF EXPERIMENTS (DOE)



6.1 INTRODUCTION

Introducing DOE as a data mining technique for this study is to categorise DOE as an effective data mining technique, enhancing the awareness of using DOE, not only as a traditional statistical technique but also to complement existing methods and methodologies used for statistical data analysis for process improvement. Also, incorporating DOE as a data mining technique within the DMAIC process will limit traditional guesswork in selecting independent variables for experimentation.

For this research Design of Experiments (DOE) is an approach within the data transformation process that focuses on applying DOE on historical data to predict future process improvement. This approach serves as a pre-screening process, using historic

process data and not the typical approach by subjectively selecting design parameters based on experience and personal preferences.

Design of experiments (DOE) is traditionally used for controlled experimentation to determine the effects of variables in a process with the primary goal to extract the maximum amount of unbiased information regarding factors affecting a production process from as few (costly) observations as possible. In industrial settings, complex interaction among many factors that influence a product is rarely known and may have underlying negative processing effects. Generally no-one is interested in them, but they may complicate the process of identifying significant factors in experiments with many factors that would not be possible or practical to identify.

Experimentation (DOE) with the unknown is a high risk factor for failure with a possible huge negative cost implication because no benchmarks exists for a frame of reference; most are dependent on instinct and experience. Experimentation is also characterised as hit-or-miss experiences. This approach, although procedurally dependent, has a high uncertainty level regarding success or failure as well as the risk of the unknown because it is based mainly on experience.

6.2 DISCUSSION ON DESIGN OF EXPERIMENTS (DOE)

DOE tests processes by changing multiple factors in a controlled sequence; results are collected in a controlled way. By doing this, relationships between the changes in factors and the associated responses are identified. It uses a systematic method to determine **relationships amongst factors** affecting a process and the output of that process. In addition, it also complements cause-and-effect analysis to determine **relationships amongst independent variables**.

The primary goal for DOE is to extract the maximum amount of unbiased information regarding the factors affecting a production process from as few (costly) observations as possible.

Although DOE is an advanced statistical technique, Kleppman (2014:51) cautions that the goal of DOE is to acquire process knowledge by scientifically designing a model to measure variables in a controlled way to gain process understanding by experimenting with limited experiments **and not** a statistical technique that will solve all process problems.

Khuri and Cornell (1987:1) believe that most exploratory analysis is twofold, **firstly** quantifying relationships between dependent variables and independent variables to determine settings for experimental runs and **secondly** determining the optimal settings for the independent variables for all experimental runs to ensure optimum values for dependent variables when these experimental runs are run.

These two purposes above are also applicable to this study in the sense that the purpose for analysing the selected process for this study is to find the minimal number of variables scientifically affecting the process outcome by varying the input levels in specific experimental runs and then measuring the corresponding results. Once these variables have been identified, we endeavour to determine at which settings these variables should operate for optimum process performance. The empirical analysis for the DOE application for this study illustrating these purposes is in section 7. 3.

Phadke (1989:11) discusses the effect of variation beyond the control of the designer when formulating a DOE analysis. These factors induce process variation, which is difficult to control, but is part of the natural process variation. Contributors for this variation are called “noise factors” which are normally external of nature or part of process deterioration. The analyst can only statistically measure the effect on process outcomes for understanding.

Although Launsby and Schmidt (1991:2) state that factor changes for DOE design and experimentation give little or no consideration in accommodating historical data that led to changes, these are new, high risk of failure, high cost experimental exercises. In contrast to this approach, the main contribution of this study is an experimental design approach **based on historical data** for industrial DM. One of the reasons for utilising historical data is to avoid costly experimentation into the unknown with a false sense of process optimization based on current and subjective data provided by “experts”. Working with historical data may assist the analyst to gain knowledge about true process behaviour in its natural environment.

6.2.1 Common DOE terminologies

When using DOE as a statistical technique, terminologies used when setting the design, analysing the results or discussing findings are unique to this technique. Common terminologies used for experimental design studies as summarized by Mason *et al*. (1989:92) are shown below, with the terminologies used for this study highlighted with

an asterisk (*) and found during the analysis section 6.3 of this chapter. Those terminologies not marked are not applicable to this study.

Blocking - Groups of homogeneous experimental runs.

Confounding − One or more effects that cannot ambiguously be attributed to a single factor or interaction.

Covariate − An uncontrollable variable that influences the response but is unaffected by any other experimental factors.

Design* − Complete specification of experimental runs

Effect* − Change in the average response due to factor-level combinations

Experimental region* – The defined window of experimentation upon which design is built

Factor* − A controllable factor that is thought to influence the response

Interaction* − Existence of joint factor effects in which the effect of one factor depends on the level of the other factor

Level* − Specific experimental value for a factor

Repeat tests* − Two or more observations for the same experimental factor levels

Replication − Repetition of an entire experiment under two or more sets of conditions by an analyst.

"There should be genuine replication, and it should be done in such a way that variation amongst replicates can provide an accurate measure of errors that affect comparisons amongst different runs." (Box *et al.,* 1978:105)

Replications improve the chance of detecting statistically significant effects in a process. If the objective is to determine a signal-to-noise ratio that guides the number of runs, control charts types and process capability studies, then typically, ANOVA tables combined with experience will determine the level of process noise.

Response* − Outcomes or results of experiment test run − Single combination of factor level that yields an observation on the response.

6.2.2 Principles of controlling DOE

According to Hedges (2008:13), experimental design controls background variability so that systematic effects of treatments are observed. He summarizes three basic principles of controlling design of experiments:

Control by matching − Known sources of variation may be limited and matching is only possible on visible physical characteristics. For this study, measuring variation through historical data, matching was part of the industrial process given by the data set. Because historical data are used, experimental run outcomes also represent matching in a way because various experimental combinations represent different periods of process variation physical process characteristics.

Control by randomisation – Randomizing experimental runs provides a way to assess whether differences in outcomes are due to assumptions or inherent process patterns or process deterioration. Because this study uses historical data (2005-2009), randomization of experimental runs occurs naturally because selecting experimental runs over this period accommodates all external and internal process variation influencing process outcomes.

Control by statistical adjustment − Statistical control is important for high accuracy but it is the weakest of the three experimental design principles because its validity depends on knowing a statistical model for responses. Therefore, sound knowledge of statistical control methods is critical. For this purpose, Statistical Process Control (SPC) utilised potential independent and dependent variables represented by the historical data set to evaluate which will be the critical few for the DOE model. SPC is also a good method to evaluate the stability of outcomes.

6.2.3 Advice for successful DOE

Set good objectives. Any study has to have a purpose, a reason, a direction or a clear defined objective. It may be to reduce variability, or to reduce independent variables that do not significantly influence the process, and to optimise only the few critical variables or any clear defined objectives. The success of a DOE study correlates directly to the quality of set objectives.

Researchers should **be cautious** against trying to study **too many factors**; this is a common mistake by experimenters, which leads to a complex model, often unexplainable for the users to understand. When this happens, the model masks the true contribution of the responses. The original raw database for this study contained 44 dependent and independent variables. From a DOE design perspective, 44 variables were too many and therefore needed to reduce these variables to a critical few for DOE analysis. After screening, only seven independent and one dependent variables were selected for the final DOE analysis. The variable screening process is discussed in chapter 4. Give special attention ensuring that the **appropriate responses are measured**. If non-value added responses are measured, the study may lose its impact due to wasted time. Use the Pareto principle of 80/20, which is always better than to over-analyse with too many variables.

Measure responses quantitatively. A quantitative analysis is always better than a qualitative analysis. If the model requires qualitative measures, transform inputs and responses to a quantitative basis if possible. There was no need to include qualitative analysis for the proposed Experimental design approach because all statistical analyses were performed with quantitative data and all results were represented in a numeric format.

Always randomise the run order. Box and Draper (1969:75) state that randomising experimental runs is important since it ensures that procedures are less dependent on assumptions. It also ensures that if systematic patterns or trends do occur from unsuspecting variables, they are not mistaken for effects of deliberately induced variables. When randomising, the robustness of the model increases because time, different shifts, different operators, tool wear, temperature changes and raw material variability are taken into account in the design. Robustness against the effects of uncontrollable variables that are not chance based should be the objective of the design. This an important feature of DOE because the effect of uncontrollable system noise can lead to poor inferences from DOE results. For this study, randomizing of experimental runs occurred naturally, because the DOE design and analysis are based on historical data representing a period form 2005 – 2009. Selecting experimental runs over this period accommodates all external and internal process variation influencing process outcomes. For this reason, there was no need for randomizing the selected experimental runs.

Screen out known sources of variation by blocking. When grouping experimental runs into homogeneous blocks, variations such as raw materials, machine differences or shift changes screen out noise by known variation. Blocking was not applicable for this study because raw process data represented one operation and homogenous blocking of any variable that might contribute to noise variation were extremely difficult to accommodate. Consider the influence of blocking with the DOE design with current data for future work.

Experimentation should be iterative by nature. Follow up experiments based on results of the first designed experiment results used for the next. Usually experimenters start with an initial screening design to identify significant factors.

Confirmation of critical findings is imperative. Do not assume results are correct and change processes accordingly. Always verify results against known standards, procedures, and process parameters based on normal variation or experience. Results are always within some degree of confidence and, therefore, will always have some degree of uncertainty.

6.3 APPLICATION FOR DESIGN OF EXPERIMENTS (DOE)

Mason *et al.* (1989:97) discuss the importance of the experience region (experimental region, factor space) for DOE model design. The experimental region accommodates all possible levels (high, low) of each quantitative factor, that may be part of the DOE design for which experimentation is possible.

The defined experienced region for this study is important to illustrate the scope of the study for the analyst and the observer. For this study, it is defined as the selected seven critical independent variables with one dependent variable, which give a full factorial DOE two level design of 128 experimental runs. The screening of 44 variables to reach the critical seven independent variables is not considered as the DOE design experimental region but as part of the DB cleaning process for effective data analysis.

Applying designed experiments requires planning skills, statistical skills, teamwork skills and engineering skills. Understanding the process to be analysed is critical because processes are different, therefore the designed experiment should different too. The experiments of a chemical process, for example, would differ from the experiments of a

mechanical process, but one would still go through the methodology in order to apply DOE (Antony, 2014:36).

6.3.1 Determine minimum and maximum values for each independent variable for DOE analysis

Tables 6.1 to 6.3 show the splitting of the database into different formats, a 10/90, 25/75 and a 30/70 percentile to evaluate which to use for the base DOE model. There could be many more derivatives still to be explored, but none of the tables below was chosen because of too many missing values for individual experimental runs between the percentile values of the three tables. When subdividing the database into the percentiles it causes a huge reduction of full DOE runs. The only way full DOE runs are possible would be is if the database were much larger.

To overcome the percentile constraint, a median split for all independent variables was calculated for this study to represent the minimum and maximum levels. By doing this the database will theoretically be split 50/50. See appendix 3 for a sample of the median split database.

Variable	Descriptive Statistics										
	Valid N	**Mean**	**Median**	**Minimum**	**Maximum**	**Low Quartile**	**Upper Quartile**	**Percentile 10**	**Percentile 90**	**Quartile Range**	**Std. Dev**
Mix discharge temp	9772	163.6	164.0	158.0	166.0	163.0	164.0	162.0	165.0	1.00	0.97
Cool begin temp	9772	155.5	156.0	130.0	171.0	153.0	158.0	151.0	160.0	5.00	4.18
Actual cool time	9772	19.3	19.0	11.0	57.0	18.0	20.0	18.0	21.0	2.00	1.70
Actual dump temp	9772	107.2	107.0	95.0	116.0	107.0	108.0	105.0	109.0	1.00	1.63
Actual tamp pressure	9772	5.9	5.9	5.7	6.1	5.8	5.9	5.8	6.0	0.09	0.07
Actual extrusion rate	9772	49.9	50.0	27.0	60.0	50.0	50.0	47.0	54.0	0.00	2.20
Actual extrusion speed	9772	15.6	16.0	8.0	19.0	15.0	16.0	14.0	16.0	1.00	1.06

Table 6.1: 10 and 90 Percentile data cut

Variable	Descriptive Statistics										
	Valid N	**Mean**	**Median**	**Minimum**	**Maximum**	**Low Quartile**	**Upper Quartile**	**Percentile 25**	**Percentile 75**	**Quartile Range**	**Std. Dev**
Mix discharge temp	9772	163.6	164.0	158.0	166.0	163.0	164.0	163.0	164.0	1.00	0.97
Cool begin temp	9772	155.5	156.0	130.0	171.0	153.0	158.0	153.0	158.0	5.00	4.18
Actual cool time	9772	19.3	19.0	11.0	57.0	18.0	20.0	18.0	20.0	2.00	1.70
Actual dump temp	9772	107.2	107.0	95.0	116.0	107.0	108.0	107.0	108.0	1.00	1.63
Actual tamp pressure	9772	5.9	5.9	5.7	6.1	5.8	5.9	5.8	5.9	0.09	0.07
Actual extrusion rate	9772	49.9	50.0	27.0	60.0	50.0	50.0	50.0	50.0	0.00	2.20
Actual extrusion speed	9772	15.6	16.0	8.0	19.0	15.0	16.0	15.0	16.0	1.00	1.06

Table 6.2: 25 and 75 Percentile data cut

Variable	Descriptive Statistics										
	Valid N	**Mean**	**Median**	**Minimum**	**Maximum**	**Low Quartile**	**Upper Quartile**	**Percentile 30**	**Percentile 70**	**Quartile Range**	**Std. Dev**
Mix discharge temp	9772	163.6	164.0	158.0	166.0	163.0	164.0	163.0	164.0	1.00	0.97
Cool begin temp	9772	155.5	156.0	130.0	171.0	153.0	158.0	154.0	158.0	5.00	4.18
Actual cool time	9772	19.3	19.0	11.0	57.0	18.0	20.0	18.0	20.0	2.00	1.70
Actual dump temp	9772	107.2	107.0	95.0	116.0	107.0	108.0	107.0	108.0	1.00	1.63
Actual tamp pressure	9772	5.9	5.9	5.7	6.1	5.8	5.9	5.9	5.9	0.09	0.07
Actual extrusion rate	9772	49.9	50.0	27.0	60.0	50.0	50.0	50.0	50.0	0.00	2.20
Actual extrusion speed	9772	15.6	16.0	8.0	19.0	15.0	16.0	16.0	16.0	1.00	1.06

Table 6.3: 30 and 70 Percentile data cut

Calculating minimum and maximum values for the remaining seven independent variables shown in tables 6.1 to 6.3, the following became evident:

To base the selection of an independent variable on the normality assumption should not be the only criteria. The distribution may look normal, but must be evaluated in combination with a trend graph, which will show the trend pattern.

A normal distribution could be evident, but the distribution of the individual points in time sequence (line graph) may show a different representation.

A fundamental criterion for DOE to be effective is that it is database size dependent. The larger the database, the lower the risk that some DOE runs will have no data, because every time a level for variable is selected, the database is halved. For each independent variable that forms the unique combination of an experimental run, the database is halved. The reason for this is that each variable consists of approximately equal high and low values.

The less normal individual points are distributed, the larger the database must be. There is a ratio between database size and sequential data points not normally distributed.

Skewed distributions are difficult to accommodate in the model. Min and max values do not correspond with other independent variables in the model. Again this points to database size dependency.

There is no clear-cut formula to determine the percentile percentage to trim the extreme values. This is based on experience and the amount of independent variables that want to be retained in the model.

The size of the database needed is proportionate to the number of replicates for each experimental run extracted from the database for the model. Replicated runs have the advantage of predicting variability and a more representative average outcome for each experimental run.

Tables 6.1 to 6.3 show that that none of the above methods of splitting independent variables represents a viable database for DOE analysis. Too many DOE conditions will result into missing values due to skewed distributions.

A median split of each independent variable was the best option because the lower half that represents low values, represents (-1) and the higher half that represents high values, represents (+1). This model was used for DOE and regression analysis for this study. To force the data-set split into equal halves reduces the risk of missing values.

6.3.2 Design [2** (7-0) resolution Full (128 Runs)]. A complete experimental region

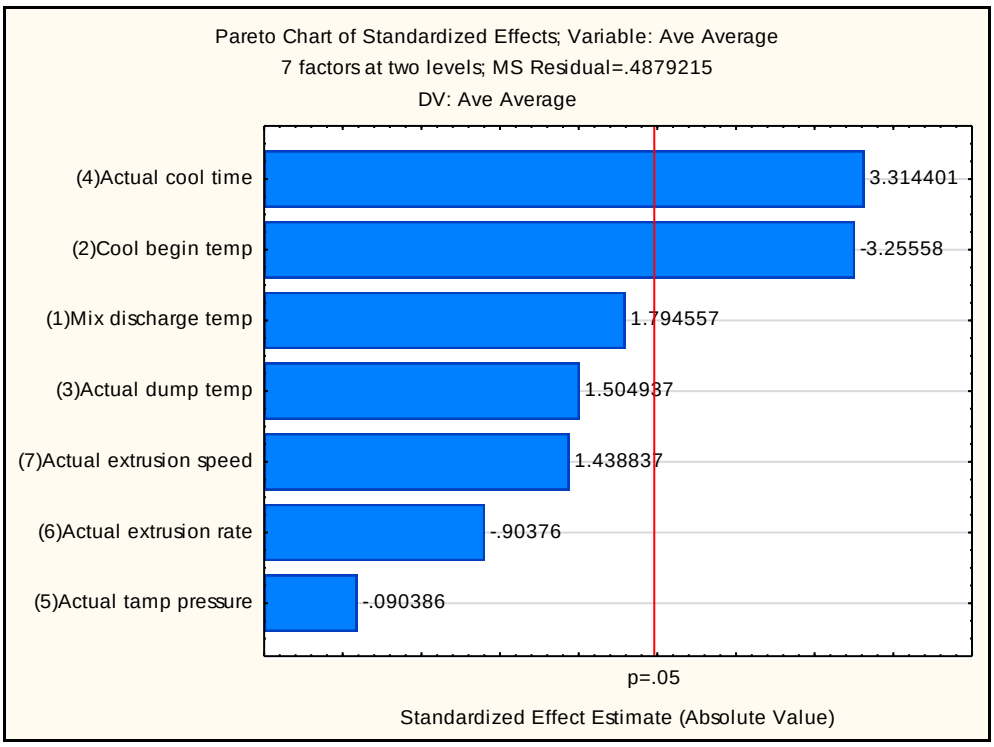
A full resolution model with 128 runs for seven variables at two levels produced nine runs with missing values (7%). These runs are:

Run	Mix discharge temp	Cool begin temp	Actual cool time	Actual dump temp	Actual tamp pressure	Actual extrusion rate	Actual extrusion speed
4	1.00000	1.00000	-1.00000	-1.00000	-1.00000	-1.00000	-1.00000
19	-1.00000	1.00000	-1.00000	-1.00000	1.00000	-1.00000	-1.00000
35	-1.00000	1.00000	-1.00000	-1.00000	-1.00000	1.00000	-1.00000
36	1.00000	1.00000	-1.00000	-1.00000	-1.00000	1.00000	-1.00000
43	-1.00000	1.00000	-1.00000	1.00000	-1.00000	1.00000	-1.00000
51	-1.00000	1.00000	-1.00000	-1.00000	1.00000	1.00000	-1.00000
59	-1.00000	1.00000	-1.00000	1.00000	1.00000	1.00000	-1.00000
67	-1.00000	1.00000	-1.00000	-1.00000	-1.00000	-1.00000	1.00000
75	-1.00000	1.00000	-1.00000	1.00000	-1.00000	-1.00000	1.00000

Table 6.4: Experimental runs with no values (missing values) – 128 runs

Table 6.4 shows gaps in the dataset that do not represent experimental combination runs. These gaps are not necessarily bad, but may indicate that independent variables are not normally distributed or have cyclical data trends within data. The experimental runs that represent gaps in full experimental runs are referred to future work, and should be treated as opportunities measuring the potential impact on the study. These gaps by default limit the experimental region to 93% of the potential region.

The analysis for the full resolution 128 runs, illustrated with graphs and tables follows:



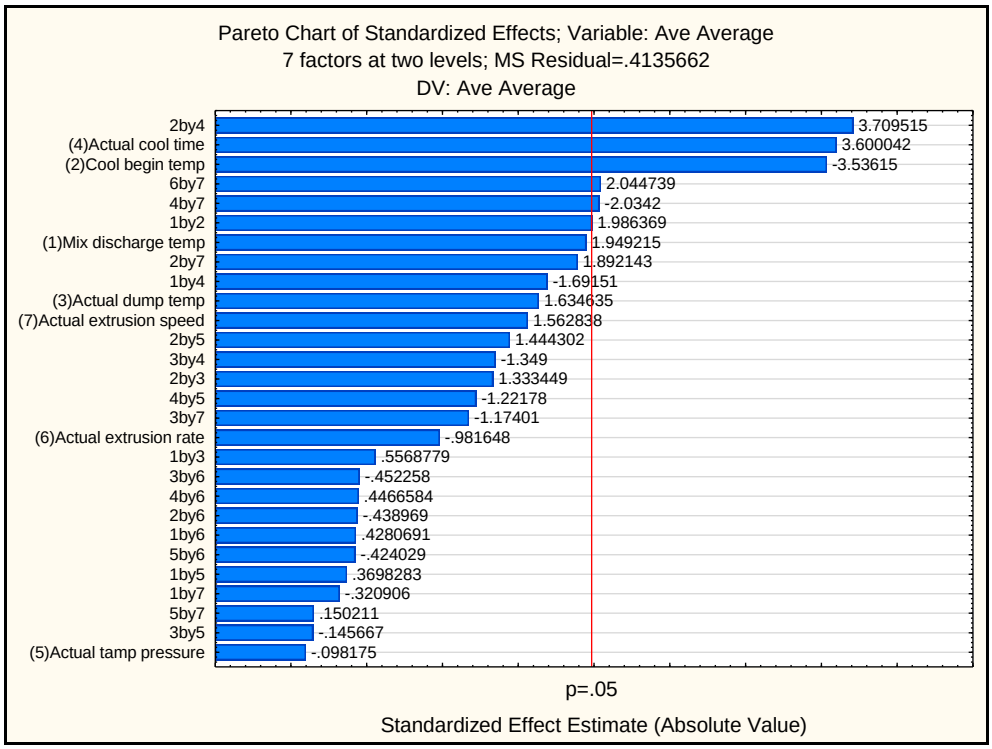
Graph 6.1: DOE standardized main effects 128 runs

Graph 6.1 represents the Pareto graph showing standardized effects for all independent variables not including 2nd or 3rd order interactions for the full factorial, showing two main effects, Actual cool time and Cool begin time as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the two identified main effects are good determinants for the dependent variable is 95%. The residual five variables do have an effect but not statistically significantly.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.19981; Adj:.15313 (Thesis data - First period) DB 128 runs) 7 factors at two levels; MS Residual=.4879215 DV: Ave Average			
	Effect	Std.Err.	t(120)	p
Mean/Interc.	2.72055	0.06174	44.0643	0.00000
(1)Mix discharge temp (x)	0.22159	0.12348	1.7945	0.07524
(2)Cool begin temp (x)	-0.40200	0.12348	-3.2555	0.00147
(3)Actual dump temp (x)	0.18583	0.12348	1.5049	0.13496
(4)Actual cool time (x)	0.40926	0.12348	3.3144	0.00121
(5)Actual tamp pressure (x)	-0.01116	0.12348	-0.0903	0.92813
(6)Actual extrusion rate (x)	-0.11159	0.12348	-0.9037	0.36793
(7)Actual extrusion speed (x)	0.17766	0.12348	1.4388	0.15280

Table 6.5: DOE main effect estimates summary estimated period 128 runs – Average pressure

Following the Pareto Graph 6.1, Table 6.5 shows the same two variables (Actual cool time and Cool begin time) as statistically significant main effects with p values of 0.004171 & 0.001215 respectively. When analysing only main effects, not including 2nd or 3rd order interactions for the full factorial design, Actual cool time and Cool begin time are significant main effects determining average extrusion pressure, calculated for p=0.05.



Graph 6.2: DOE standardized effects for main & two order interactions 128 runs

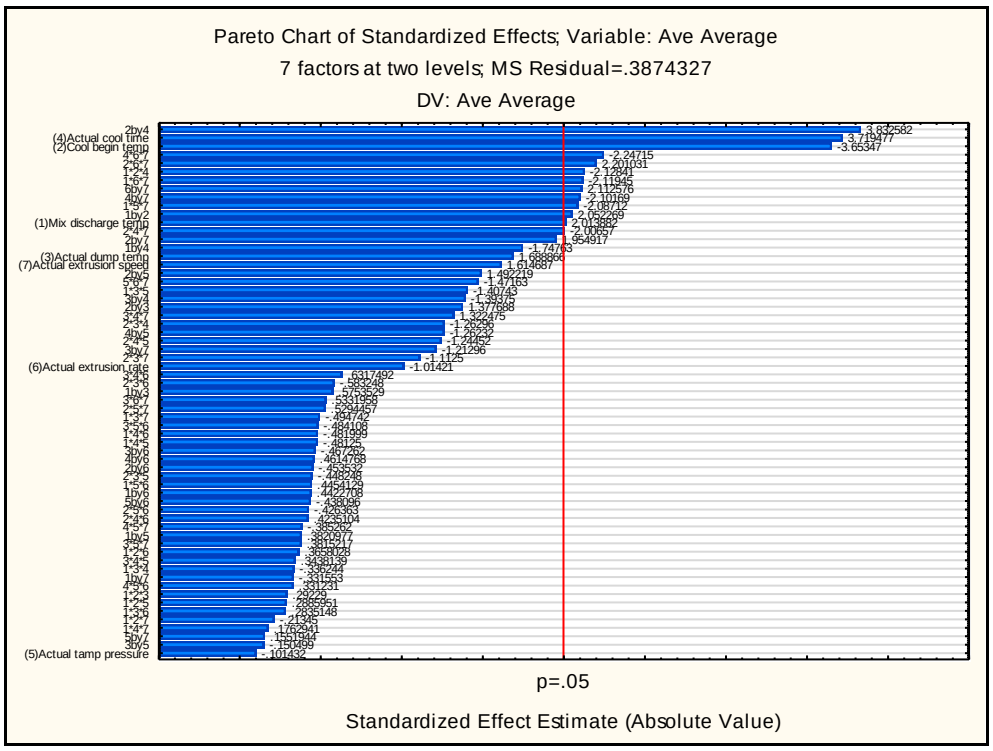
Graph 6.2 represents the Pareto graph showing standardized effects for all independent variables including 2nd but not 3rd order interactions for the full factorial, showing the same two main effects, Actual cool time and Cool begin time, as for main effects but with an additional four significant 2nd order interactions as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the two identified main effects and the four 2nd order interactions are good determinants for the dependent variable is 95%.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.44045; Adj:.28219 (Thesis data - First period) DB 128 runs) 7 factors at two levels; MS Residual=.4135662 DV: Ave Average			
	Effect	Std.Err.	t(99)	p
Mean/Interc.	2.720556	0.056842	47.86193	0.000000
(1)Mix discharge temp (x)	0.221594	0.113684	1.94921	0.054100
(2)Cool begin temp (x)	-0.402002	0.113684	-3.53615	0.000619
(3)Actual dump temp (x)	0.185831	0.113684	1.63463	0.105302
(4)Actual cool time (x)	0.409265	0.113684	3.60004	0.000499
(5)Actual tamp pressure (x)	-0.011161	0.113684	-0.09818	0.921992
(6)Actual extrusion rate (x)	-0.111597	0.113684	-0.98165	0.328666
(7)Actual extrusion speed (x)	0.177669	0.113684	1.56284	0.121282
1 by 2	0.225817	0.113684	1.98637	0.049757
2 by 4	0.421711	0.113684	3.70951	0.000343
4 by 7	-0.231255	0.113684	-2.03420	0.044606
6 by 7	0.232453	0.113684	2.04474	0.043535

Table 6.6: DOE main and two order effect estimates summary estimated period 128 runs – Average

Following the Pareto Graph 6.2, Table 6.6 shows the same two variables (Actual cool time and Cool begin time) as statistically significant main effects with p values of 0.000499 & 0.000619 respectively, but also includes four significant 2^{nd} order interactions. This confirms that the two main effects stays significant irrespective if 2^{nd} order interactions are included. In addition to the main effects, four 2^{nd} order interactions were also significant.

An interaction between the two main effects is also significant and shows a positive effect, but for the individual main effects, cool begin temperature has a negative effect and actual cool time has a positive effect. The reason for this should be investigated and is referred to future work. When analysing main effects, including 2^{nd} but no 3^{rd} order interactions for the full factorial design, Actual cool time and Cool begin time remain significant main effects determining average extrusion pressure, calculated for p=0.05. The four interactions are also significant as determinants for average extrusion pressure.



Graph 6.3: DOE standardized effects for main, two and three order interactions 128 runs

Graph 6.3 represents the Pareto graph showing standardized effects for all independent variables including 2nd and 3rd order interactions for the full factorial. It shows three main effects, Actual cool time, Cool begin time and Mix discharge temperature for main effects with four significant 2nd order interactions and six significant 3rd order interactions as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the three main effects, four 2nd order interactions and six 3rd order interactions are good determinants for the dependent variable, is 95%.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.66113; Adj:.32755 (Thesis data - First period) DB 128 runs) 7 factors at two levels; MS Residual=.3874327 DV: Ave Average			
	Effect	Std.Err.	t(64)	p
Mean/Interc.	2.720556	0.055017	49.44980	0.000000
(1)Mix discharge temp (x)	0.221594	0.110033	2.01388	0.048230
(2)Cool begin temp (x)	-0.402002	0.110033	-3.65347	0.000523
(3)Actual dump temp (x)	0.185831	0.110033	1.68887	0.096112
(4)Actual cool time (x)	0.409265	0.110033	3.71948	0.000423
(5)Actual tamp pressure (x)	-0.011161	0.110033	-0.10143	0.919524
(6)Actual extrusion rate (x)	-0.111597	0.110033	-1.01421	0.314299
(7)Actual extrusion speed (x)	0.177669	0.110033	1.61469	0.111298
1 by 2	0.225817	0.110033	2.05227	0.044238
2 by 4	0.421711	0.110033	3.83258	0.000292
4 by 7	-0.231255	0.110033	-2.10169	0.039522
6 by 7	0.232453	0.110033	2.11258	0.038544
1*2*4	-0.234195	0.110033	-2.12841	0.037159
1*5*7	-0.229652	0.110033	-2.08712	0.040865
1*6*7	-0.233210	0.110033	-2.11945	0.037937
2*4*7	-0.220789	0.110033	-2.00657	0.049025
2*6*7	0.242186	0.110033	2.20103	0.031348
4*6*7	-0.247261	0.110033	-2.24715	0.028086

Table 6.7: DOE main, two and three order effect estimates summary estimated period 128 runs – Average

Following the Pareto Graph 6.3, Table 6.7 shows the same two variables (Actual cool time and Cool begin time) with a third variable (mix discharge temperature) as statistically significant main effects with p values of 0.000422, 0.000523 and 0.04823 respectively, but also includes four significant 2nd order interactions and six significant 3rd order interactions. This confirms that at least two main effects stays consistently significant, irrespective if 2nd or 3rd order interactions are included. In addition to the main effects, the same four 2nd order interactions with six additional 3rd order interactions are also significant.

An interaction between the two main effects is also significant and still shows a positive effect, but for the individual main effects, cool begin temperature has a negative effect and actual cool time has a positive effect. All 3rd interactions, that includes the significant main effect, are all negative. The reason for this should be investigated and is referred to future work. When analysing main effects, including 2nd and 3rd order interactions for the full factorial design, Actual cool time and Cool begin time remain significant main effects but include mix discharge temperature for determining average

extrusion pressure, calculated for p=0.05. The four 2nd order interactions with the six 3rd order interactions are also significant as determinants for average extrusion pressure.

6.3.2.1 Discussion

Interaction	Main	2 way	3 way	Main	2 way	3 way	Main	2 way	3 way
# Runs	128	128	128	64	64	64	32	32	32
Intersection	2.721	2.721	2.721	2.745	2.745	2.745	2.922	2.922	2.922
Var1 (x1)	0.222	0.222	0.222	0.168	0.168	0.168	0.032	0.032	0.032
Var2 (x2)	-0.402	-0.402	-0.402	-0.351	-0.351	-0.351	-0.018	-0.018	-0.018
Var3 (x3)	0.186	0.186	0.186	0.359	0.359	0.359	-0.007	-0.007	-0.007
Var4 (x4)	0.409	0.409	0.409	0.036	0.036	0.036	0.064	0.064	0.064
Var5 (x5)	-0.011	-0.011	-0.011	0.032	0.032	0.032	-0.181	-0.181	-0.181
Var6 (x6)	-0.112	-0.112	-0.112	-0.245	-0.245	-0.245	-0.081	-0.081	-0.081
Var7 (x7)	0.178	0.178	0.178	0.125	0.125	0.125	-0.060	-0.060	-0.060
1 by 2		0.226	0.226						
2 by 4		0.422	0.422						
4 by 7		-0.231	-0.231						
6 by 7		0.232	0.232		0.380				
2 by 3					0.374				
1*2*4			-0.234						
1*5*7			-0.230						
1*6*7			-0.233						
2*4*7			-0.221						
2*6*7			0.242						
4*6*7			-0.247						
Missing values	7%			6%			0%		

Table 6.8: Experimental design outcome summary for 128, 64 and 32 runs

For the full factorial of 128 runs, there were 7% missing values that represent 9 experimental runs, see Table 6.4. It seems that variables 1-3 in Table 6.4 contribute the most towards missing values because the factor levels are either high or low, which means these variables are not normally distributed and that causes “gaps” in the data set. The effect of these missing values on the outcomes is not clear but offers an opportunity to explore in future work.

Referring to summary of effects Table 6.8, **for main effects**, the same two variables are significant for the no interaction model, the 2nd order interaction model with a third significant variable as a main effect for the 3rd order interaction model.

When including **2nd order interactions**, the same four 2nd order interactions are significant for 2nd and 3rd order interaction models.

When including **3rd order interactions,** six 3rd order interactions are significant for the 3rd order interaction model. The effects for these 3rd order interactions are all negative which needs to be investigated further to have a deeper understanding of the dataset.

Although significant main, two-way and three-way effects are present for the 128 run model, the impact of missing values, sign changes for 2nd and 3rd order interactions on the results is unknown and therefore will not be considered for this study. A more comprehensive analysis is necessary to evaluate the impact sensitivity to missing values and sign changes on significant effects. This is part of future work for this study.

6.3.3 Design [2** (7-1) resolution VII (64 Runs)]. A partial experimental region

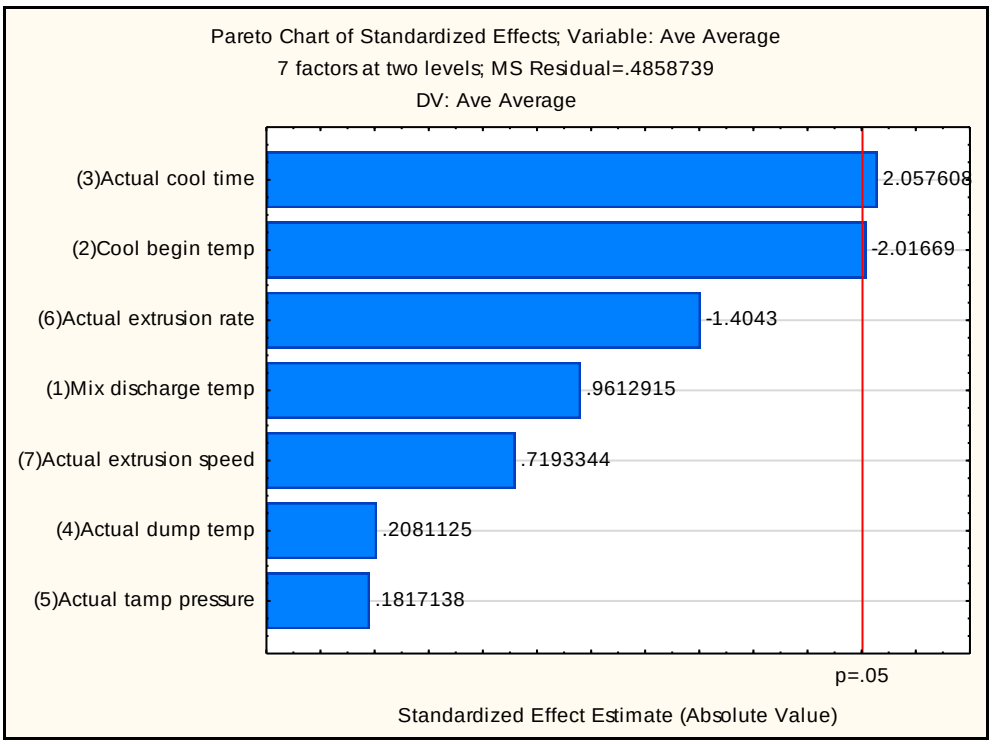
A VII resolution with 64 runs for seven variables at two levels produced 4 runs with missing values (6%). These runs are:

Run	Mix discharge temp	Cool begin temp	Actual cool time	Actual dump temp	Actual tamp pressure	Actual extrusion rate	Actual extrusion speed
11	-1	1	-1	1	-1	-1	1
36	1	1	-1	-1	-1	1	-1
43	-1	1	-1	1	-1	1	-1
51	-1	1	-1	-1	1	1	-1

Table 6.9: Experimental runs with no values (missing values) – 64 runs

Table 6.9 shows gaps in the dataset that do not represent experimental combination runs. These gaps are not necessarily bad but may indicate that independent variables are not normally distributed or have cyclical data trends within data. The experimental runs that represent gaps in the 2** (7-1) resolution VII (64 RUNS) DOE design is referred to future work, and should be treated as opportunities measuring the potential impact on the study. These gaps by default limit the experimental region to 94% of the potential region.

The analysis for the VII resolution 64 runs, illustrated with graphs and tables, follows:



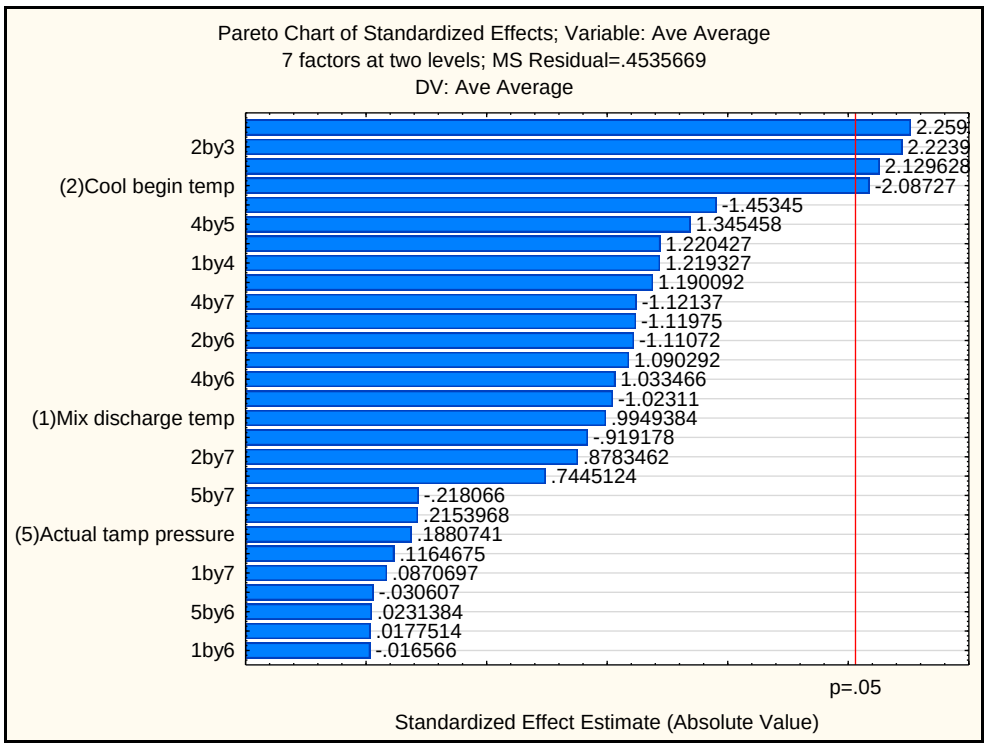
Graph 6.4: DOE standardized main effects 64 runs.

Graph 6.4 represents the Pareto graph showing standardized effects for all independent variables not including 2nd or 3rd order interactions for the 2** (7-1) resolution VII (64 RUNS) DOE factorial design, showing two main effects, Actual cool time and Cool begin time as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the two identified main effects are good determinants for the dependent variable is 96%. The residual five variables do have an effect but not statistically significantly.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.17393; Adj:.07067 (Thesis data - First period) DB 64 runs) 7 factors at two levels; MS Residual=.4858739 DV: Ave Average			
	Effect	Std.Err.	t(56)	p
Mean/Interc.	2.745	0.087	31.51	0.000
(1)Mix discharge temp (x)	0.168	0.174	0.96	0.341
(2)Cool begin temp (x)	-0.351	0.174	-2.02	0.049
(3)Actual cool time (x)	0.359	0.174	2.06	0.044
(4)Actual dump temp (x)	0.036	0.174	0.21	0.836
(5)Actual tamp pressure (x)	0.032	0.174	0.18	0.856
(6)Actual extrusion rate (x)	-0.245	0.174	-1.40	0.166
(7)Actual extrusion speed (x)	0.125	0.174	0.72	0.475

Table 6.10: DOE main effect estimates summary estimated period 64 runs – Average pressure

Following the Pareto Graph 6.4, Table 6.10 shows the same two variables (Actual cool time and Cool begin time) as statistically significant main effects with p values of 0.044295 & 0.048533 respectively. When analysing only main effects, not including 2nd or 3rd order interactions for the 2** (7-1) resolution VII (64 RUNS) DOE factorial design, Actual cool time and Cool begin time are significant main effects determining average extrusion pressure, calculated for p=0.05.



Graph 6.5: DOE standardized effects for main & two order interactions 64 runs.

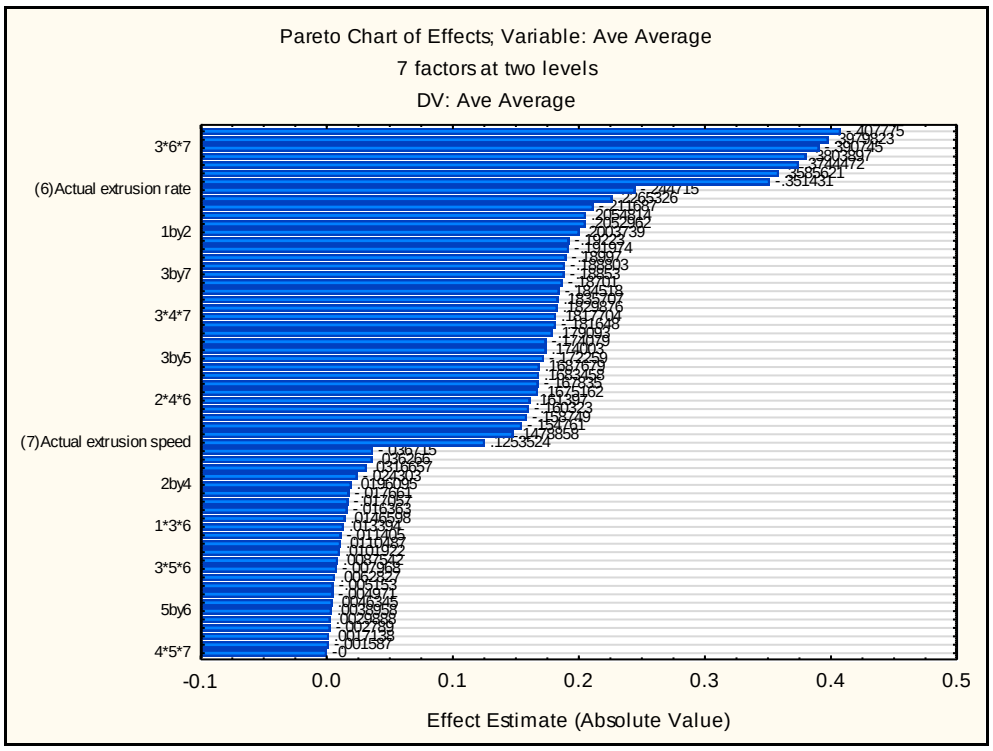
Graph 6.5 represents the Pareto graph showing standardized effects for all independent variables including 2nd but not 3rd order interactions for the 2** (7-1) resolution VII (64 RUNS) DOE factorial design. The graph shows the same two main effects, Actual cool time and Cool begin time as for main effects, but with an additional two significant 2nd order interactions as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the two identified main effects and the two 2nd order interactions are good determinants for the dependent variable is 96%.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.51803; Adj:.13246 (Thesis data - First period) DB 64 runs) 7 factors at two levels; MS Residual=.4535669 DV: Ave Average			
	Effect	Std.Err.	t(35)	p
Mean/Interc.	2.745	0.084	32.61	0.000
(1)Mix discharge temp (x)	0.168	0.168	0.99	0.327
(2)Cool begin temp (x)	-0.351	0.168	-2.09	0.044
(3)Actual cool time (x)	0.359	0.168	2.13	0.040
(4)Actual dump temp (x)	0.036	0.168	0.22	0.831
(5)Actual tamp pressure (x)	0.032	0.168	0.19	0.852
(6)Actual extrusion rate (x)	-0.245	0.168	-1.45	0.155
(7)Actual extrusion speed (x)	0.125	0.168	0.74	0.462
2 by 3	0.374	0.168	2.22	0.033
6 by 7	0.380	0.168	2.26	0.030

Table 6.11: DOE main and two order effect estimates summary estimated period 64 runs – Average

Following the Pareto Graph 6.5, Table 6.11 shows the same two variables (Actual cool time and Cool begin time) as statistically significant main effects with p values of 0.040311 & 0.044211 respectively. This confirms that the two main effects stay significant irrespective if 2nd order interactions are included. In addition to the main effects, two 2nd order interactions ae also significant.

An interaction between the two main effects is also significant and shows a positive effect, but for the individual main effects, cool begin temperature has a negative effect and actual cool time has a positive effect. The reason for this should be investigated and is referred to future work. When analysing main effects, including 2nd but no 3rd order interactions for the 2** (7-1) resolution VII (64 RUNS) DOE factorial design, Actual cool time and Cool begin time remain significant main effects determining average extrusion pressure, calculated for p=0.05. The two interactions are also significant as determinants for average extrusion pressure.



Graph 6.6: DOE standardized effects for main, two and three order interactions 64 runs.

Graph 6.6 represents the Pareto graph showing standardized effects for all independent variables including 2nd and 3rd order interactions for the 2** (7-1) resolution VII (64 RUNS) DOE factorial design. It shows no main effects, no significant 2nd order or significant 3rd order interactions for determinants of average extrusion pressure. This means that with p=0.05, the certainty that no variables are good determinants for the dependent variable is 96%.

Interaction	Main	2 way	3 way	Main	2 way	3 way	Main	2 way	3 way
# Runs	128	128	128	64	64	64	32	32	32
Intersection	2.721	2.721	2.721	2.745	2.745	2.745	2.922	2.922	2.922
Var1 (x1)	0.222	0.222	0.222	0.168	0.168	0.168	0.032	0.032	0.032
Var2 (x2)	-0.402	-0.402	-0.402	-0.351	-0.351	-0.351	-0.018	-0.018	-0.018
Var3 (x3)	0.186	0.186	0.186	0.359	0.359	0.359	-0.007	-0.007	-0.007
Var4 (x4)	0.409	0.409	0.409	0.036	0.036	0.036	0.064	0.064	0.064
Var5 (x5)	-0.011	-0.011	-0.011	0.032	0.032	0.032	-0.181	-0.181	-0.181
Var6 (x6)	-0.112	-0.112	-0.112	-0.245	-0.245	-0.245	-0.081	-0.081	-0.081
Var7 (x7)	0.178	0.178	0.178	0.125	0.125	0.125	-0.060	-0.060	-0.060
1 by 2		0.226	0.226						
2 by 4		0.422	0.422						
4 by 7		-0.231	-0.231						
6 by 7		0.232	0.232		0.380				
2 by 3					0.374				
1*2*4			-0.234						
1*5*7			-0.230						
1*6*7			-0.233						
2*4*7			-0.221						
2*6*7			0.242						
4*6*7			-0.247						
Missing values	7%			6%			0%		

Table 6.12: Experimental design outcome summary for 128, 64 and 32 runs

6.3.3.1 Discussion

Interaction	Main	2 way	3 way	Main	2 way	3 way	Main	2 way	3 way
# Runs	128	128	128	64	64	64	32	32	32
Intersection	2.721	2.721	2.721	2.745	2.745	2.745	2.922	2.922	2.922
Var1 (x1)	0.222	0.222	0.222	0.168	0.168	0.168	0.032	0.032	0.032
Var2 (x2)	-0.402	-0.402	-0.402	-0.351	-0.351	-0.351	-0.018	-0.018	-0.018
Var3 (x3)	0.186	0.186	0.186	0.359	0.359	0.359	-0.007	-0.007	-0.007
Var4 (x4)	0.409	0.409	0.409	0.036	0.036	0.036	0.064	0.064	0.064
Var5 (x5)	-0.011	-0.011	-0.011	0.032	0.032	0.032	-0.181	-0.181	-0.181
Var6 (x6)	-0.112	-0.112	-0.112	-0.245	-0.245	-0.245	-0.081	-0.081	-0.081
Var7 (x7)	0.178	0.178	0.178	0.125	0.125	0.125	-0.060	-0.060	-0.060
1 by 2		0.226	0.226						
2 by 4		0.422	0.422						
4 by 7		-0.231	-0.231						
6 by 7		0.232	0.232		0.380				
2 by 3					0.374				
1*2*4			-0.234						
1*5*7			-0.230						
1*6*7			-0.233						
2*4*7			-0.221						
2*6*7			0.242						
4*6*7			-0.247						
Missing values	7%			6%			0%		

Table 6.13: Experimental design outcome summary for 128, 64 and 32 runs

For the 2** (7-1) resolution VII (64 RUNS) DOE factorial design there are 6% missing values that represent 4 experimental runs, see Table 6.9. It seems that variables 1-3 in Table 6.11 **also** contribute the most towards missing values because the factor levels are either high or low as for the full resolution of 128 runs, which means these variables are not normally distributed and that causes “gaps” in the data set. The effect of these missing values on the outcomes is not clear but offers an opportunity to explore in future work.

Referring to summary of effects Table 6.13, **for main effects**, the same two variables are significant for the no interaction model and 2nd order interaction model with no significant variables for the three-way interaction model.

When including 2nd order interactions, two variables are significant for the 2nd order interaction model and no significant variables for the three-way interaction model.

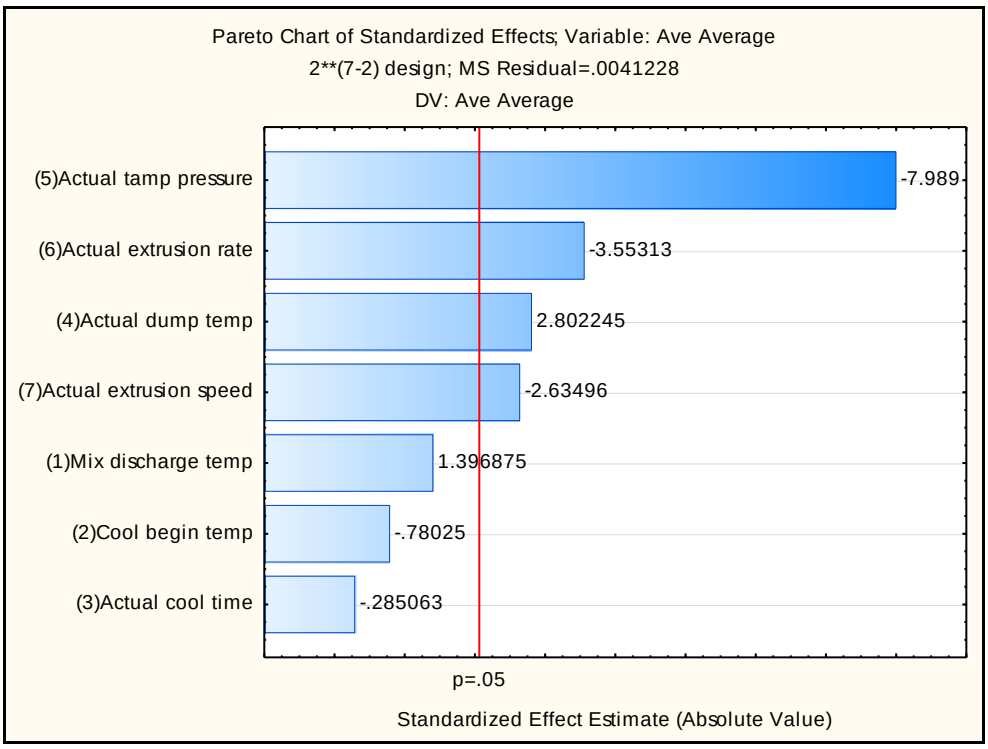
When including 3rd order **interactions,** no variables are significant for 1st, 2nd or 3rd order interactions.

Although significant main, two-way and no three-way effects are present, the impact of missing values on the results is unknown and therefore will not be considered for this study. A more comprehensive analysis is necessary to evaluate the impact sensitivity to missing values on significant effects. This is part of future work for this study

6.3.4 Design [2** (7-2) resolution IV (32 Runs)]

A 2** (7-2) resolution IV (32 RUNS) DOE factorial design for seven variables at two levels produced no missing values.

Because no gaps exist, the experimental region is 100% of the potential region for this design. The analysis for the 2** (7-2) resolution IV (32 RUNS) DOE design, illustrated with graphs and tables, follows:



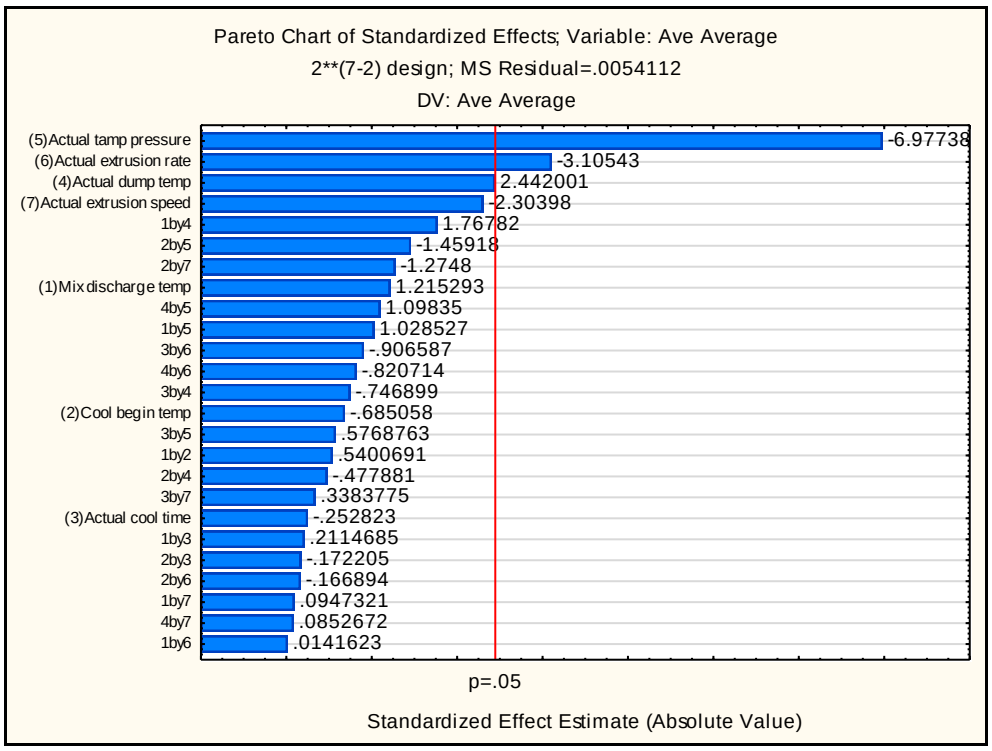
Graph 6.7: DOE standardized main effects 32 runs

Graph 6.7 represents the Pareto graph showing standardized effects for all independent variables, not including 2nd or 3rd order interactions for the 2** (7-2) resolution IV (32 RUNS) DOE factorial design, showing four main effects, Actual tamp pressure, Actual extrusion rate, Actual dump temperature and Actual extrusion speed as the most important determinants of average extrusion pressure. This means that with p=0.05, the certainty that the four identified main effects are good determinants for the dependent variable is 95%. The residual three variables do have an effect but not statistically significantly.

Factor	Effect Estimates; Var.:Ave Average; R-sqr=.79641; Adj:.73703 (Thesis data) extended DB) 2**(7-2) design; MS Residual=.0041228 DV: Ave Average			
	Effect	Std.Err.	t(24)	p
Mean/Interc.	2.922	0.011	257.5	0.000
(5)Actual tamp pressure (x)	-0.181	0.023	-8.0	0.000
(6)Actual extrusion rate (x)	-0.081	0.023	-3.6	0.002
(4)Actual dump temp (x)	0.064	0.023	2.8	0.010
(7)Actual extrusion speed (x)	-0.060	0.023	-2.6	0.015
(1)Mix discharge temp (x)	0.032	0.023	1.4	0.175
(2)Cool begin temp (x)	-0.018	0.023	-0.8	0.443
(3)Actual cool time (x)	-0.006	0.023	-0.3	0.778

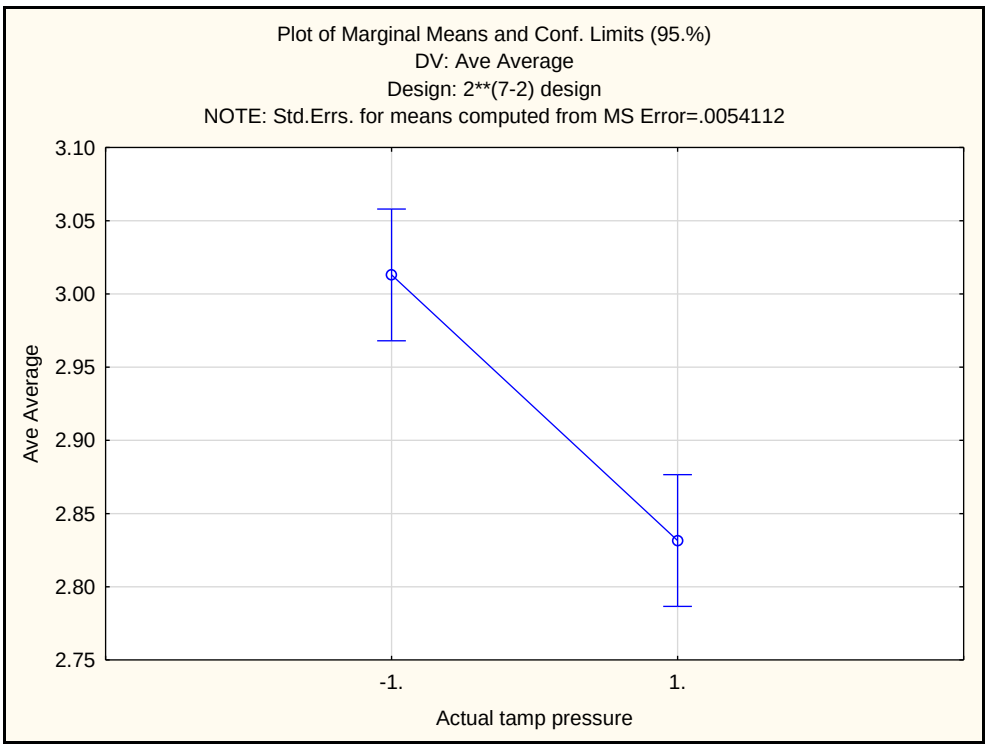
Table 6.14: DOE effect estimates summary estimated period 32 runs – Average pressure

Following the Pareto Graph 6.7, Table 6.14 shows the same four variables (Actual tamp pressure, Actual extrusion rate, Actual dump temperature and Actual extrusion speed) as statistically significant main effects with p values of 0.00, 0.001615, 0.009878 and 0.014508 respectively. When analysing only main effects, not including 2^{nd} or 3^{rd} order interactions for the 2** (7-2) resolution IV (32 RUNS) DOE factorial design, Actual tamp pressure, Actual extrusion rate, Actual dump temperature and Actual extrusion speed are significant main effects determining average extrusion pressure, calculated for p=0.



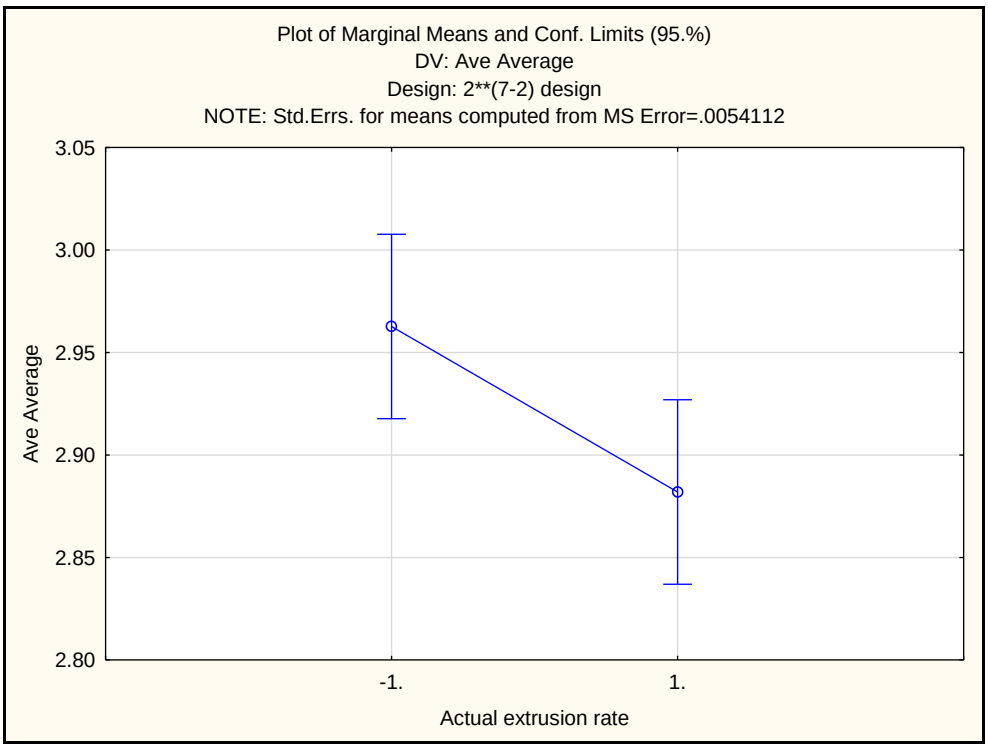
Graph 6.8: DOE standardized effects for main & two order interactions 32 runs

Graph 6.8 represents the Pareto graph showing standardized effects for all independent variables including 2nd but not 3rd order interactions for the 2** (7-2) resolution IV (32 RUNS) DOE factorial design, showing the same four main effects, Actual tamp pressure, Actual extrusion rate, Actual dump temperature and Actual extrusion speed as for main effects. No significant 2nd order interactions are present for determinants of average extrusion pressure. This means that with p=0.05, the certainty that the four identified main effects are good determinants for the dependent variable is 95%. A possible reason for no significant interactions is that the minimum and maximum values selected from the historical data set were too narrow and therefore could not calculate real significant interactions.



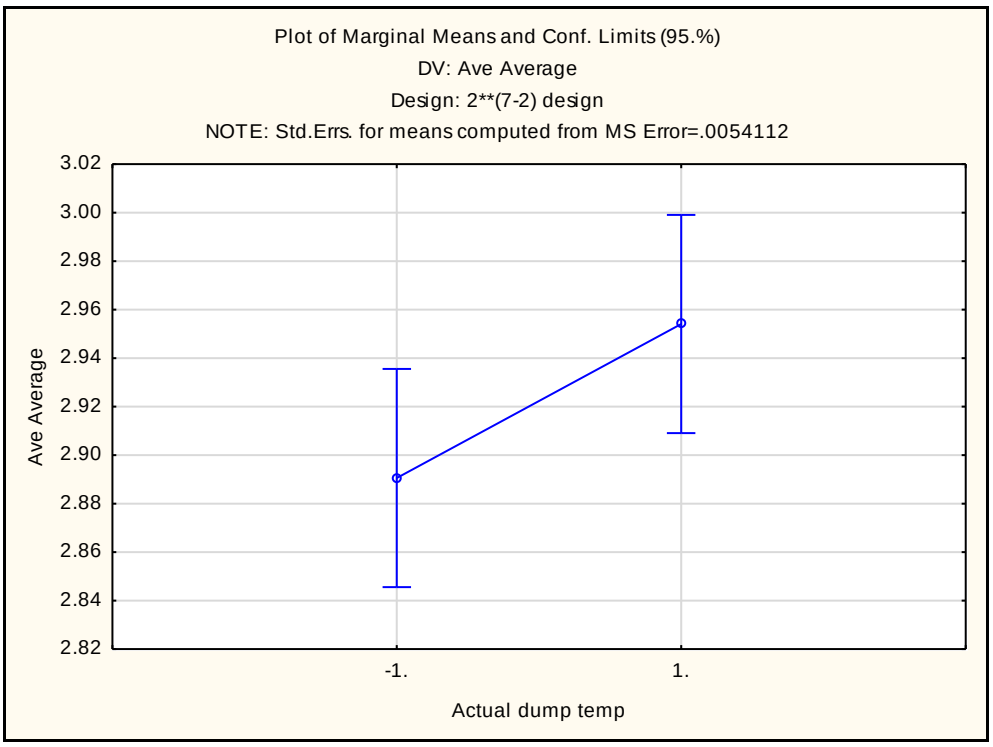
Graph 6.9: DOE marginal means effects for Actual tamp pressure

Graph 6.9 shows the marginal effect on the dependent variable with a change from a low to high-level for an independent variable. Actual tamp pressure has the biggest negative effect (-0.181466) on the average extrusion pressure, refer to Table 6.14. This means that the average extrusion pressure reduces with 0.181466 with a change from minimum to maximum value for actual tamp pressure. This variable was also recognised as the red “x” during the regression analysis, refer to chapter 8. Both DOE and regression recognise variable 5 as the main driver for predicting average extrusion pressure.



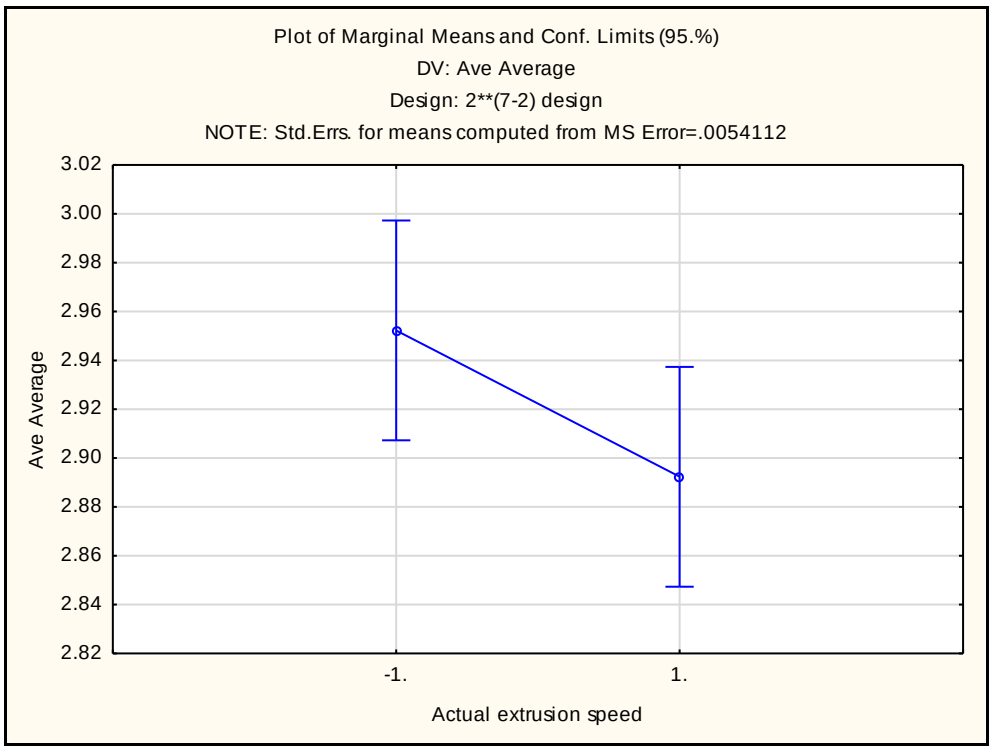
Graph 6.10: DOE marginal means effects for actual extrusion rate

Graph 6.10 shows the marginal effect on the dependent variable with a change from a low to high-level change for an independent variable. Actual extrusion rate has the second highest negative effect (-0.080765) on the average extrusion pressure, refer to Table 6.14. This means that the average extrusion pressure reduces with 0.080765 with a change from minimum to maximum value for actual extrusion pressure.



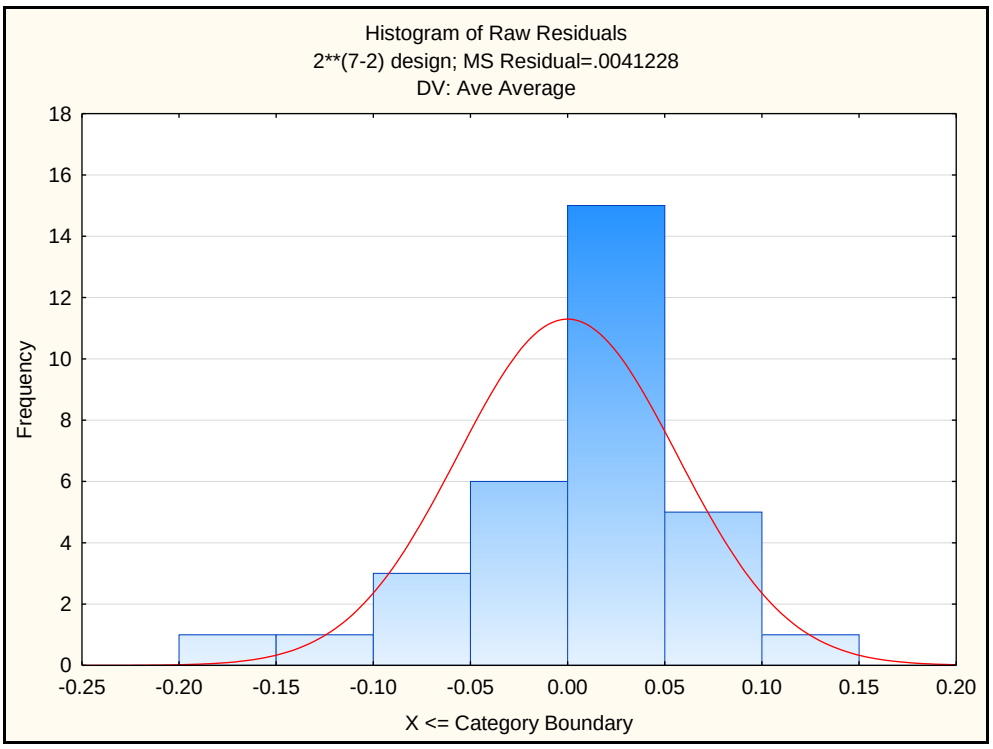
Graph 6.11: DOE marginal means effects for actual dump temperature

Graph 6.11 shows the marginal effect on the dependent variable with a change from a low to high-level change for an independent variable. Actual dump temperature has the third highest positive effect (0.0.063511) on the average extrusion pressure; refer to Table 6.14. This means that the average extrusion pressure reduces with 0.181466 with a change from minimum to maximum value for actual dump temperature.



Graph 6.12: DOE marginal means effects for actual extrusion speed

Graph 6.12 shows the marginal effect on the dependent variable with a change from a low to high-level change for an independent variable. Actual extrusion speed has the fourth highest negative effect (-0.059921) on the average extrusion pressure; refer to Table 6.14. This means that the average extrusion pressure reduces with 0.059921, a change from minimum to maximum value for actual extrusion speed.



Graph 6.13: Raw residual histogram for 32 runs

Graph 6.13 represents the residual histogram, which shows a normal distribution for the dependent variable, average extrusion pressure, which indicates that the predictive DOE model can be used for predicting. Being not skewed, it does not show extreme over- or under-prediction estimates.

6.3.5 Discussion

Interaction	Main	2 way	3 way	Main	2 way	3 way	Main	2 way	3 way
# Runs	128	128	128	64	64	64	32	32	32
Intersection	2.721	2.721	2.721	2.745	2.745	2.745	2.922	2.922	2.922
Var1 (x1)	0.222	0.222	0.222	0.168	0.168	0.168	0.032	0.032	0.032
Var2 (x2)	-0.402	-0.402	-0.402	-0.351	-0.351	-0.351	-0.018	-0.018	-0.018
Var3 (x3)	0.186	0.186	0.186	0.359	0.359	0.359	-0.007	-0.007	-0.007
Var4 (x4)	0.409	0.409	0.409	0.036	0.036	0.036	0.064	0.064	0.064
Var5 (x5)	-0.011	-0.011	-0.011	0.032	0.032	0.032	-0.181	-0.181	-0.181
Var6 (x6)	-0.112	-0.112	-0.112	-0.245	-0.245	-0.245	-0.081	-0.081	-0.081
Var7 (x7)	0.178	0.178	0.178	0.125	0.125	0.125	-0.060	-0.060	-0.060
1 by 2		0.226	0.226						
2 by 4		0.422	0.422						
4 by 7		-0.231	-0.231						
6 by 7		0.232	0.232		0.380				
2 by 3					0.374				
1*2*4			-0.234						
1*5*7			-0.230						
1*6*7			-0.233						
2*4*7			-0.221						
2*6*7			0.242						
4*6*7			-0.247						
Missing values	7%			6%			0%		

Table 6.15: Experimental design outcome summary for 128, 64 and 32 runs

For the 2** (7-2) resolution IV (32 RUNS) DOE factorial design there were no missing values, see Table 7.15. Because no gaps exist, the experimental region is 100% of the potential region for this DOE design.

Referring to summary of effects Table 6.15, **for main effects** four variables are significant for 1st order, two variables are significant for the 2nd order interaction model and no significant variables for the 3rd order interaction model.

When including **2nd and 3rd order interactions**, no significant variables were present.

For the 2** (7-2) resolution IV (32 RUNS) DOE factorial design, no additional analysis is necessary for missing value impact on DOE design outcomes.

6.4 PROPOSED DOE MODEL

Referring to Table 6.15, the resolution IV 2** (7-2) design was selected (Box *et al*., 1978:408) which provided runs with no missing values for each of the 32 runs. The main reason for selecting this design for the study going forward is that no missing values

were produced and the risk of using false factor interactions caused by missing values is minimal.

This design is typically used in a processing environment where financial and processing constraints prevent full factorial experimental runs to measure effects of low and high settings for independent variables to determine which independent variables contribute the greatest effect on a dependent variable as the operating level is deliberately changed.

Design: 2(7-2) design. Resolution : IV**							
Run number	**Mix discharge temp**	**Cool begin temp**	**Actual cool time**	**Actual dump temp**	**Actual tamp pressure**	**Actual extrusion rate**	**Actual extrusion speed**
1	-1	-1	-1	-1	-1	1	1
2	1	-1	-1	-1	-1	-1	-1
3	-1	1	-1	-1	-1	-1	-1
4	1	1	-1	-1	-1	1	1
5	-1	-1	1	-1	-1	-1	1
6	1	-1	1	-1	-1	1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	-1	-1	-1	1
9	-1	-1	-1	1	-1	-1	-1
10	1	-1	-1	1	-1	1	1
11	-1	1	-1	1	-1	1	1
12	1	1	-1	1	-1	-1	-1
13	-1	-1	1	1	-1	1	-1
14	1	-1	1	1	-1	-1	1
15	-1	1	1	1	-1	-1	1
16	1	1	1	1	-1	1	-1
17	-1	-1	-1	-1	1	1	-1
18	1	-1	-1	-1	1	-1	1
19	-1	1	-1	-1	1	-1	1
20	1	1	-1	-1	1	1	-1
21	-1	-1	1	-1	1	-1	-1
22	1	-1	1	-1	1	1	1
23	-1	1	1	-1	1	1	1
24	1	1	1	-1	1	-1	-1
25	-1	-1	-1	1	1	-1	1
26	1	-1	-1	1	1	1	-1
27	-1	1	-1	1	1	1	-1

Design: 2**(7-2) design. Resolution : IV							
Run number	**Mix discharge temp**	**Cool begin temp**	**Actual cool time**	**Actual dump temp**	**Actual tamp pressure**	**Actual extrusion rate**	**Actual extrusion speed**
28	1	1	-1	1	1	-1	1
29	-1	-1	1	1	1	1	1
30	1	-1	1	1	1	-1	-1
31	-1	1	1	1	1	-1	-1
32	1	1	1	1	1	1	1

Table 6.16: Summary of standard design standard 2 (7-2) resolution IV design**

Table 6.16 represents the seven selected independent variables for the DOE analysis. The 32 runs sliced the database into high and low values to represent each run respectively, based on median values. For each experimental run, the average extrusion pressure was the dependent variable.

The 2** (7-2) resolution IV (32 RUNS) DOE factorial design was evaluated for the estimated period only because the goal is to determine how accurately the estimated period of the historical data, based on DOE regression, predicts the validation period. For a predictive model the DOE regression model was run to determine the coefficients for a predictive model.

	Regr. Coefficients; Var.:Ave Average; R-sqr=.79661; Adj:.73729 (Thesis data - First period) DB ANN) 2**(7-2) design; MS Residual=.0041222 DV: Ave Average			
Factor	Regressn Coeff.	Std.Err.	t(24)	p
Mean/Interc.	2.922	0.011	257.5	0.000
(1)Mix discharge temp (x)	0.016	0.011	1.4	0.177
(2)Cool begin temp (x)	-0.009	0.011	-0.8	0.440
(3)Actual cool time (x)	-0.003	0.011	-0.3	0.775
(4)Actual dump temp (x)	0.032	0.011	2.8	0.010
(5)Actual tamp pressure (x)	-0.091	0.011	-8.0	0.000
(6)Actual extrusion rate (x)	-0.040	0.011	-3.6	0.002
(7)Actual extrusion speed (x)	-0.030	0.011	-2.6	0.014

Table 6.17: DOE regression summary estimated period 32 runs – Average pressure

Referring to Table 6.17, independent variables 1-3 had no significant effect on average extrusion pressure for the estimated period. Only variables 4-7 are significant, same as for main effects. The R-square or the coefficient of determination of 0.796 shows that the residual variability is lower than the explained variability, which indicates a good predictive model.

All the DOE regression coefficients, irrespective of whether the independent variable is significant or not, are used for predicting the next production period to evaluate accuracy. The reason is that, from a practical perspective, all of the independent variables form part of the process output. Non-significant variables contribute negligible effects on the dependent variable, and therefore they are kept as part of this study.

The DOE regression prediction model with the average pressure as the dependent variable performed the best across all evaluations. For this model, the average extrusion pressure will be used as the dependent variable as the evaluating statistic.

6.5 PROPOSED PREDICTION MODEL – DOE REGRESSION WITH AVERAGE STATISTIC

The goal was in the first place to identify which significant independent variables can be used for a DOE design to predict the next period of the same process, and secondly, to adjust these independent variables beyond the calculated minimum and maximum value of the model DOE runs for the next period, then evaluating the effect to direct process improvement.

The proposed DOE regression prediction model for process development is:

- $y = 2.922313 + 0.015804*x_1 - 0.008908*x_2 - 0.003288*x_3 + 0.031755*x_4 – 0.090733*x_5 - 0.040383*x_6 - 0.029961*x_7$.
- Variable 1-7 corresponds with Factor 1-7 in the regression statistical summary Table 6.18 below.

Factor	Regr. Coefficients; Var.:Ave Average (y); R-sqr=.79661; Adj:.73729 (Thesis data - First period) DB ANN) 2**(7-2) design; MS Residual=.0041222 DV: Ave Average			
	Regressn Coeff.	Std.Err.	t(24)	p
Mean/Interc.	2.922	0.011	257.5	0.000
(1)Mix discharge temp (x)	0.016	0.011	1.4	0.177
(2)Cool begin temp (x)	-0.009	0.011	-0.8	0.440
(3)Actual cool time (x)	-0.003	0.011	-0.3	0.775
(4)Actual dump temp (x)	0.032	0.011	2.8	0.010
(5)Actual tamp pressure (x)	-0.091	0.011	-8.0	0.000
(6)Actual extrusion rate (x)	-0.040	0.011	-3.6	0.002
(7)Actual extrusion speed (x)	-0.030	0.011	-2.6	0.014

Table 6.18: DOE regression summary estimated period 32 runs – Average pressure

6.6 DISCUSSION OF PROPOSED MODEL

All seven independent variables are part of the regression model, irrespective if they are significant or not. This defies the basis for traditional regression to use only the significant variables, which have the highest correlation with a dependent variable. The non-significant independent variables are critical for this process and cannot be discarded based on statistical significance. Their low impact on the process according to the model should then have a minimal effect if used, and can be viewed as trivial.

An objective of this study is also to identify which independent variables correlate the highest and are significant towards a selected dependent variable, and then only vary those operating levels for process development. For this reason, only operating levels are changed for only independent variables 4, 5 and 6 for process development when applying the proposed model.

The final DOE run selection of which runs to test, whether for low or high outcomes for process development, is a financial decision together with the lowest risk impact on production continuity and throughput. This is not part of this study and should be part of future work, but is critical for selecting the final experimental runs.

6.7 CONCLUSIONS

Using the validation process, variables 4 to 7 support the DOE proposed model for process improvement. Variables 1 to 3 are either not significant or need more evaluation for understanding.

Because the dependent variable is an average value, variables may be discarded as non-significant due to small variances and narrow confidence intervals, and therefore should be treated cautiously before excluding from a predictive model.

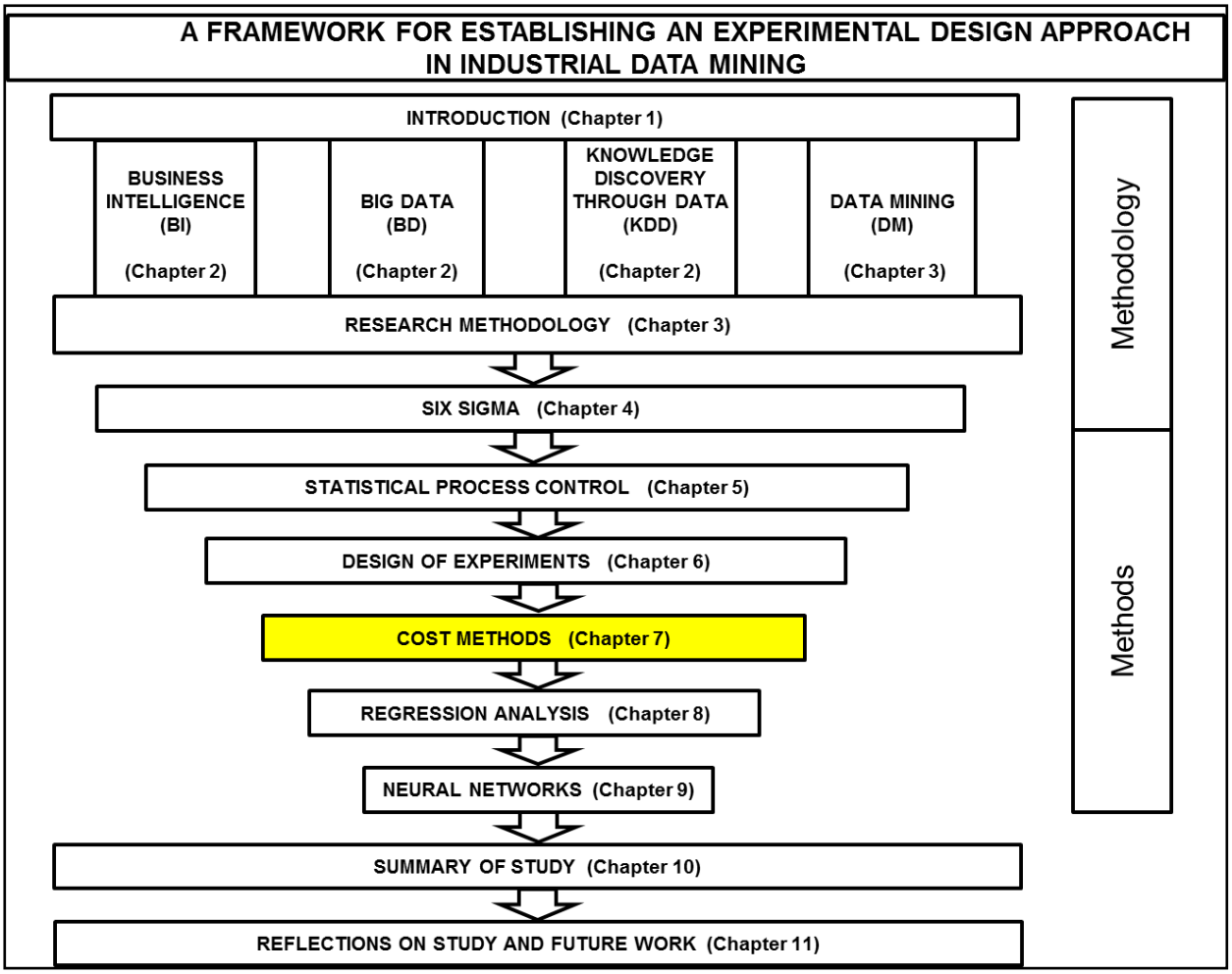
The selection of minimum and maximum values seems to have major effect on significant effects for variables. The impact of boundary changes to interaction sensitivity needs careful analysis.

The 2** (7-0) resolution FULL (128 RUNS) DOE factorial design and the 2** (7-1) resolution VII (64 RUNS) DOE factorial design provide two-way and three-way interactions, which is an advantage of DOE modelling; but for this study, these interactions include missing values which may mask true interactions. For this reason, these two models were not selected for going forward with this study.

Comparing significant main effects for the three models, the 2** (7-2) resolution IV (32 RUNS) DOE factorial design provided four different significant variables than the 2** (7-0) resolution FULL (128 RUNS) DOE factorial design and the 2** (7-1) resolution VII (64 RUNS) DOE factorial design (refer to Table 6.15). In theory, similar main effects should be evident for all models. For this reason, because the 2** (7-2) resolution IV (32 RUNS) DOE factorial design have no missing values, the significant independent variables may be more credible and reduce the risk of masked factor effects due to missing values.

CHAPTER 7

COST METHODS



7.1 INTRODUCTION

Lochner and Matar (1990:20) discuss some basic elements of quality control that include DOE, as the total loss generated by a product to society will be known. Continuous improvement and cost reductions are necessary for staying in business, quality improvement programs must include the reduction of product performance variance as well as the optimisation of the central target values. Deviation from the target results is expressed as a loss to the customer expressed in the quadratic quality/loss function, quality and cost of a product are determined by the engineering design and manufacturing process.

Measuring the cost for a process whilst trying to improve process and product quality is usually put on the back burner because analysts are mostly focused on improving

throughput and product quality as a quick fix with little regard to the associated costs involved. The associated cost for this improvement becomes apparent at a later stage as a huge surprise. For this study the associated cost in determining the best experimental run is important because not only does each experimental run deviate from the standard, it also has a cost implication element for each run.

The validation period of the database was used as the comparative base to evaluate the DOE and normal regression prediction accuracy compared to the estimated period in terms of DOE target level accuracy as well as the associated cost and signal to noise ratio when product quality moves away from the target value for each experimental run or process condition.

This a guide of how cost analysis could be used to guide analysts in finding the sound solutions.

7.2 DESCRIPTION OF COST METHODS

Three quadratic quality loss functions, each with its comparative signal to noise ratio for each experimental run, were used for determining the total cost for deviating from the process target standard. The goal is to minimize the variability in the product's performance in response to noise factors that influence the product performance while maximizing the variability in response to noise factors.

Noise factors are not under the control of the operator of a product and normally form part of the environmental variations that influence a process. Signal factors are controlled by the operator, which influences the product directly. The goal of quality improvement is to find the best settings of factors under your control that are involved in a production process, in order to maximize the S/N ratio because by this, the factors in the experiment will represent the real control factors.

Diagram 7.1 represents the traditional cost model that is based on the principle that a company only starts to lose money if products are produced outside the process specifications. This principle was used for many years until Taguchi changed the belief that money is lost as the process starts to deviate from the set average specification.

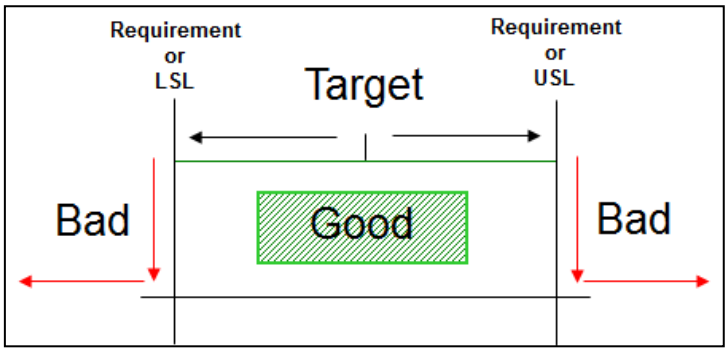


Diagram 7.1: Traditional cost diagram

Quality enthusiasts acknowledge that Taguchi changed the thinking of using the traditional loss function, Diagram 7.1, in designing three new quadratic quality loss functions and their respective signal to noise ratio's, that represents loss of quality more realistically. These new loss functions are as follows:

- Nominal the best
- Smaller the better
- Larger the better

7.2.1 Nominal the best

For nominal the best, there is a defined target quality value for the product that has to be achieved. This quality target is set by the manufacturer to ensure that the production process produces the required quality set by the consumer. There is a specified upper and lower processing limit beyond which the product will be scrapped or re-worked. The target quality value is the middle point between these two limits. Quality is in this case is defined in terms of deviation from the target value.

The quadratic equation that describe the loss function of one unit of product as it deviates from the target value is:

$$L = k(y - m)^2$$

Where:

L = Financial Loss expressed in a currency

y = Output Value

m = Target Value of Output

k = Proportionality constant, representing the cost factor associated with loss.

k for nominal-the-best is defined as:

$$k = \frac{A_0}{\Delta_0^2}$$

A_0 = Consumer Currency Loss

Δ_0 = Maximum deviation from the quality target value allowed by Consumer

The S/N ratio equation for nominal the best is a fixed signal value (nominal value), and the variances around this value are the result of noise factors:

Eta (S/N) = 10 * $\log_{10}$ (Mean^2/Variance)

Use the Eta ratio in combination with the nominal value when a target quality characteristic is pursued. See Diagram 7.2 for a graphical representation of the Nominal the best loss function.

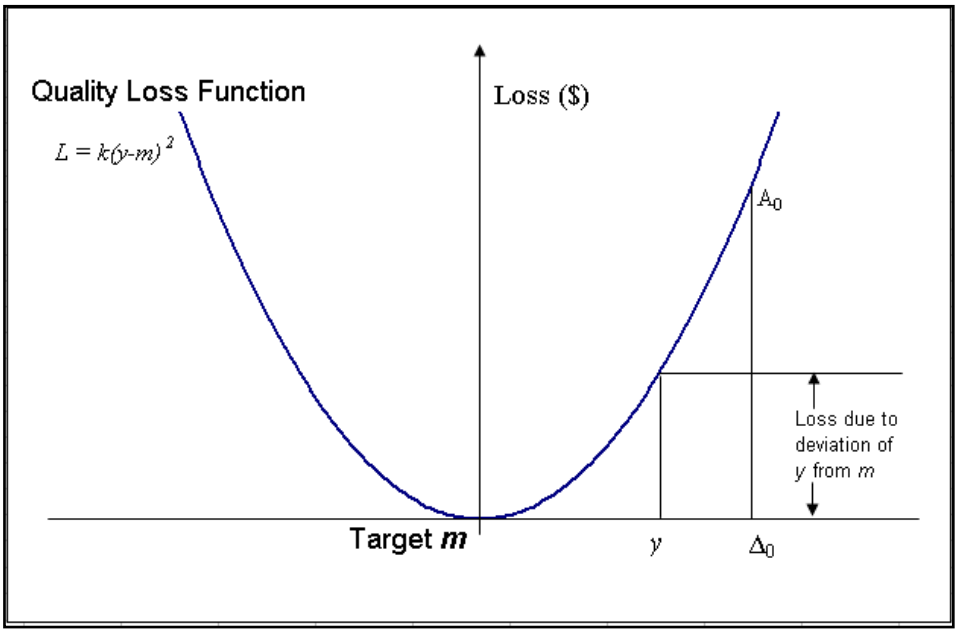


Diagram 7.2: Loss function for Nominal the best (Sharman *et al*., 2007)

7.2.2 Smaller the better

For Smaller the better, the ideal target value is zero. This quality target requirement is set by the consumer to ensure that the production process produces the required quality. Here, the ideal value is zero and as the value increases the subsequent loss increases due to the progressively worsening of product performance until it reaches the upper limit where the product will be scrapped or re-worked. Quality is in this case is

defined in terms of deviation from the target value which is zero. Minimizing this characteristic as much as possible would produce a more desirable product.

The quadratic equation for the loss functions of one unit of product:

$$L(y) = k(y)^2$$

Where:
L = Financial Loss expressed in a currency
y = Output Value
k = Proportionality Constant , representing a cost factor associated with loss.
y = Output Value

k for smaller the better is defined as:

$$k = \frac{A_0}{y_0^2}$$

A_0 = Consumer Currency Loss
y_0 = Maximum deviation from the quality target value allowed by Consumer.

The S/N ratio equation for Smaller the better is when you want to minimize the occurrences of some undesirable product characteristics.

$Eta(S/N) = -10 * log_{10} [(1/n) * S(y_i^2)]$

Eta is the resultant S/N ratio where *n* is the number of observations on a specific product, and *y* is the respective quality characteristic. Thus, maximizing this ratio will increase quality.

See Diagram 7.3 for a graphical representation of the smaller the best loss function.

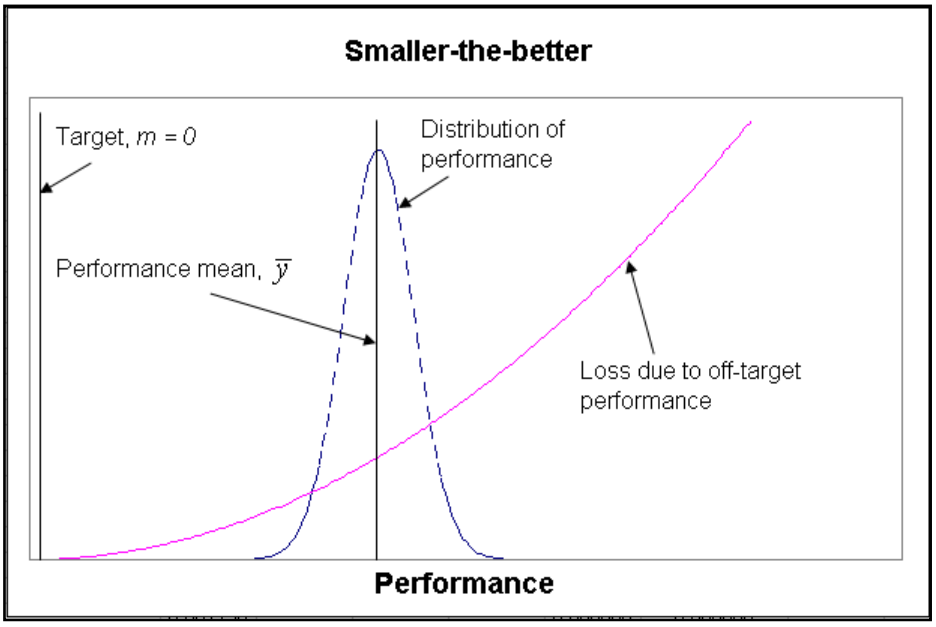


Diagram 7.3: Loss function for Smaller the better (Sharman *et al.,* 2007)

7.2.3 Larger the better

The larger the better characteristic is just the opposite of the Smaller the better characteristic. For this characteristic, it is preferred to maximize the result, and the ideal target value is infinity.

The quadratic equation for the loss functions of one unit of product:

$$L = \frac{k}{y_0^2}$$

Where:

k = Proportionality Constant , representing a cost factor associated with loss.

y_0 = Minimum deviation from the quality target value allowed by Consumer.

k for larger the better is the same as for smaller the better:

$$k = \frac{A_0}{y_0^2}$$

The only difference between the two is the definition of y_0.

Signal to noise ratio equation for Larger the better is:

$Eta(S/N) = -10 * log_{10} [(1/n) * S(1/y_i^2)]$

S/N ratio for Smaller the better is similar with only the exception that the reciprocal of the quality characteristic is used.

See Diagram 7.4 for a graphical representation of the Larger the better loss function.

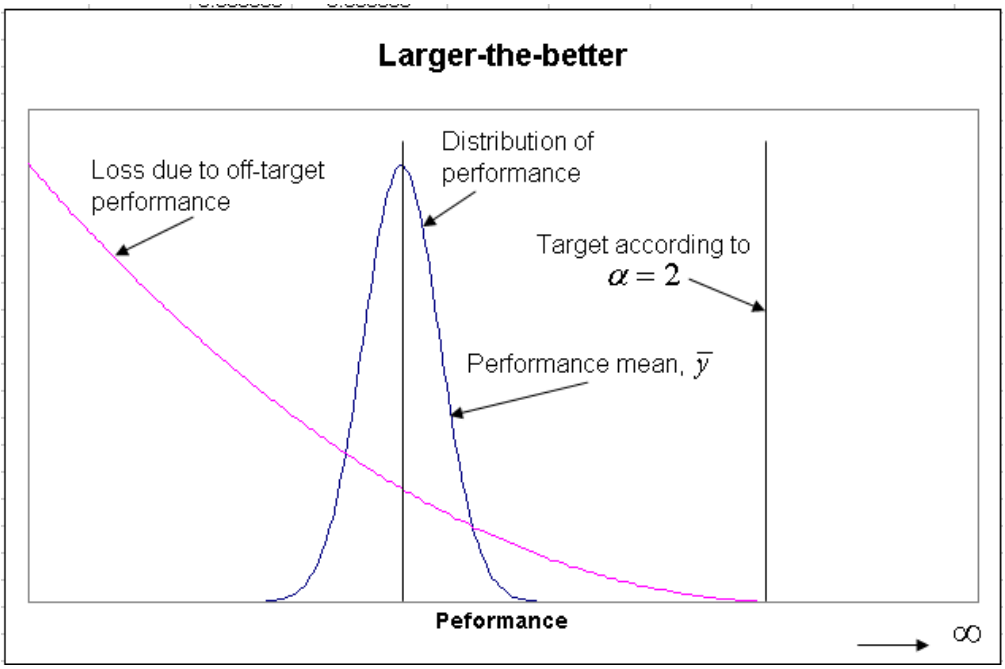


Diagram 7.4: Larger the better loss function (Sharman *et al.*, 2007)

7.3 DATABASE ANALYSIS

7.3.1 DOE target value analysis

The 32 experimental runs for both periods were ranked from lowest to highest. The ranked DOE runs are used to determine which DOE experimental runs or conditions are the same across the two periods in this study. Comparing associated ranked outcomes eliminates bias towards outcome levels for the test period because it simply compares low and high outcomes. For this study the level of process performance is of lesser importance than outcome levels.

DOE runs are ranked from the lowest to highest output value for each run. For evaluation purposes, the associated DOE outcomes were divided arbitrarily into four zones, namely best, good, fair and poor. For this study, the focus is not to compare individual ranked runs but only to compare the zones, namely best, good, fair and poor across the three periods depicting in Table 7.5. **The main purpose of this evaluation is to evaluate if the prediction models will at least distinguish amongst different**

levels of outcomes irrespective of the target value. By colour coding the associated ranked outcomes that fall within the selected four zones a graphical representation emerges showing where each experimental run fits as its represented outcome deviates from the target outcome.

Each of the four zones with the representative target outcomes was colour coded as:

- Best = green (3.25 – 3.75)
- Good = yellow (3.0 – 3.25, 3.75 – 4.0)
- Fair = orange (2.75 – 3.0, 4.0 – 4.25)
- Poor = red (2.5 – 2.75, 4.25 – 4.5)

For this production process, the closer an outcome moves towards the red zone, the higher the risk becomes either to re-work or to scrap the represented product for that specific outcome. There is also an associated cost to the consumer based on the quadratic loss function as an outcome moves towards the red zone.

A summarized graphical representation follows showing to which zone each applicable outcome belongs for each category. Table 7.1 and graphs 7.1 - 7.3 below represent the impact the position of outcomes for each period have on the risk of producing non-conforming products.

Dependent variable - Average pressure										
	First period				Second period				Second period - Predicition	
	1 st Period DOE Base		1 st Period DOE regression	1 st Period regression	2 nd Period DOE Base		2 nd Period DOE regression	2 nd Period regression	2nd period prediciton DOE regression	2nd period prediciton normal regression
DOE run number	Outcome	S/N	Outcome	Outcome	Test outcome	S/N	Test outcome	Test outcome	Test outcome	Test outcome
1*	2.962	19.506	2.907	2.962	3.327	21.160	3.256	3.169	2.907	2.878
2*	2.924	20.481	3.079	3.003	3.487	17.857	3.504	3.406	3.079	3.000
3*	3.160	24.902	3.030	3.081	3.661	18.372	3.704	3.533	3.030	**3.011**
4*	2.908	23.780	2.921	2.926	3.316	19.632	3.357	3.195	2.921	**2.863**
5*	3.044	20.796	2.981	3.011	3.332	17.517	3.339	3.271	2.981	2.985
6*	2.983	20.316	2.992	2.958	3.398	23.434	3.327	3.254	2.992	2.919
7*	2.965	21.712	2.943	2.963	3.648	18.932	3.527	3.451	2.943	2.913
8*	3.020	23.627	2.995	3.026	3.518	18.904	3.440	3.451	2.995	2.997
9*	3.080	20.739	3.111	3.069	3.498	19.246	3.560	3.426	3.111	**3.040**
10	3.004	20.933	3.003	2.990	3.172	19.811	3.213	3.144	3.003	2.929
11*	2.970	22.527	2.953	2.972	3.351	18.823	3.413	3.227	2.953	2.901
12*	3.184	20.341	3.125	3.056	3.804	17.541	3.662	3.549	3.125	**3.062**
13*	2.940	18.801	3.024	3.015	3.274	18.773	3.383	3.159	3.024	2.981
14*	3.088	21.086	3.077	3.084	3.195	17.300	3.297	3.301	3.077	3.056
15*	2.964	20.662	3.027	3.036	3.510	18.122	3.497	3.483	3.027	**3.046**
16*	3.015	22.676	3.038	3.012	3.474	21.736	3.485	3.400	3.038	2.961
17	2.810	18.426	2.786	2.779	3.034	19.012	3.098	3.152	2.786	**2.698**
18	2.843	21.628	2.838	2.794	3.021	20.797	3.012	3.128	2.838	**2.760**
19	2.654	22.388	2.789	2.773	3.196	18.027	3.212	3.257	2.789	**2.785**
20	2.814	26.946	2.800	2.728	3.165	18.970	3.200	3.183	2.800	2.647
21	2.862	19.716	2.860	2.886	3.317	19.692	3.182	3.149	2.860	2.629
22	2.720	20.646	2.751	2.763	2.806	19.235	2.835	2.950	2.751	2.624
23	2.726	21.426	2.702	2.775	2.937	19.298	3.035	3.119	2.702	**2.661**
24	2.857	21.806	2.874	2.849	3.148	20.040	3.283	3.325	2.874	2.679
25	2.896	23.970	2.870	2.894	3.133	20.459	3.068	3.113	2.870	2.767
26	2.940	23.384	2.881	2.852	3.077	17.516	3.056	3.114	2.881	2.677
27	2.769	26.197	2.832	2.814	3.317	18.965	3.256	3.213	2.832	**2.677**
28*	2.893	22.841	2.884	2.856	3.183	19.483	3.170	3.251	2.884	2.793
29	2.786	21.132	2.783	2.783	2.955	18.550	2.891	3.014	2.783	**2.742**
30	3.019	18.981	2.955	2.922	3.134	17.515	3.140	3.233	2.955	2.854
31	2.917	21.980	2.906	2.865	3.273	19.413	3.340	3.365	2.906	2.735
32	2.802	22.699	2.797	2.801	3.074	18.352	2.992	3.155	2.797	**2.728**

Table 7.1: DOE run ranking Average pressure

For the estimated period, a total 21 of the 32 experimental runs have the same colour, subdivided into sixteen fair and five good codes across the estimated period DOE base, estimated period DOE regression and estimated period regression. It means that 66% of the same colour codes for the DOE base for the estimated period are also predicted by both the DOE regression and normal multiple regression. It indicates a good stability for the data in the base DOE for the estimated period, not necessarily a low risk of potential re-working or scrapping.

For the validation period, 19 of the 32 experimental runs have the same colour, subdivided into twelve best, six good and one fair codes across the validation period DOE base, validation period DOE regression and validation period regression. It means that 59% of the same colour codes for the DOE base for the validation period are also

predicted by both the DOE regression and normal multiple regression. It indicates a fair stability for the data in the base DOE for the validation period, not necessarily a low risk of potential re-work or scrap.

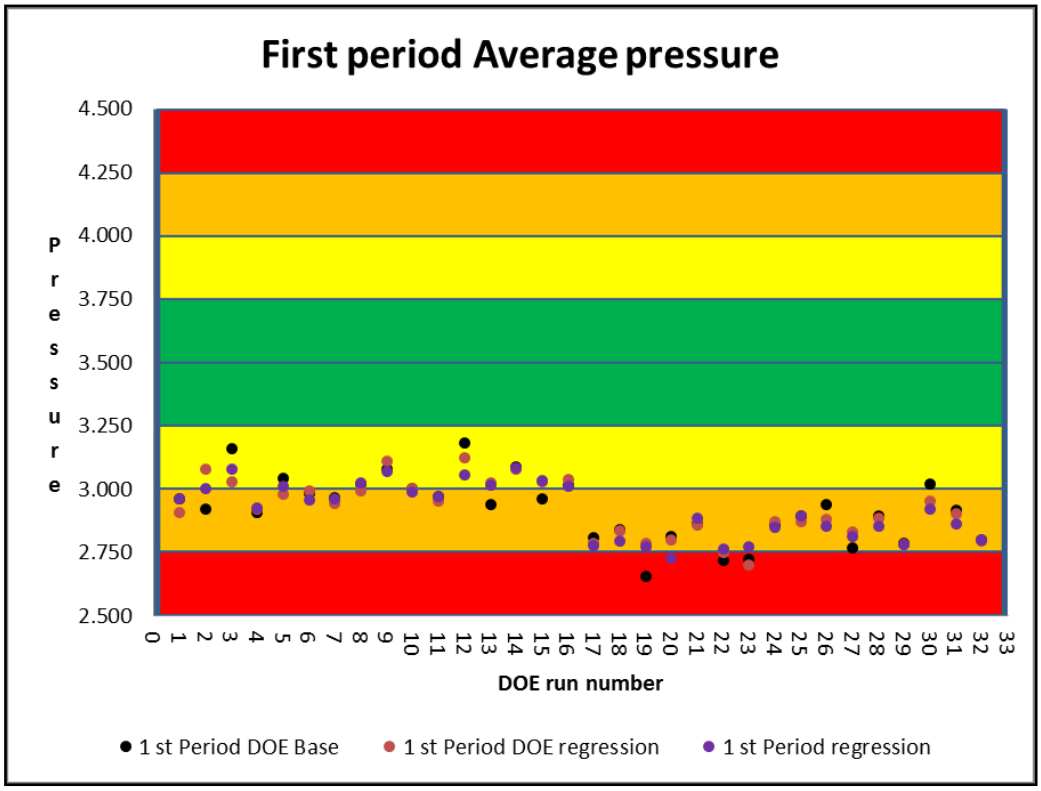
For the validation period, we choose the experimental runs that have similar risk profiles across all three periods. These experimental runs are selected on the following colour zone risk criteria:

- Only green
- Green and yellow
- Green, yellow and orange
- Green and orange
- No red
- No yellow and orange

The above selection criteria minimize the risk for not producing out of specification products as well as reduce the cost passed onto the consumer.

Experimental runs marked with an asterisk in Table 7.1 comply with the above criteria. They are 1 – 9, 11 – 16 and 28. These sixteen runs are the runs with the lowest non-conformance risk and cost profile.

Below are the graphical risk profiles for each of the three periods represented in Table 7.1.

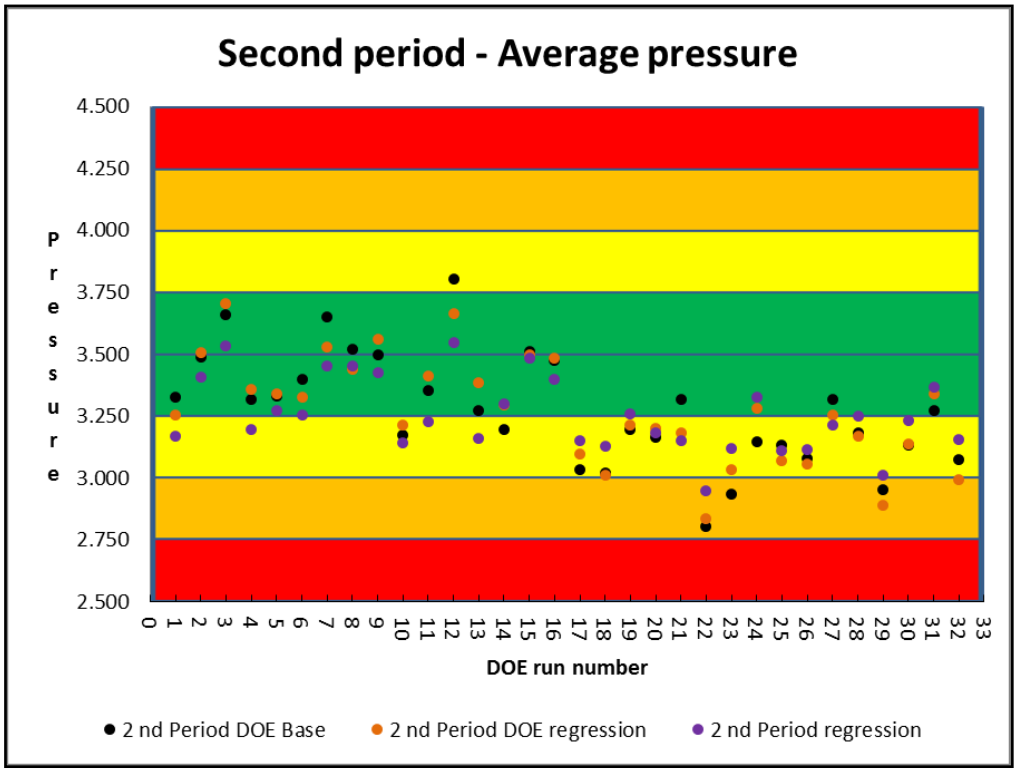


Graph 7.1: Zone risk profile of DOE outcomes for period 1

Period 1 shows that the process operating level for the DOE runs for each operating category (estimated period base DOE, estimated period DOE regression and estimated period regression) is below target and remains fairly stable between good and fair with a small portion in the poor zone.

Al three categories are close to one another indicating a stable production process, which is a prerequisite when used for future prediction.

The associated potential costs for operating in these zones are discussed in the next section.



Graph 7.2: Zone risk profile of DOE outcomes for validation period

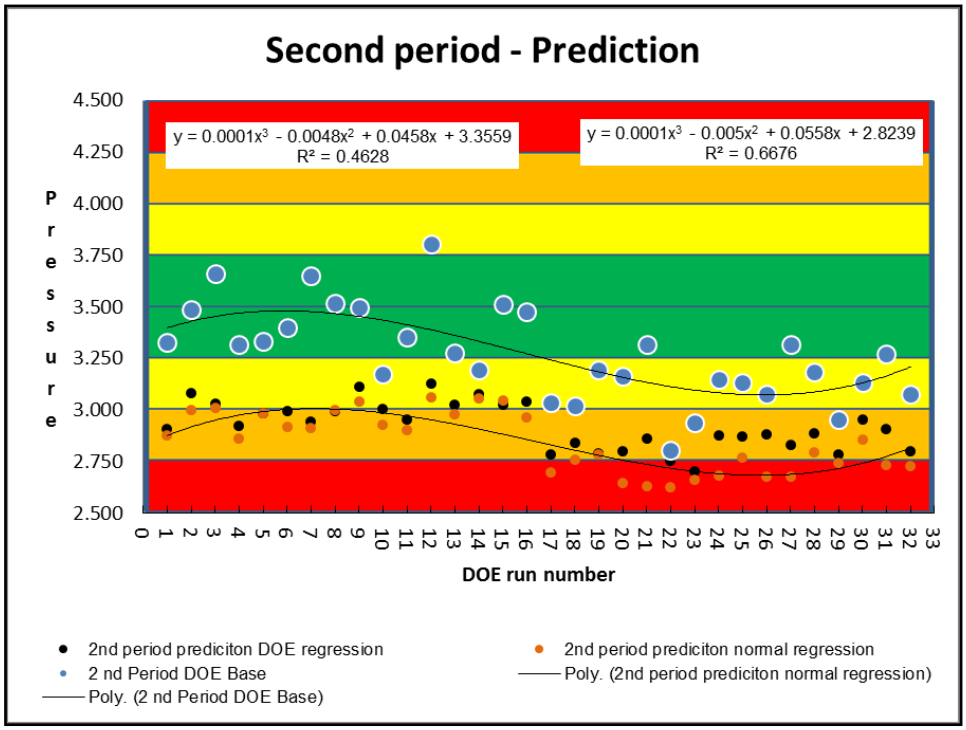
Validation period shows that the process operating level for the DOE runs for each operating category (validation period base DOE, validation period DOE regression and validation period regression) shifted towards target and also shows a shift form the best to the good and fair zones based on sequential experimental run number.

Although a shift occurred approximately halfway through the experimental runs, the processing level remains stable. This shift was due to a raw material change discussed earlier.

For validation period, the three categories are also close to one another but not as close as the estimated period. The shift in the processing level due to a raw material shift influenced prediction variation for validation period DOE regression and validation period regression.

The production process is still stable and within specification but with a higher variability, therefore it can still be used for predicting validation period.

The associated potential costs for operating in these zones are discussed in the next section.



Graph 7.3: Zone risk profile of DOE predicted outcomes for validation period

Predicting the validation period with DOE regression and normal multiple regression with regression equations that used the data for estimated period shows a predicted operating level for the validation period similar than the estimated period.

Although the predicted levels are similar, the normal regression predicts a large portion in the red zone which increases the risk of producing out of specification products.

For comparative purposes, the actual DOE base outcomes were also plotted on this graph. The prediction model that is closest to the actual, even with the process shift, should be considered as the better predictive model. In this case, it is the DOE regression model.

The shift that is identified in Graph 7.2 is also evident in Graph 7.3 for the validation periods and the actual DOE base outcomes. Two fitted trend lines through the predicted data and the actual DOE outcomes show the same operating pattern but on a different level.

If an adjustment is made towards the actual DOE base data, the best predictor is still the DOE regression with a close approximation of the validation period.

The associated potential costs for operating in these zones are discussed in the next section.

7.3.2 Cost analysis

A primary objective of this study was to compare the estimated period DOE base to the validation period DOE base, which is a prediction, using DOE and multiple regression equations that are based on the estimated period DOE database. The success will be how close the estimated period replicates the same experimental outcomes for the next period.

These ranked experimental DOE run outcomes were divided arbitrarily into four groups, following the sequential ranked outcomes from low to high, *inter alia*, a best, good, fair and a poor group. As outcomes deviate from the production target, the colours give a risk profile and associated loss passed on to the consumer for producing non-conforming products.

The grouping is the basis for evaluating the associated cost, which is passed on to the consumer in terms of risk producing non-conforming products as well as to select experimental runs that provide the best independent variable combination for process improvement. Taguchi states “We measure the quality of a product in terms of the total loss to society due to functional variation and harmful side effects” (Phadke, 1989:4). To measure this loss and outcomes deviated from a target outcome, these four groups were colour coded as:

- Best = green
- Good = yellow
- Fair = orange
- Poor = red

For the study we chose Nominal the best loss function with its applicable signal to noise ratio for calculating potential cost occurred by the customer. The reason for choosing this function is that the aim for the production process is to start at a specific target value for every run and then to maintain that level during a run. Traditionally the start of each process run is irrelevant as long as it starts within the specified engineering tolerances. From a product quality point of view, it is better to start close to the centre or

lower than the target, because that the subsequent processes are more tolerant to products produced at a lower than a higher output, even if both are within engineering tolerances.

Diagram 7.5 shows the cost impact by the Nominal the best quality loss function as product quality moves away from the target.

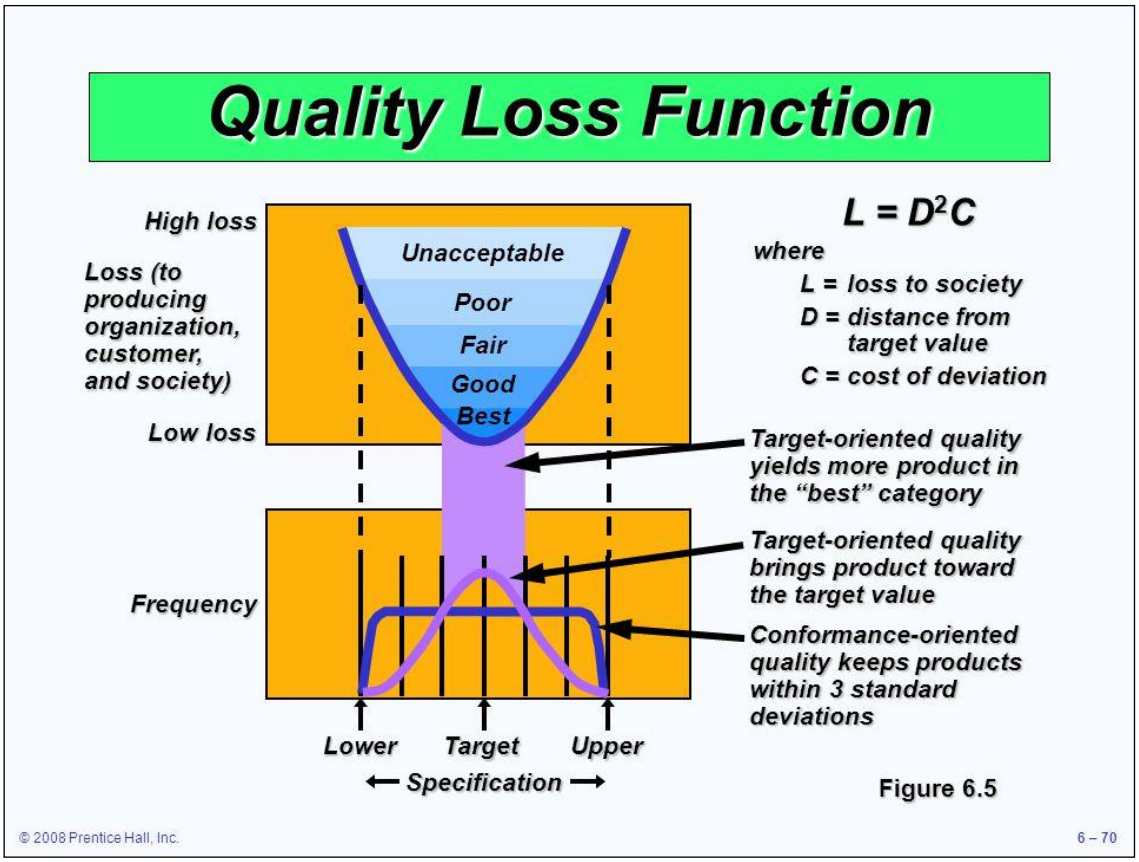
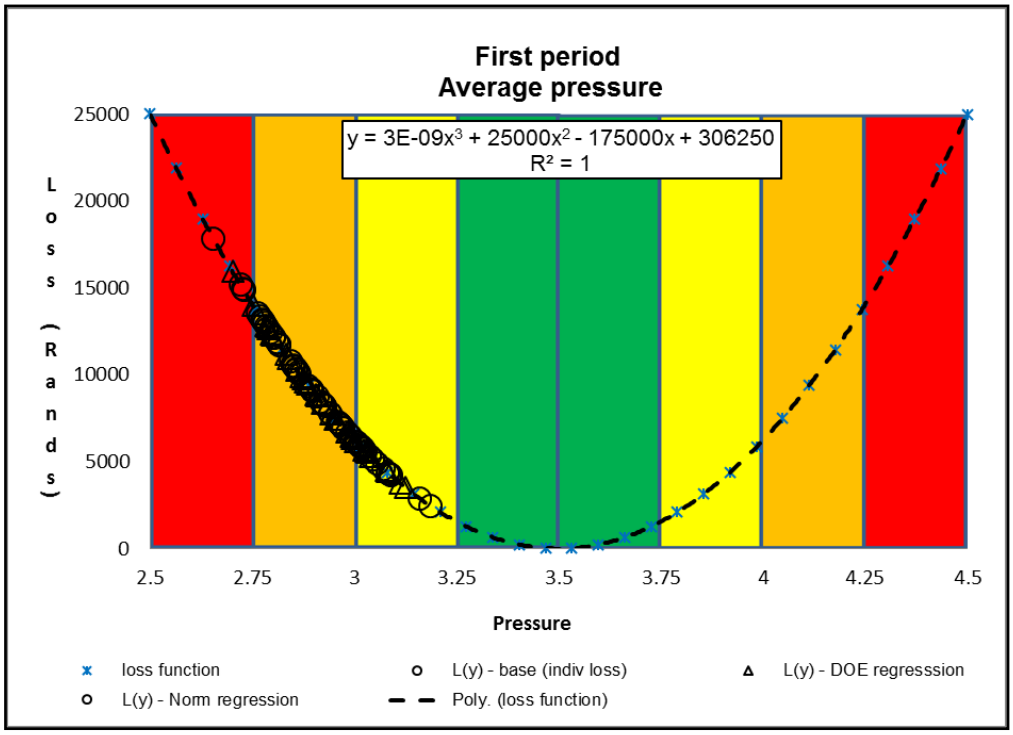


Diagram 7.5: Quality loss function (Baran, 2011)

When producing a product there is always a risk of producing non-conforming products as well as applicable costs towards re-work or scrapping products. For this study, the focus will be to evaluate that risk in terms of coloured zones that represent increased risk as an outcome deviates from the target value across period 1, validation period and predicted validation period. The quality loss function that is superimposed on these zones reflects the cost passed to the customer as outcomes deviate from the target value. It forms an integral part of the decision process because the cost component for each corresponding experimental run for both periods in combination with probable experimental runs must be part of the final analysis.

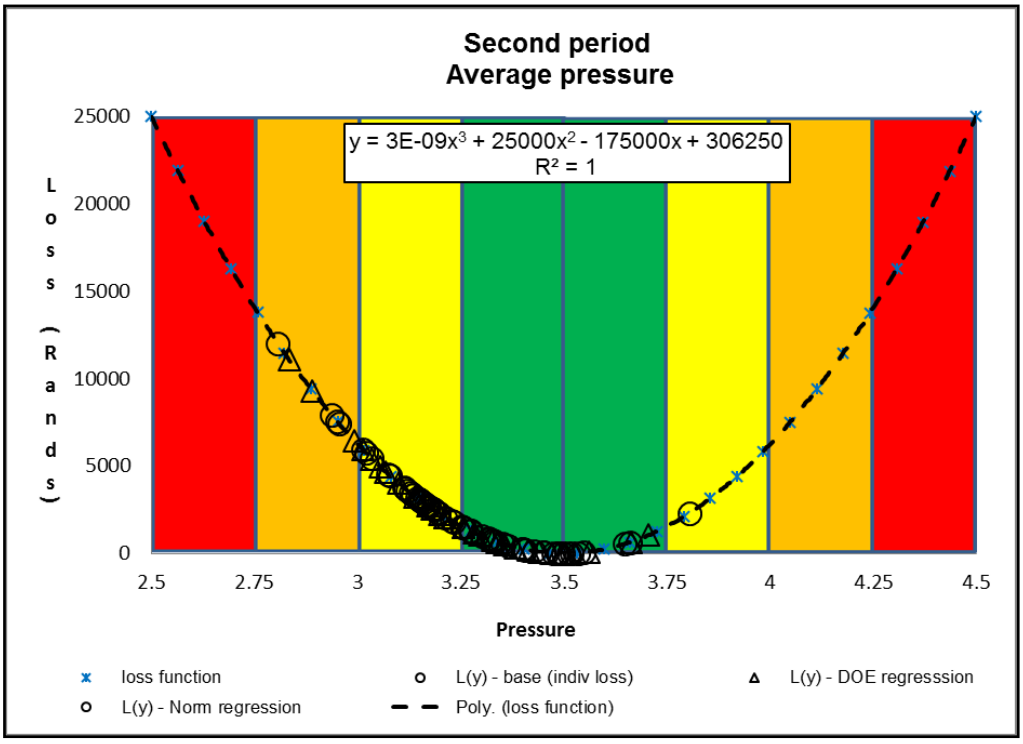


Graph 7.4: Zone cost and risk profile of DOE outcomes for period 1

For the zone cost and risk profile cost analysis the cost associated to the DOE ranking and zone risk profile for the estimated period is:

- DOE base is approximately R 279 000
- DOE regression is approximately R 276 600
- Multiple regression is approximately R 282 400

The cost is similar for all three categories but is mostly on the lower side of the target value which in terms of product performance is better than operating towards the high side of the target value.



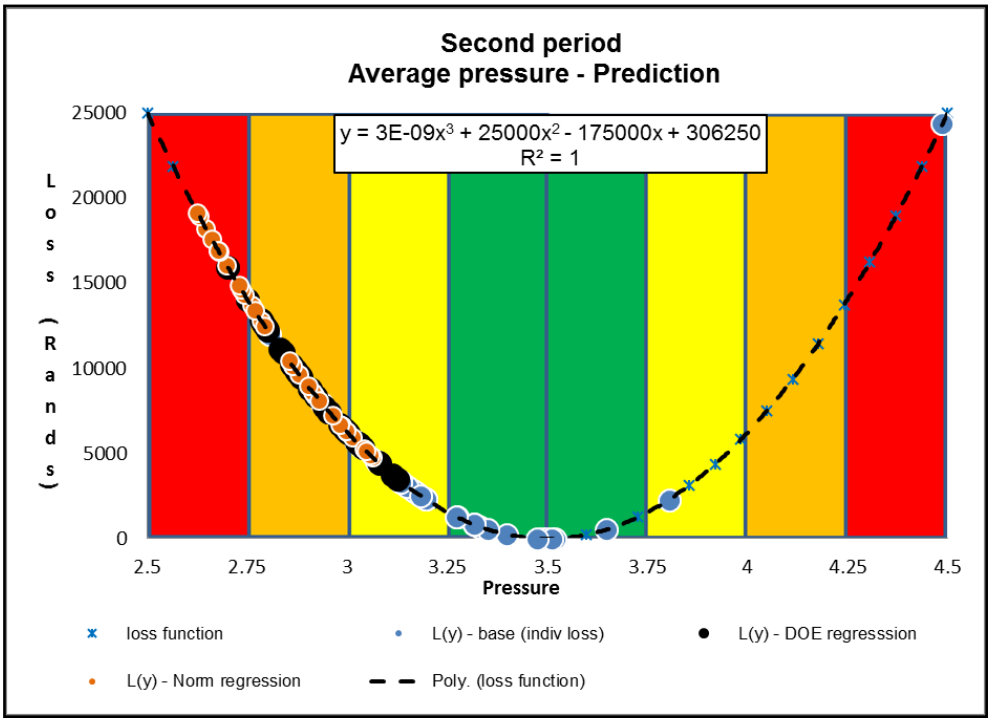
Graph 7.5: Zone cost and risk profile of DOE outcomes for validation period

For the zone cost and risk profile cost analysis for the validation period, the cost associated to the DOE ranking and zone risk profile for the validation period is:

- DOE base is approximately R 81 000
- DOE regression is approximately R 76 800
- Multiple regression is approximately R 65 300

The costs are also similar to the estimated period but much lower because the experimental outcomes are close to the target value.

The lower costs for the validation period in terms of product performance have a small risk of producing out of specification.



Graph 7.6: Zone cost and risk profile of DOE predicted outcomes for validation period

For the zone cost and risk profile cost analysis for the predicted validation period the cost associated to the DOE ranking and zone risk profile for the second validation period is:

- For DOE regression is approximately R 276 600
- For multiple regression is approximately R 360 700

The costs are significantly different for the two prediction categories but are significantly higher than the estimated and validation period.

The difference in costs is explained by the fact that both the prediction models were calculated form the estimated period's data, but the actual process in the validation period changed. However, the important issue is that the DOE regression model caused significantly lower costs than for the multiple regressions.

Table 7.2 below is a cost summary for the three periods representing the unseen cost handed to the consumer even if the process is within specification.

Cost summary table							
First period			Second period			Second period -	
Doe base	DOE regression	Multiple regression	DOE base	DOE regression	Multiple regression	Prediction - with 1st period DOE regression	Prediction - with 1st period multiple regression
R 279 000	R 276 600	R 282 400	R 801 000	R 76 800	R 65 300	R 276 600	R 360 700

Table 7.2: Cost summary

7.3.3 Signal to noise ratio analysis

This ratio measures how the outcome varies relative in meeting the target value influenced by different noise factors. In Table 7.1, the experimental runs 1 – 9, 11 – 16 and 28 were selected based on cost profiles across the three periods. Signal to noise ratios for of these runs have been calculated, for cost factors must be taken into account for final selection. The ideal is to select runs with high ratios because the risk of being influenced by external variation is minimized.

Dependent variable - Average pressure										
	First period				Second period				Second period - Predicition	
	1 st Period DOE Base		1 st Period DOE regression	1 st Period regression	2 nd Period DOE Base		2 nd Period DOE regression	2 nd Period regression	2nd period prediciton DOE regression	2nd period prediciton normal regression
DOE run number	Outcome	S/N	Outcome	Outcome	Test outcome	S/N	Test outcome	Test outcome	Test outcome	Test outcome
20	2.814	26.946	2.800	2.728	3.165	18.970	3.200	3.183	2.800	2.647
27	2.769	26.197	2.832	2.814	3.317	18.965	3.256	3.213	2.832	**2.677**
3*	3.160	24.902	3.030	3.081	3.661	18.372	3.704	3.533	3.030	**3.011**
25	2.896	23.970	2.870	2.894	3.133	20.459	3.068	3.113	2.870	2.767
4*	2.908	23.780	2.921	2.926	3.316	19.632	3.357	3.195	2.921	**2.863**
8*	3.020	23.627	2.995	3.026	3.518	18.904	3.440	3.451	2.995	2.997
26	2.940	23.384	2.881	2.852	3.077	17.516	3.056	3.114	2.881	2.677
28*	2.893	22.841	2.884	2.856	3.183	19.483	3.170	3.251	2.884	2.793
32	2.802	22.699	2.797	2.801	3.074	18.352	2.992	3.155	2.797	**2.728**
16*	3.015	22.676	3.038	3.012	3.474	21.736	3.485	3.400	3.038	2.961
11*	2.970	22.527	2.953	2.972	3.351	18.823	3.413	3.227	2.953	2.901
19	2.654	22.388	2.789	2.773	3.196	18.027	3.212	3.257	2.789	**2.785**
31	2.917	21.980	2.906	2.865	3.273	19.413	3.340	3.365	2.906	2.735
24	2.857	21.806	2.874	2.849	3.148	20.040	3.283	3.325	2.874	2.679
7*	2.965	21.712	2.943	2.963	3.648	18.932	3.527	3.451	2.943	2.913
18	2.843	21.628	2.838	2.794	3.021	20.797	3.012	3.128	2.838	**2.760**
23	2.726	21.426	2.702	2.775	2.937	19.298	3.035	3.119	2.702	**2.661**
29	2.786	21.132	2.783	2.783	2.955	18.550	2.891	3.014	2.783	**2.742**
14*	3.088	21.086	3.077	3.084	3.195	17.300	3.297	3.301	3.077	3.056
10	3.004	20.933	3.003	2.990	3.172	19.811	3.213	3.144	3.003	2.929
5*	3.044	20.796	2.981	3.011	3.332	17.517	3.339	3.271	2.981	2.985
9*	3.080	20.739	3.111	3.069	3.498	19.246	3.560	3.426	3.111	**3.040**
15*	2.964	20.662	3.027	3.036	3.510	18.122	3.497	3.483	3.027	**3.046**
22	2.720	20.646	2.751	2.763	2.806	19.235	2.835	2.950	2.751	2.624
2*	2.924	20.481	3.079	3.003	3.487	17.857	3.504	3.406	3.079	3.000
12*	3.184	20.341	3.125	3.056	3.804	17.541	3.662	3.549	3.125	**3.062**
6*	2.983	20.316	2.992	2.958	3.398	23.434	3.327	3.254	2.992	2.919
21	2.862	19.716	2.860	2.886	3.317	19.692	3.182	3.149	2.860	2.629
1*	2.962	19.506	2.907	2.962	3.327	21.160	3.256	3.169	2.907	2.878
30	3.019	18.981	2.955	2.922	3.134	17.515	3.140	3.233	2.955	2.854
13*	2.940	18.801	3.024	3.015	3.274	18.773	3.383	3.159	3.024	2.981
17	2.810	18.426	2.786	2.779	3.034	19.012	3.098	3.152	2.786	**2.698**

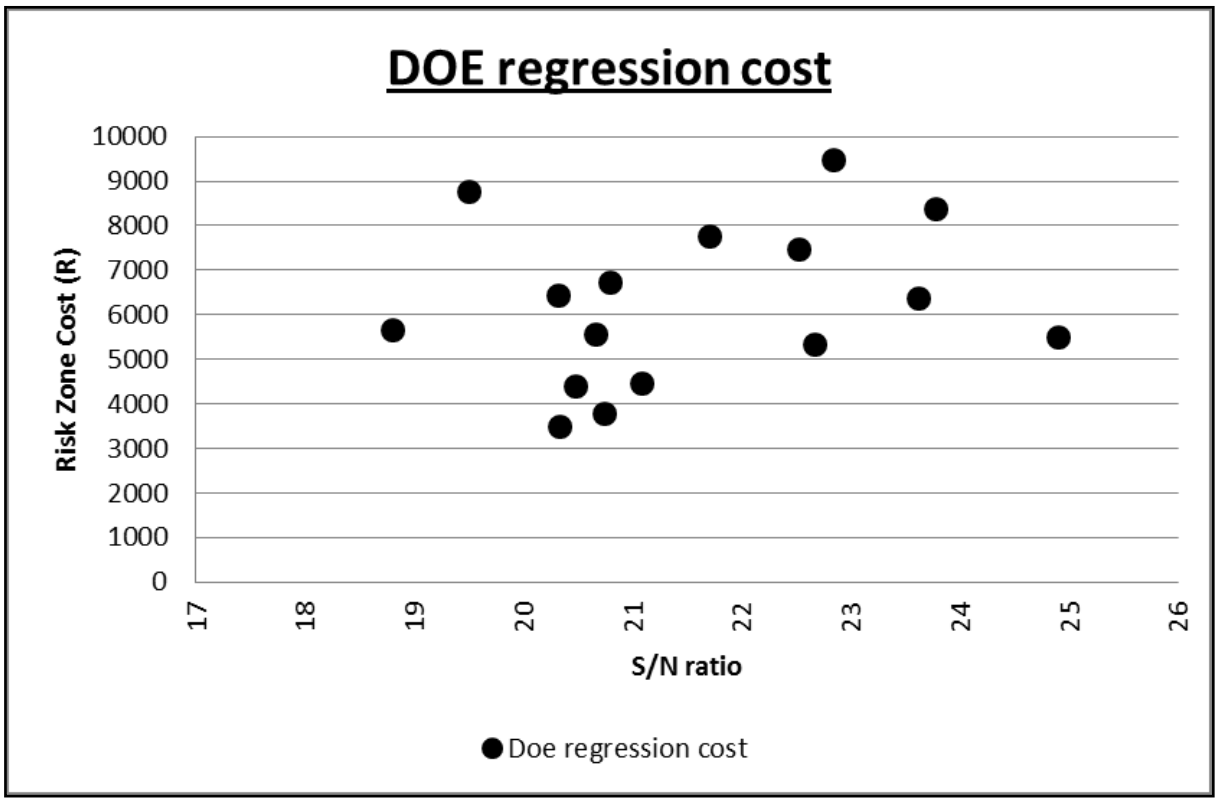
Table 7.3: Signal to noise ratio ranking

Table 7.3 is ranked from highest to lowest S/N ratio based on the estimated period, because the estimated period is the base used for predicting the validation period. The ranked table assist in selecting the experimental runs in a ranked order in terms of minimizing noise factors.

Table 7.4 below shows the summary for the ranked experimental runs with the applicable cost for the third validation period.

First period Base		Second predicted period	
DOE run number	Ranked S/N	Doe regression cost	Normal multiple regression cost
3*	24.90	5522	5974
4*	23.78	8374	10135
8*	23.63	6365	6315
28*	22.84	9482	12513
16*	22.68	5332	7253
11*	22.53	7476	8970
7*	21.71	7760	8629
14*	21.09	4478	4925
5*	20.80	6723	6626
9*	20.74	3777	5283
15*	20.66	5585	5162
2*	20.48	4422	6249
12*	20.34	3509	4795
6*	20.32	6444	8448
1*	19.51	8784	9664
13*	18.80	5660	6728

Table 7.4: Ranked S/N ratio with validation period cost



Graph 7.7: S/N ratio vs risk zone cost

Graph 7.7 shows that there is no relationship between S/N ratio and costs associated with the different zones. For this reason ranking of cost Vs S/N ratio is irrelevant and will only be a guide for selecting experimental runs with highest S/N ratio and lowest cost.

DOE regression S/N ratio with low cost will be used for decision making because normal multiple regression cost is consistently higher than DOE regression, see Table 7.4.

7.4 COST METHODS RESULTS

7.4.1 DOE zone risk ranking analysis

For the estimated period, 66% of the same colour codes for the DOE base for the estimated period are also predicted by both the DOE regression and normal multiple regression. It indicates a good stability for the data in the base DOE for the estimated period.

For the validation period, 59% of the same colour codes for the DOE base for the validation period are also predicted by both the DOE regression and normal multiple regression. It indicates a fair stability for the data in the base DOE for the validation period.

For the second validation period, the experimental runs that have similar risk profiles across all three periods were selected, based on a colour combination criteria. The selected 16 runs are 1 – 9, 11 – 16 and 28, which represent the lowest non-conformance risk and cost profile.

7.4.2 Zone risk profile analysis per period

The zone risk profile for period 1 shows that the process operating level for the DOE runs for each operating is below target and remains fairly stable between good and fair with a small portion in the poor zone. Al three categories are similar within each zone, indicating a stable production process for period 1.

The zone risk profile for validation period shows that the process operating level for the DOE runs for each operating category shifted towards target, and also shows a shift form the best to the good and fair zones, based on sequential experimental run number. The shift occurred approximately halfway through the experimental runs but the

processing level remains stable. This shift was due to a raw material change which was identified as part of normal process variation.

For validation period, the three categories are also close to each other but not as close as the estimated period. The shift in processing level influenced prediction variation for validation period DOE regression and validation period regression, but with a simple process adjustment, a similar process operating level can be achieved.

Predicting the validation period with DOE regression and normal multiple regression with regression equations that used the data for the estimated period shows a predicted operating level for the validation period similar to the estimated period.

Although the predicted levels are similar, the normal regression predicts a large portion in the red zone, which increases the risk of producing out of specification products.

Compared to the period 1 DOE base, the prediction model that is closest to the actual even with the process shift is the DOE regression model.

Even with a simple process adjustment to the actual DOE base data, the best predictor is still the DOE regression with a close approximation of the validation period.

7.4.3 Cost analysis

Cost summary table							
First period			Second period			Second period -	
Doe base	DOE regression	Multiple regression	DOE base	DOE regression	Multiple regression	Prediction - with 1st period DOE regression	Prediction - with 1st period multiple regression
R 279 080	R 276 607	R 282 428	R 80 908	R 76 830	R 65 353	R 276 607	R 360 739

Table 7.5: Cost summary table for two periods

The cost for the estimated period is similar for all three prediction categories but is mostly on the lower side of the target value which in terms of product performance is better than operating towards the high side of the target value.

The costs are also similar for the validation period compared to the estimated period, but much lower because the experimental outcomes are close to the target value. The lower costs for the validation period in terms of product performance have a small risk of producing out of specification.

The costs are significantly different for the two prediction categories but are significantly higher than the first and validation period. The difference in costs is explained by the fact that both the prediction models were calculated from the estimated period's data but the actual process in the validation period changed. However, the important issue is that the DOE regression model caused significantly lower costs than for the multiple regressions.

7.4.4 Signal to noise ratio

In Table 7.4, the ideal is to select runs with the maximum ratio because the risk of being influenced by external variation is minimized.

The 16 experimental runs 1 – 9, 11 – 16 and 28 selected were based on cost profiles across the three periods with no consideration to Signal to noise ratios. Each of these runs has a specific Signal to noise ratio that must be taken into account for final selection.

Table 7.6 below shows the summary for the ranked S/N ratios with the applicable cost for the third validation period.

First period Base		Second predicted period	
DOE run number	Ranked S/N	Doe regression cost	Normal multiple regression cost
3*	24.90	5522	5974
4*	23.78	8374	10135
8*	23.63	6365	6315
28*	22.84	9482	12513
16*	22.68	5332	7253
11*	22.53	7476	8970
7*	21.71	7760	8629
14*	21.09	4478	4925
5*	20.80	6723	6626
9*	20.74	3777	5283
15*	20.66	5585	5162
2*	20.48	4422	6249
12*	20.34	3509	4795
6*	20.32	6444	8448
1*	19.51	8784	9664
13*	18.80	5660	6728

Table 7.6: Ranked S/N ratio with validation period cost

There is no relationship between S/N ratio and costs associated with the different risk zones. For this reason ranking of cost Vs S/N ratio is irrelevant and will only be a guide for selecting experimental runs with highest S/N ratio and lowest cost.

DOE regression S/N ratio with associated cost will be used for decision-making because normal multiple regression cost is consistently higher than DOE regression, see Table 7.6.

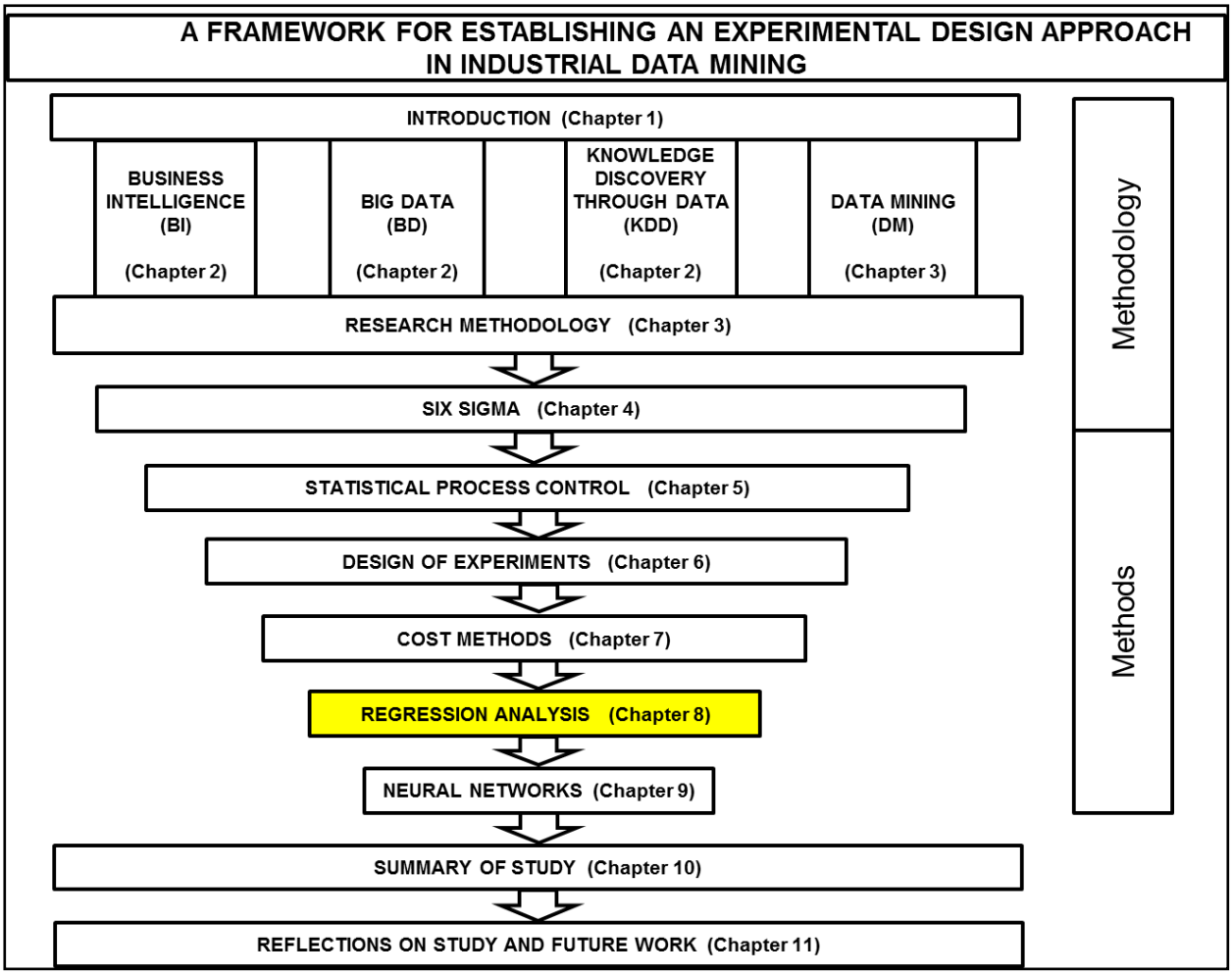
7.5 CONCLUSIONS

Costs are significantly different for the two prediction categories but are significantly higher than the first and validation period. The difference in costs is explained by the fact that both prediction models were calculated form the estimated period's data but the actual processing parameters in the validation period changed. However, the important issue is that the DOE regression model caused significantly lower costs than for the multiple regressions.

Cost analysis is an important factor when analysing data bases. To find optimal process solutions provided by an analytical technique is not enough. The associated costs to a proposed solution not only provide a benchmark for financial evaluations but can also serve as a control element in the analytical phases of DMAIC.

CHAPTER 8

REGRESSION ANALYSIS



8.1 INTRODUCTION

Regression analysis consists of a collection of techniques used to explore relationships between variables. The basis for regression analysis is to fit models for a dependent variable as a function of one or more independent variables. Regression analysis compliments designed experiments in predicting the behaviour of a dependent variable through selected independent variables. In this research a comparison between regression analysis and designed experiment (DOE) model regression is the basis for this chapter.

8.2 DESCRIPTIONS OF TECHNIQUES AND APPLICATIONS

8.2.1 Regression analysis

Regression analysis is not new and has been a critical part of statistical techniques used through the years, specifically when trying to find relationships amongst independent variables that could affect a dependent variable.

Multiple regression measures relationships between multi independent variables and a dependent variable. It sets a platform for measuring the numerical scale for group or individual relationships based on statistical assumptions and measurements (Moeinaddini *et al*., 2014:3485).

For this study we focus on linear relationships between multi independent variables and dependent variables. These linear relationships expressed as generalized linear relationships may also manifest in multi dimensions. For this reason, multiple regression serves as a very useful multivariate statistical tool (Weisburg, 1985:1).

Ryan (1989:264) explains that there are various procedures within a wide area of linear regression applications that have a direct implication in quality improvement work, and regression analysis is a complementary statistical approach of analysing data from designed experiments' model outcomes.

Brightman (1999:364) refers to applying regression analysis effectively as not an easy exercise, but adds that it is even more difficult to interpret the results, so that it makes sense in terms of both quantitative and qualitative variables. Explaining regression models in terms of parameters, dependent and independent variables and predictor variables, formulating the regression prediction model remains a challenge for the analyst to ensure user confidence.

8.2.2 Scatter plots

The simplest way to study correlations and/or identify patterns is to plot a bi-variate scatter diagram. AT & T Technologies (1985:143*)* describes it as obtaining values for two variables, x and y, in pairs. That is, measure x on a certain unit and y on the same unit, identifying them as a pair. One point on the scatter diagram represents one pair of x and y values.

StatTrek (2014) defines a scatter plot as a kind of mathematical diagram utilising Cartesian coordinates to show values for bi-variables for a set of data using horizontal and vertical axes. If the plotted data points have a relationship, it is called correlation.

The more closely the data points get when plotted to make a straight line, the higher the correlation and the stronger the relationship. Correlation represents two ways, a negative or positive correlation between the data points plotted.

A scatter plot is used when a variable exists that is influenced by another variable. The dependency of variables in plotted data points determines the relationship of the data to be analysed.

Figure 8.1 shows illustrations how scatter plots show correlations and relationships on plotted data.

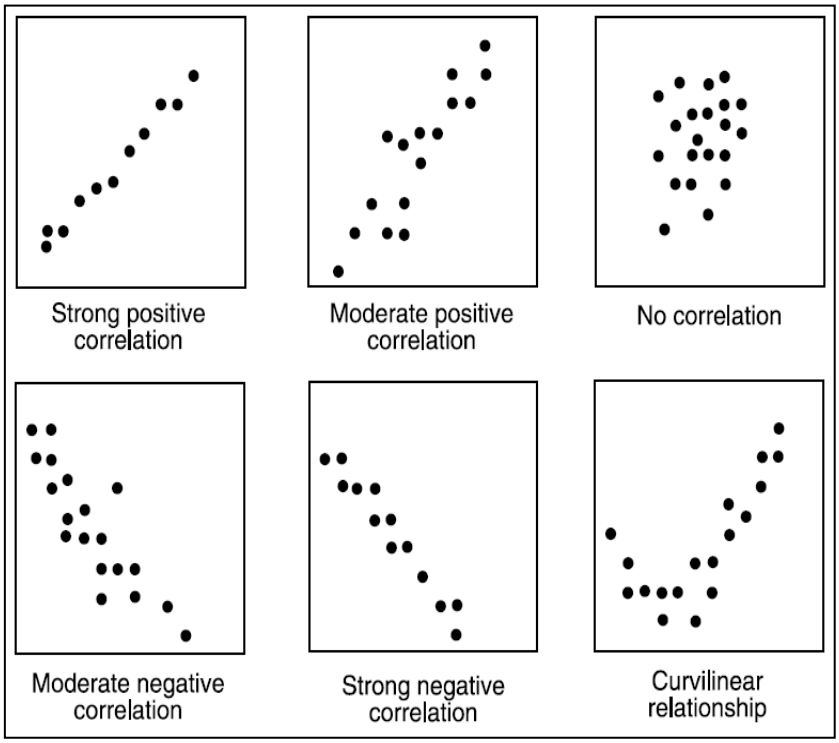


Figure 8.1: Example of a scatter plot outputs

8.2.3 Prediction error

Prediction error statistics are useful in providing a scientific guideline in prediction accuracy. Care should be taken not to take calculated accuracies as 100% true, they only give a theoretical guide of how accurate your prediction model is. Definitions of prediction error measurement statistics used for this study follow as:

Bias. Bias is a term that shows the error in forecasting and is defined as “the average error and tells whether the forecast tends to be too high or too low and by how much” (Render *et al*., 2012:179).

MAD. The mean absolute deviation is a popular forecasting measure and is particularly useful to compare different forecasting techniques. It is defined as “the average, absolute difference between the forecast and the actual demand” (Taylor, 2013:724).

MSE. The mean square error is another forecasting technique comparing different forecasting techniques, with the objective to have the smallest error possible. It is defined that “each individual error value is squared, and then these values are summed and averaged” (Taylor, 2013:726).

MAPE. The mean absolute percentage error is used to express the prediction error as a percentage and units. It is defined as “the average of the absolute values of the errors expressed as percentages of the actual values” (Render *et al*., 2015:174).

8.3 PREDICTION ANALYSIS – REGRESSION VS EXPERIMENTAL DESIGN

Six categories were identified to summarise the prediction variation between the first and validation periods as well as between the two regression prediction techniques.

First category: Estimation first period (2006 – 2009) - first period DOE run average Vs first period DOE regression: The first period DOE average for each experimental run is compared to the DOE regression run average for the first period with the same DOE runs. This shows DOE regression accuracy compared to the DOE run average for the first period. This analysis provides the accuracy of DOE prediction fit for the first period based on DOE average experimental runs.

Second category: Estimation first period (2006 – 2009) - first period DOE average Vs first period MR regression: The first period DOE average for each experimental run is compared to the MR run average for the first period with the same DOE runs. This shows MR accuracy compared to the DOE run average for the first period. This analysis provides the accuracy of MR prediction fit based on DOE average experimental runs.

Third category: Estimation second period (2009 – 2010) - second period DOE run average Vs second period DOE regression: The second period DOE average for each

experimental run is compared to the DOE regression run average for the second period with the same DOE runs. This shows DOE regression accuracy compared to the DOE run average for the second period. This analysis provides the accuracy of DOE prediction fit for the second period based on DOE average experimental runs.

Fourth category: Estimation second period (2009 – 2010) - second period DOE average Vs second period MR regression: The second period DOE average for each experimental run is compared to the MR run average for the second period with the same DOE runs. This shows MR accuracy compared to the DOE run average for the second period. This analysis provides the accuracy of MR prediction fit based on DOE average experimental runs.

Fifth category: Predicting second period with estimated first period (DOE regression) - DOE regression first period Vs DOE average for second period. The DOE average for each experimental run for the second period is compared to the individual estimated DOE regression for the first period with the same DOE runs. This shows the DOE regression prediction accuracy of the DOE average for second period based on DOE first period regression. This analysis provides the accuracy of prediction fit of the second period DOE average base using the first period DOE regression coefficients.

Sixth category: Predicting second period with estimated first period (MR regression) - MR regression first period Vs DOE average for second period. The DOE average for each experimental run for the second period is compared to the individual estimated MR regression for the first period with the same DOE runs. This shows the MR regression prediction accuracy of the DOE average for second period based on MR first period regression. This analysis provides the accuracy of prediction fit of the second period DOE average base using the first period MR regression coefficients.

Prediction accuracy for both periods, within a period and the prediction of the next period, was calculated and then compared to evaluate each category. The smaller the prediction error, the better prediction fit is achieved. In theory, the lowest prediction error indicates the best prediction model. Prediction error variation is useful for comparing different prediction techniques on a time series basis. It shows the magnitude of the prediction error as well as if there is stability in the prediction error.

8.3.1 Prediction error results

Table 8.1 below represents a summary of the prediction errors for six categories subdivided into three periods using average pressure as the dependent variable:

Description		Estimation: First period 2005 - 2009		Estimation: Second period 2010 - 2013		Predicting second period with estimated first period	
		Category 1	Category 2	Category 3	Category 4	Category 5	Category 6
Dependent variable	**Measure**	**DOE regression**	**Multiple regression**	**DOE regression**	**Multiple regression**	**First period DOE regression**	**Second period multiple regression**
Average pressure	**Bias (average error)**	**0.000**	**0.007**	**0.000**	**0.019**	**0.351**	**0.375**
	MSE	**0.031**	**0.040**	**0.059**	**0.103**	**0.351**	**0.375**
	MAD	**0.003**	**0.003**	**0.005**	**0.014**	**0.147**	**0.152**
	MAPE	**1.34%**	**1.38%**	**1.80%**	**3.13%**	**10.48%**	**10.68%**

Table 8.1: Prediction error for dependent variable (average pressure)

The prediction error is similar and very small between categories 1-2 and 3-4, which represent periods 1 and 2. This shows that both prediction techniques provide similar prediction error accuracies and therefore adequate predictors for each independent period.

The low prediction errors for periods 1 and 2 respectively indicate process stability for each period. For comparative analysis, a low prediction error gives confidence for process stability that is imperative to predict process behaviour.

Categories 5 and 6 are prediction errors for the validation period which shows higher prediction errors compared to the estimated period. These higher prediction errors are because of a technical change to a raw material for this period that caused the production process to operate on a different level than the estimated period, but still within product specification. However, if an adjustment is made to shift the validation period equal to the same operating level to the estimated period, the prediction errors will be similar for the first two periods. This is critical because high prediction errors are

not always attributed to the technique but profound knowledge of the process is essential in order to separate the two when process improvement is the focus.

DOE regression as a prediction technique shows a lower prediction error for all categories, irrespective of the represented period, compared to multiple linear regressions. Even though the prediction error for both techniques is low across all categories for BIAS, MSE, MAD and MAPE, DOE regression seems to be better for predicting future process behaviour.

Refer to Appendix 4 for prediction error samples.

For comparative analysis, the low prediction error across all three periods gives confidence for process stability that is imperative for predicting future process behaviour.

Prediction errors for the validation period are higher than the estimated period. These higher prediction errors are because of a technical change to a raw material for this period that caused the process to operate on a different level than the estimated period. With a simple process adjustment, the validation period will operate on the same operating level compared to the estimated period. The prediction errors will then be similar to the estimated period. This is important because the main goal is to evaluate if the validation period can be predicted accurately compared to the estimated period. If a process adjustment could complement the evaluation, then it should be taken into consideration.

DOE regression as a prediction technique shows lower prediction errors for all categories, irrespective of the represented period, compared to multiple linear regressions, because the values used for DOE are average values for each respective run. By this we do not claim that DOE regression outperforms multiple regression, only for this application it seems to be more accurate.

Even though the prediction error for both techniques is low across all categories for BIAS, MSE, MAD and MAPE, DOE regression seems to be the best option for predicting future process behaviour. Keep in mind that DOE predictions are based on average values operating on a low or high operating level.

8.4 RED "X" DISCUSSION

The red "X" concept was introduced during the 1990s by D. Shainin; it concentrated on the leading variable that causes the largest portion to process variability. Pareto analysis was the basis for identifying red "X". Since then the hunt for red "X" has evolved from Pareto analysis to multivariate analysis to identify the independent variable that is the core driver for any process. Diagram 8.1 shows an illustration for finding a red "X" within any process (Steiner & MacKay, 2005:4).

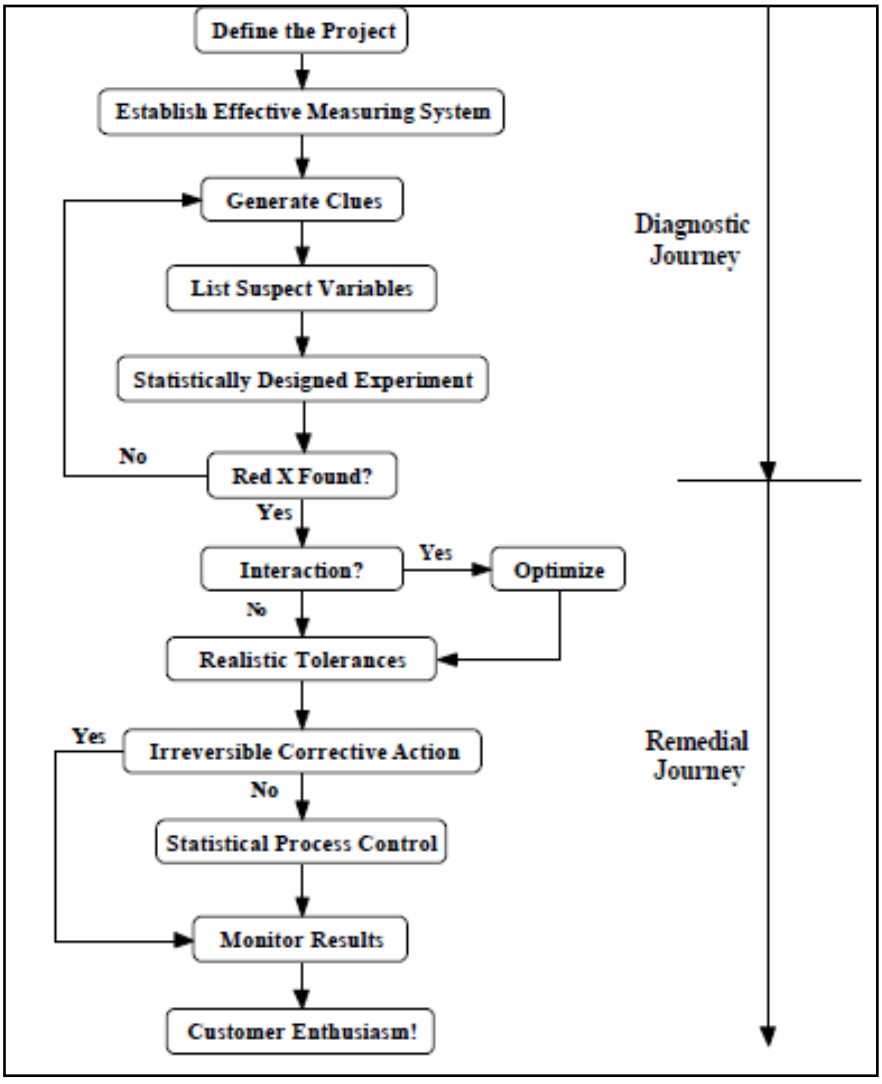
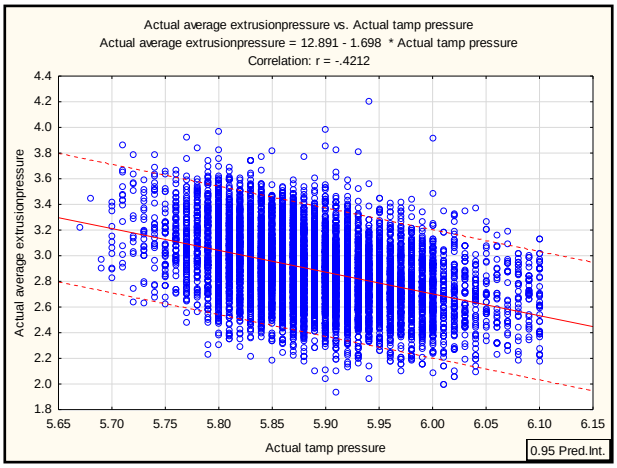
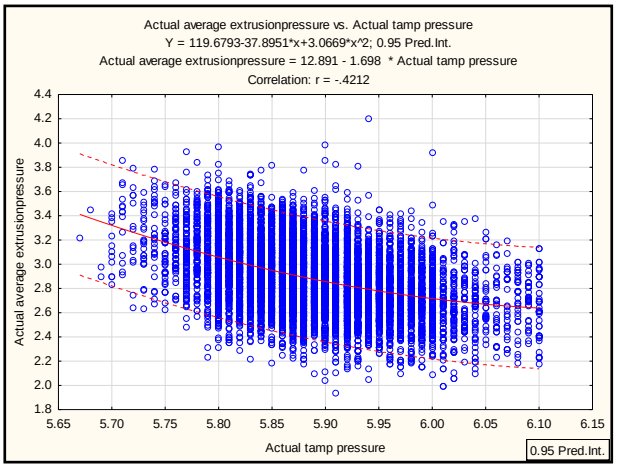


Diagram 8.1: The Shainin System for Quality Improvement (Steiner & MacKay, 2005)

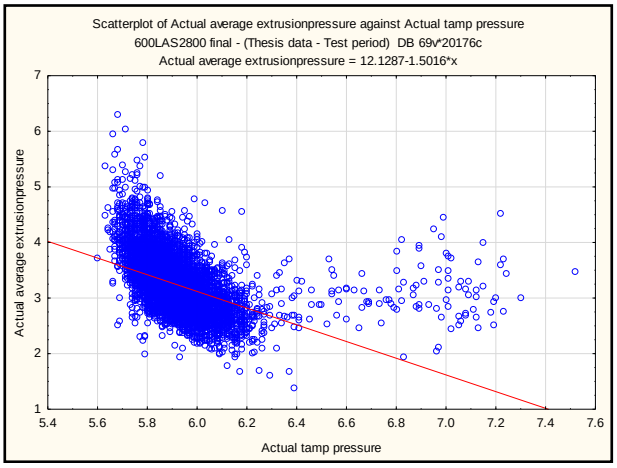
In section 8.5 variable 5 (Tamp pressure) was identified as the red "X" for this study. For this reason variable 5 is the main driver for this process and the subsequent graphs show the reduction process in finding a prediction formula beyond the validation period.

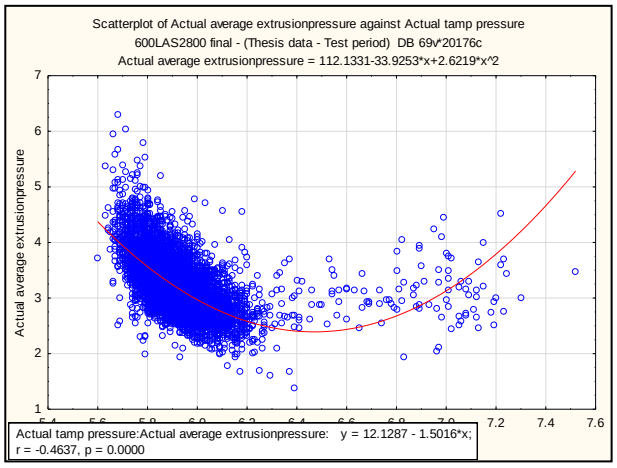




Graphs 8.1: Scatter plots estimated period: (Actual tamp pressure - Screened). Linear Vs polynomial fit

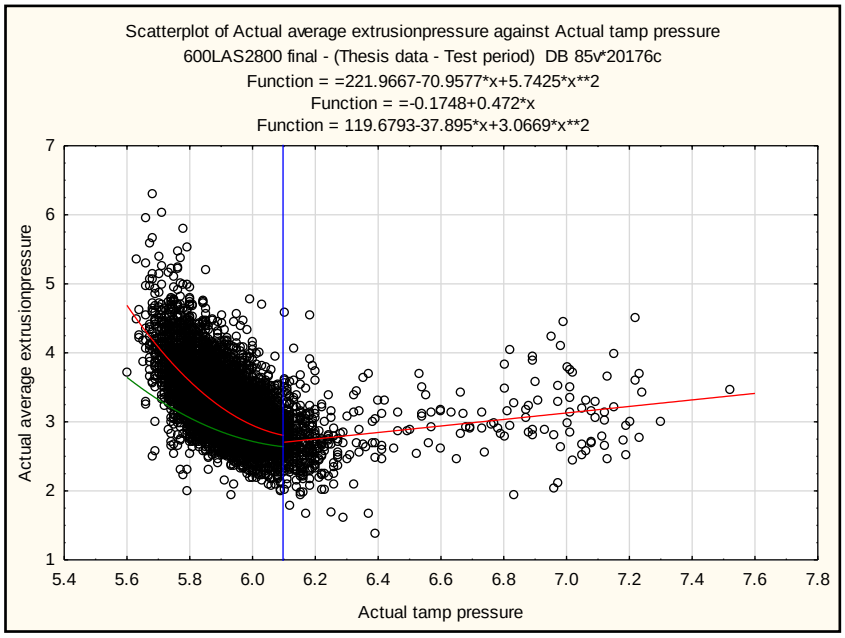
Graphs 8.1 show both a linear and a polynomial fit for the estimated period. The polynomial fit represents data better than a normal linear fit. A concern for predicting the validation period through extrapolation using a polynomial equation is that it is only accurate for the x values range between 5.7 and 6.1. Predicted values beyond 5.7 are unpredictable and inaccurate.





Graphs 8.2: Scatter plots validation period: (Actual tamp pressure - Screened). Linear Vs polynomial fit.

Graphs 8.2 show both a linear and a polynomial fit for the evaluation period. The polynomial fit represents data better than a normal linear fit. However, a polynomial fit for data larger than 6.1 does not represent the data; a linear fit will be more representative. A linear fit is more appropriate. This confirms the concern for predicting the validation period through extrapolation using a polynomial equation beyond 6.1.



Graph 8.3: Scatter plot validation period: (Actual tamp pressure - Screened) Combined polynomial & linear fit

Graph 8.3 represents three equations. Equation 1 represents the **polynomial equation for the estimated period (green)** for data below, including 6.1. It does not fit the data well. Equation 2 represents the **polynomial equation for the validation period (red)** for data below, including 6.1. It fits the data well. Equation 3 represents the **linear equation for the validation period (red)** for data above 6.1001. It fits the data well.

A proposed prediction equation for beyond the validation period is:

For tamp pressure values smaller and including 6.1: $y = 221.9667 - 70.9577*x_5 + 5.7425*(x_5)^2$; then for tamp pressure values larger than 6.101: $y = -0.1748 + 0.472*x_5$.

or

$$f(x) = \begin{cases} 221.9667 - 70.9577 * x_5 + 5.7425 * (x_5)^2 & \text{if } x_5 \le 6.1 \\ -0.1748 + 0.472 * x_5 & \text{if } x_5 > 6.1001 \end{cases}$$

8.5 APPLICATION OF REGRESSION ANALYSIS

The same data for the divided database, estimation and validation period are used for both MR and DOE. For DOE, as discussed in chapter 7, section 7.3.4, the grouped 32 DOE run design results are used for comparison to MR. For MR, raw and screened

individual data are used for MR model prediction calculations to illustrate different regression data fitting models.

Each period's database is evaluated independently to compare the significant independent variables for both periods. This comparison illustrates the level of process stability irrespective of the processing operating level, and if the same independent variables, irrespective of the period, have a significant effect on the process output.

Multiple regression analysis was applied on the estimation and validation period of the database respectively. Both periods are of equal importance in terms of evaluating consistency between the two periods in identifying significant independent variables. Similar independent variables should be statistically significant for both periods, irrespective if DOE regression or multiple regressions are used. Two regression summaries (tables 8.2 and 8.5) for the average pressure dependent variable using the remaining seven screened independent variables are as follows:

	Regression Summary for Dependent Variable: Actual average extrusionpressure(Y) R= .46712210 R^2= .21820306 Adjusted R^2= .21758974 F(7,8923)=355.78 p<0.0000 Std.Error of estimate: .24940					
N=8931	b*	Std.Err. of b*	b	Std.Err. of b	t(8923)	p-value
Intercept			11.60	0.561	20.7	0.000
1. Mix discharge temp(X_1)	-0.005	0.009	-0.00	0.003	-0.5	0.586
2. Actual cool time(X_2)	0.066	0.011	0.01	0.002	5.9	0.000
3. Cool begin temp(X_3)	-0.067	0.010	-0.00	0.001	-6.8	0.000
4. Actual dump temp(X_4)	0.164	0.011	0.03	0.002	15.2	0.000
5. Actual tamp pressure(X_5)	-0.427	0.010	-1.72	0.040	-43.4	0.000
6. Actual extrusion rate(X_6)	-0.134	0.010	-0.02	0.001	-13.6	0.000
7. Actual extrusion speed (X_7)	0.011	0.010	0.00	0.003	1.1	0.277

Table 8.2: Individual regression summary estimated period – Average pressure

Table 8.2 represents MR summary statistics for the estimated period of data. With the F statistically significant, (p>0.000), the independent variables are considered to be useful in predicting the dependent variable. Four variables in red (Cool begin temp, Actual cool time, Actual dump temp, Actual tamp pressure and Actual extrusion rate) are statistically significant, with p values of close to 0.000 for all of the five variables respectively, calculated for p = 0.05. The R = 0.4671 indicates a fair relationship of the independent variables to the dependent variable, but not ideal. An r = 1.0 is a perfect or ideal relationship. These five variables are significant for the MR model that statistically

influences the output of this model. The formula for this MR model, read for Table 8.7, is:

$Y = 11.599-0.00146*x_1 - 0.00459*x_2 + 0.0113*x_3 + 0.02793*x_4 - 1.72077*x_5 - 0.01775*x_6 + 0.00312*x_7$.

Typically, non-significant variables are excluded from a MR model equation. For this study, all variables are included in the equation because the non-significant variables have a negligible effect on the dependent variable but are a critical part of the process. In addition, the goal was not to use MR as a variable reduction process based on significance, but to have an inclusive multivariate formula for prediction.

Testing for collinearity amongst independent variables shows that all seven independent variables are independent with no or negligible collinearity. Using the Variance Inflation Factors (VIF) as described by Santana (2015:144) as our guide for if the VIF is equal to 1, there is no multicollinearity among factors, but if the VIF is greater than 1, the predictors may be moderately correlated. A VIF between 5 and 10 indicates high multicollinearity, and above 10, you can assume that the regression coefficients are poorly estimated, owing to multicollinearity. Refer to Table 8.3 for VIF values for the estimated period independent variables.

	Collinearity statistics		
Effect	Tolerance	Variance Infl fac	R square
Mix discharge temp (X_1)	0.990	1.010	0.010
Cool begin temp (X_2)	0.913	1.095	0.087
Actual cool time (X_3)	0.705	1.419	0.295
Actual dump temp (X_4)	0.757	1.321	0.243
Actual tamp pressure (X_5)	0.904	1.107	0.096
Actual extrusion rate (X_6)	0.907	1.103	0.093
Actual extrusion speed (X_7)	0.872	1.147	0.128

Table 8.3: Multicollinearity (VIF) table

Variable	Individual b_a	MR b_x	Diff: $b_a - b_x$	t Statistic: Diff/Stdev b_x	p-value	Significant α=0.05
x_1	**0.0021**	**-0.00146**	**0.00356**	**0.01427**	**0.990916**	**No**
x_2	**-0.00007**	**-0.00459**	**0.00452**	**0.01812**	**0.988466**	**No**
x_3	**0.01358**	**0.01130**	**0.00058**	**0.00232**	**0.998523**	**No**
x_4	**0.01179**	**0.02793**	**-0.01551**	**-0.06219**	**0.960460**	**No**
x_5	**-1.698**	**-1.72077**	**0.02277**	**0.09129**	**0.942044**	**No**
x_6	**-0.0223**	**-0.01775**	**-0.00455**	**-0.01824**	**0.988387**	**No**
x_7	**-0.0345**	**0.00312**	**-0.03762**	**0.15084**	**0.904691**	**No**

Table 8.4: Multicollinearity of regression coefficient based on t test

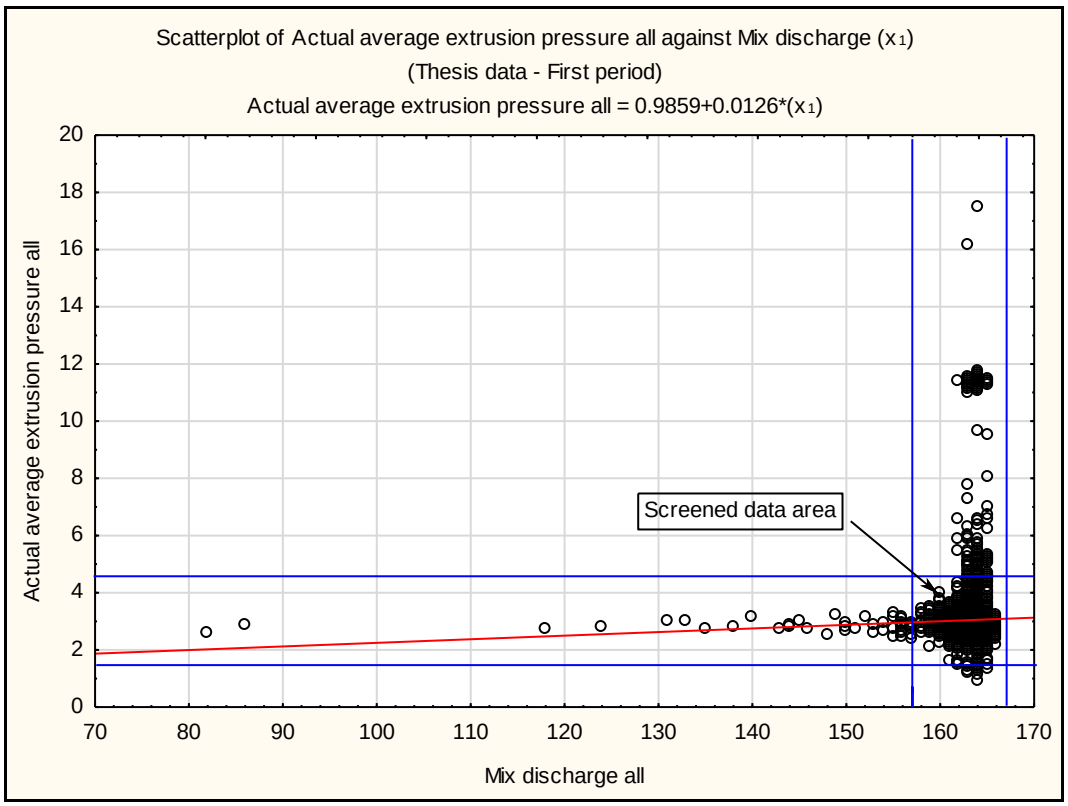
In addition to the VIF metric a regression coefficient t-test on α = 0.05 significance shows no significant difference between MR coefficients and individual regression coefficients. This confirms the VIF metric that no collinearity exists. Refer to Table 8.4

Because multicollinearity has a negligible effect amongst the independent variables, they are used independently for prediction further in the study. Hypothesis tests were done for **each screened** independent variable in Table 8.9, evaluating if a statistically significant correlation exists between individual independent variables and the selected dependent variable.

The effect of screening data following the proposed screening phases in chapter 4, section 4.4, was **firstly** to draw scatterplots for each independent variable showing raw unscreened data for each independent variable compared to raw dependent variable data. **Secondly**, to draw scatterplots for each independent variable showing screened independent data for each independent variable compared to dependent variable data.

Log transformations of the raw data for both estimation and validation period were performed, but it was decided to stay in the original scales, to keep the modelling focused. The transformations are in the appendices section.

The graphical effect follows below for the first period, **with regression analysis and hypothesis tests done on screened data. In addition, raw data scatter plots for each independent variable were done showing the screened portion form the raw data.**



Graph 8.4: Scatter plot estimated period: (Mix discharge temperature (x_1) – Raw data)

8.5.1 Hypothesis test first period – Mix discharge temperature (x_1)

Single variable null hypothesis for Mix discharge temperature

There is no correlation between Mix discharge temperature and Average extrusion pressure.

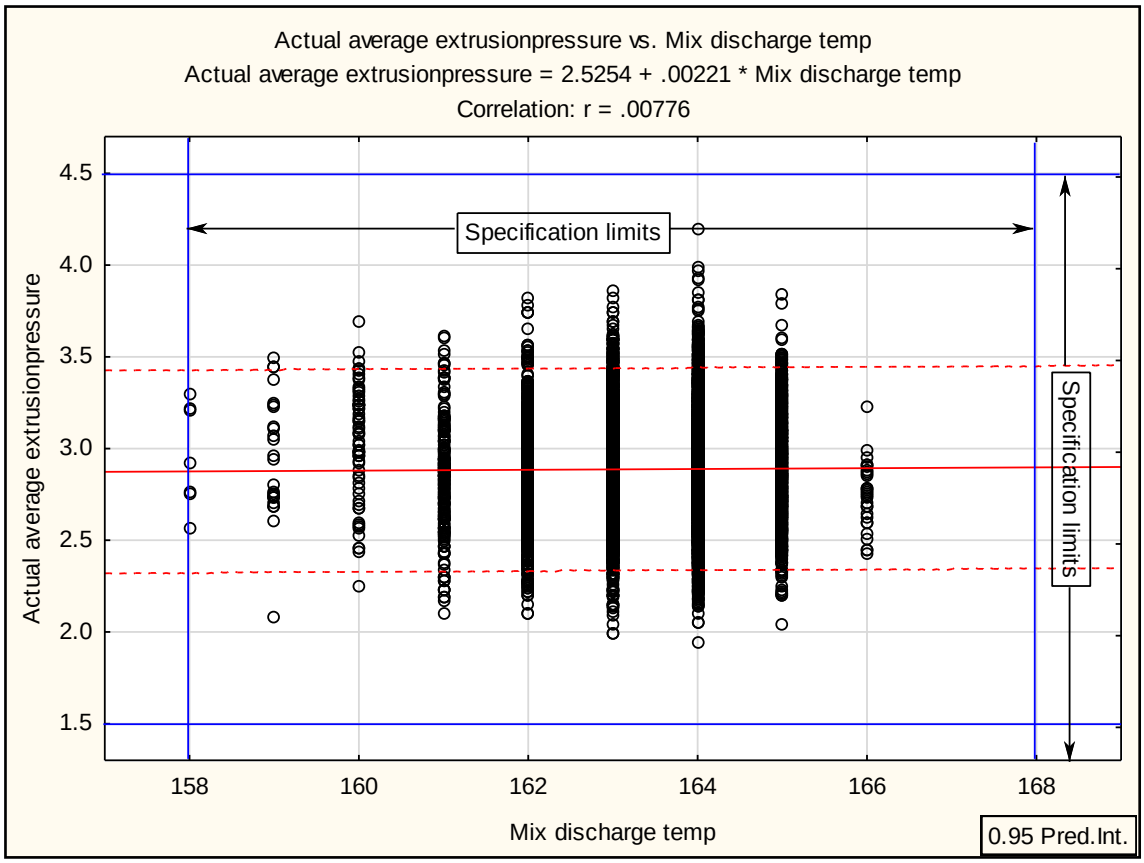
Single variable alternative hypothesis for Mix discharge temperature

There is a correlation between Mix discharge temperature and Average extrusion pressure.

Correlation between Mix discharge temperature and Average extrusion pressure is 0.77%. (See Graph 8.5). The associated p-value of no correlation between Mix discharge temperature and Average extrusion pressure is 0.463147, which is bigger than the significance level of α=0.05. The null hypothesis will not be rejected and therefore the correlation is not statistically significant for a significance level of 0.05.

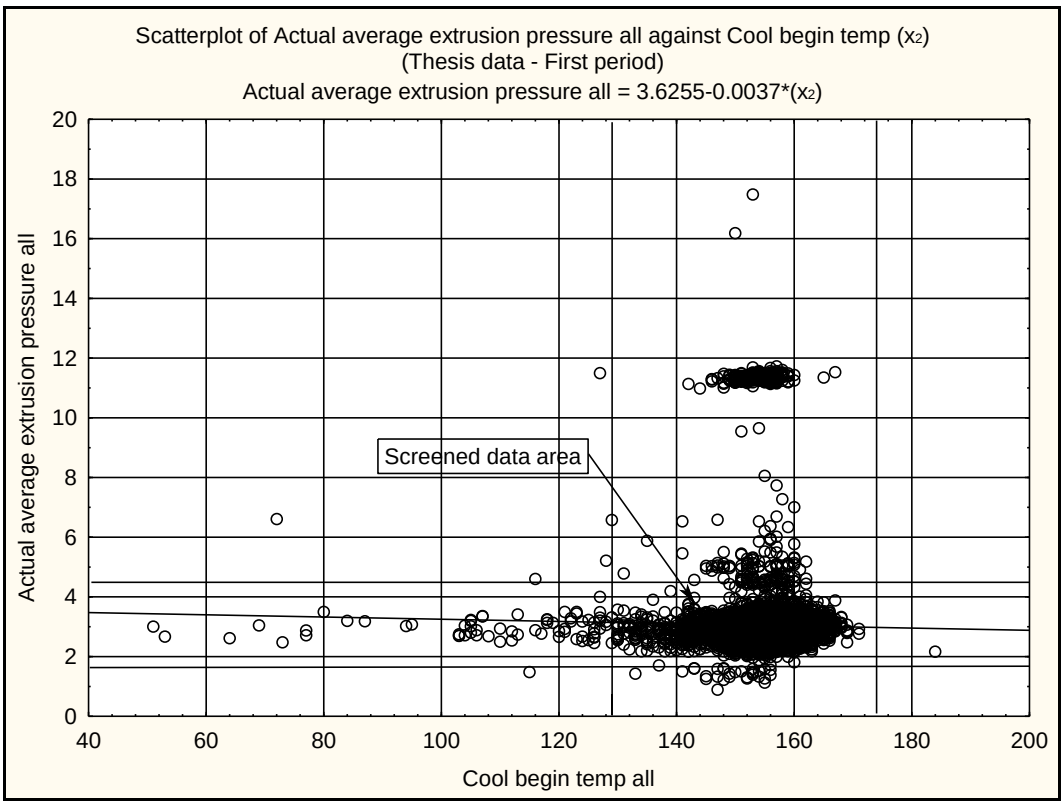
Conclusion

There is not a statistically significant correlation between Mix discharge temperature and average extrusion pressure.



Graph 8.5: Scatter plot estimated period: (Mix discharge temperature - Screened)

For variable 1, the hypothesis test shows no statistically significant correlation between Mix discharge temperature and average extrusion pressure. As an independent variable, the weak positive relationship (r = 0.00776), with a negligible **positive** influence of 0.00221 on pressure for the dependent variable with every degree increase of temperature for the independent variable, confirms a weak predictor variable. Table 8.4 for the MR model shows this variable not significant with a negligible **negative** influence of -0.00146 on pressure for the dependent variable with every degree increase of temperature for the independent variable. Therefore, variable 1 is a weak non-significant predictor for the regression model, and as an independent variable.



Graph 8.6: Scatter plot estimated period: (Cool begin temperature (x_2) - Raw)

8.5.2 Hypothesis test first period– Cool begin temperature (x_2)

Single variable null hypothesis for Cool begin temperature

There is no correlation between Cool begin temperature and Average extrusion pressure.

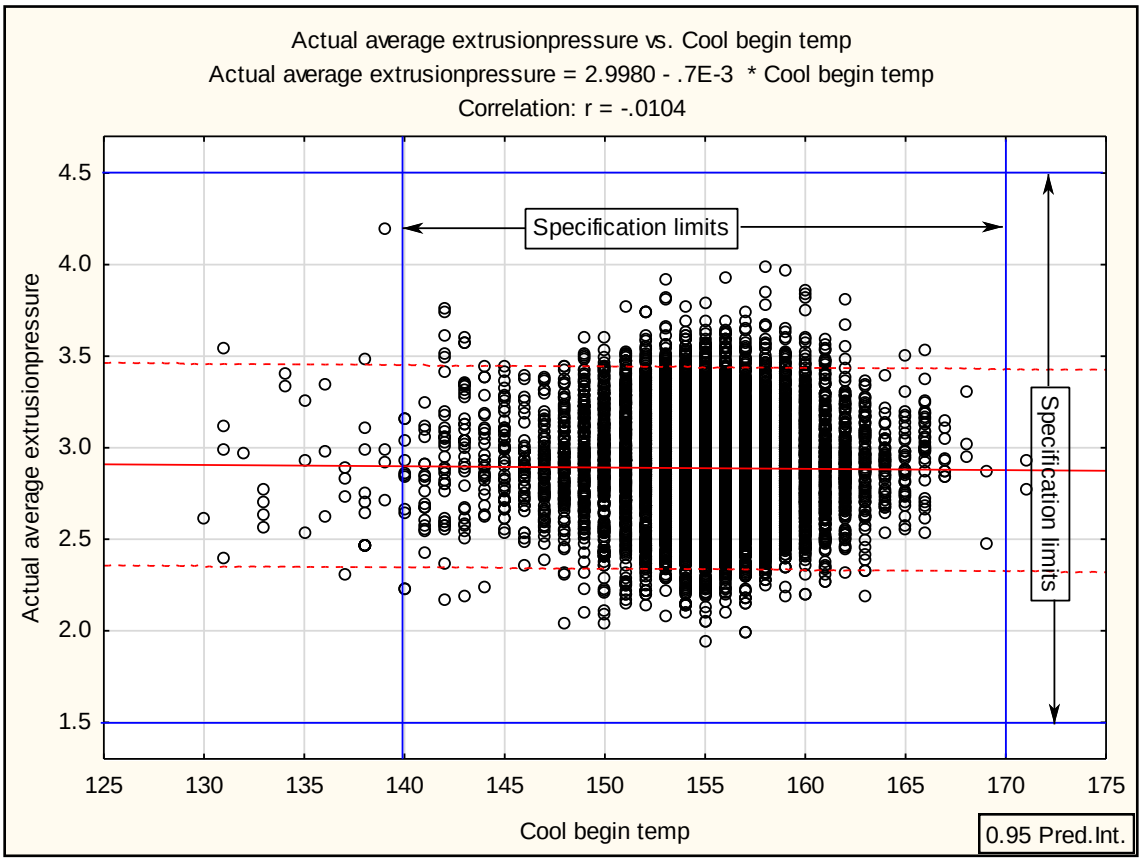
Single variable alternative hypothesis for Cool begin temperature

There is a correlation between Cool begin temperature and Average extrusion pressure.

Correlation between Cool begin temperature and Average extrusion pressure is -1.04%. (See Graph 8.7). The associated p-value of no correlation between Cool begin temperature and Average extrusion pressure is 0.32718, which is bigger than the significance level of α=0.05. The null hypothesis will not be rejected and therefore the correlation is not statistically significant for a significance level of 0.05.

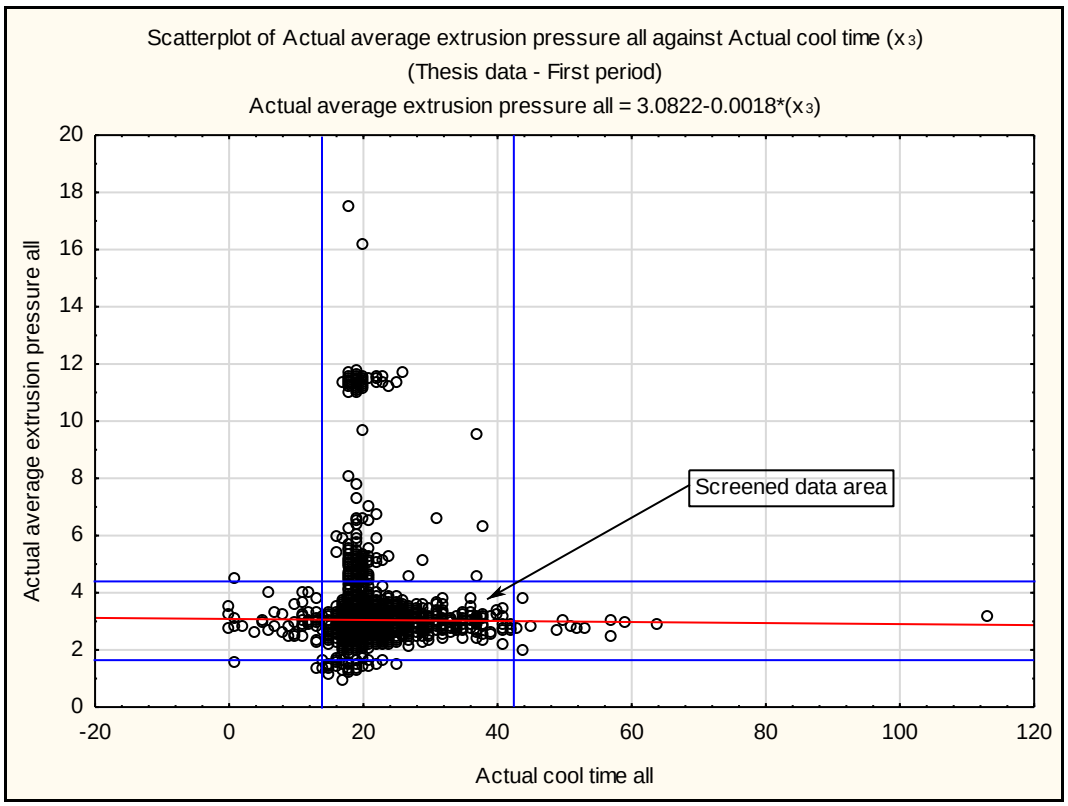
Conclusion

There is not a statistically significant correlation between Cool begin temperature and average extrusion pressure.



Graph 8.7: Scatter plot estimated period: (Cool begin temperature - Screened)

For variable 2, the hypothesis test shows no statistically significant correlation between Cool begin temperature and average extrusion pressure. As an independent variable the weak negative relationship (r = -0.0104), with negligible **negative** influence of 0.0007 on pressure for the dependent variable with every degree increase of temperature for the independent variable, confirms a weak predictor variable. Table 8.4 for the MR model shows this variable as significant with a negligible **negative** influence of -0.00459 on pressure for the dependent variable with every degree increase of temperature for the independent variable. Therefore, variable 2 is a weak predictor for the regression model, and as an independent variable, but only a significant predictor for the MR model.



Graph 8.8: Scatter plot estimated period: (Actual cooling time (x_3) - Raw)

8.5.3 Hypothesis test first period – Actual cooling time (x_3)

Single variable null hypothesis for Actual cooling time

There is no correlation between Actual cooling time and Average extrusion pressure.

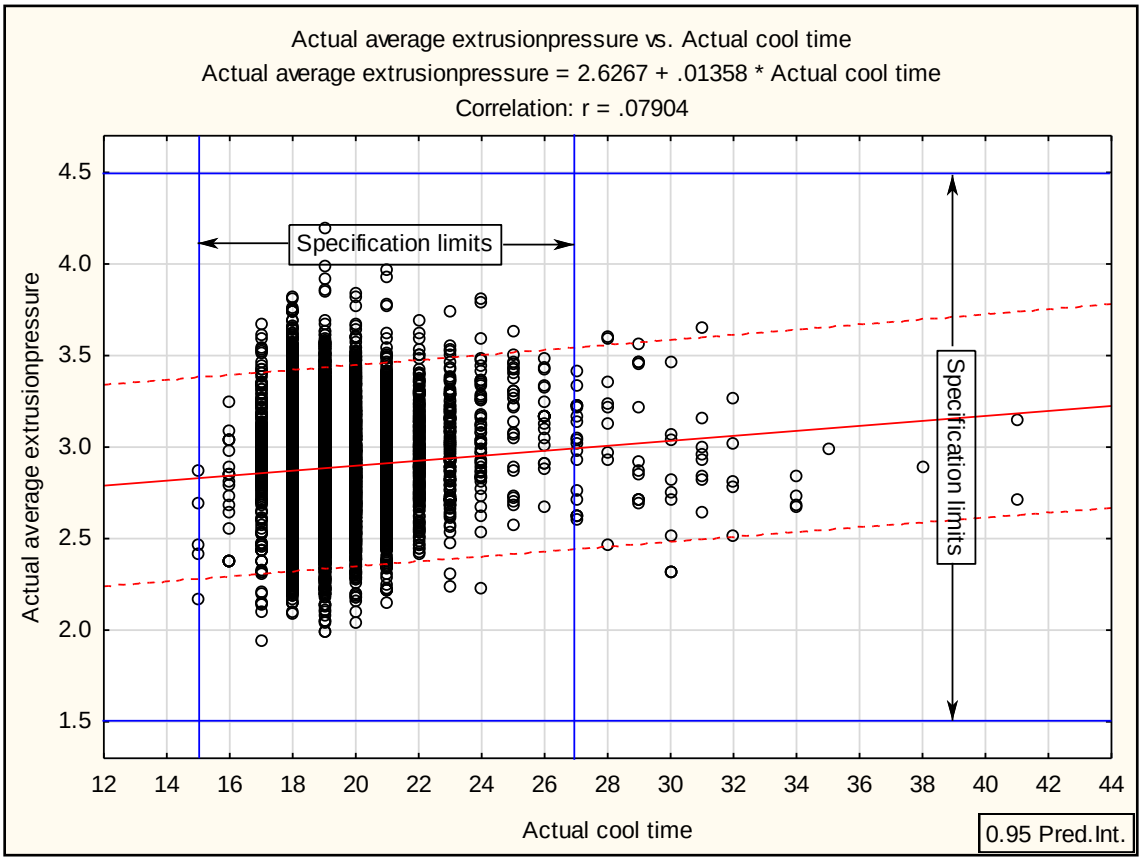
Single variable alternative hypothesis for Actual cooling time

There is a correlation between Actual cooling time and Average extrusion pressure.

Correlation between Actual cooling time and Average extrusion pressure is 7.904%. (See Graph 8.9). The associated p-value of no correlation between Actual cooling time and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

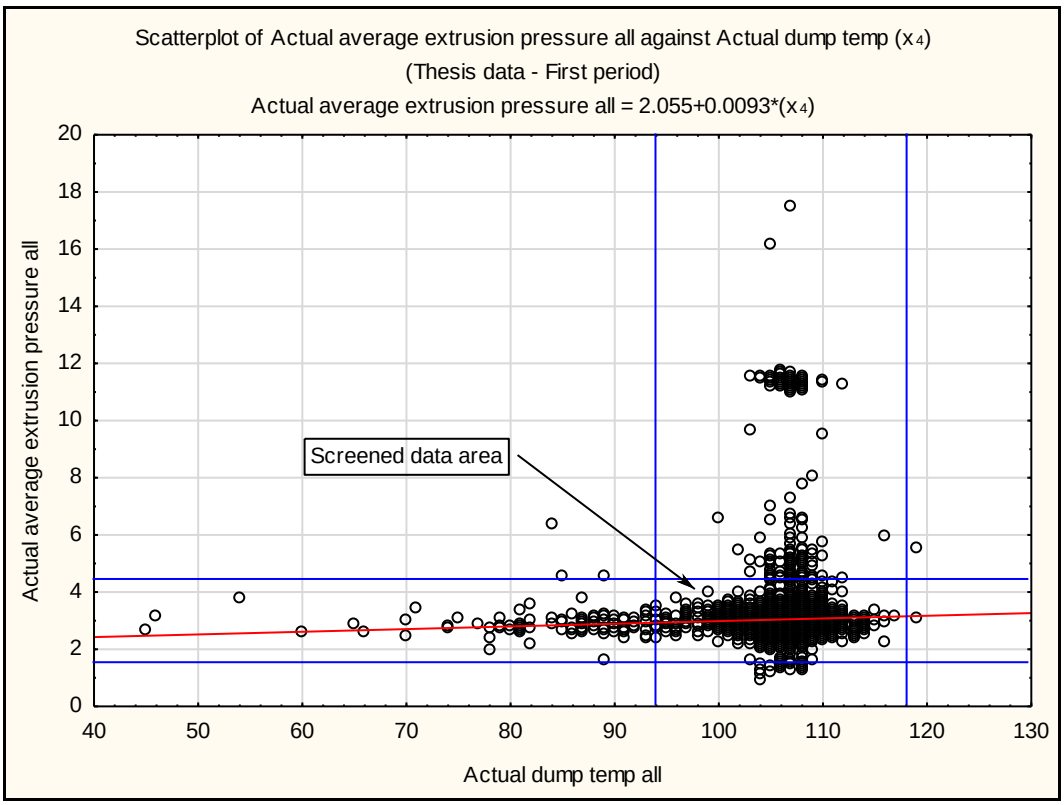
Conclusion

There is a statistically significant correlation between Actual cooling time and average extrusion pressure.



Graph 8.9: Scatter plot estimated period: (Actual cooling time - Screened)

For variable 3, the hypothesis shows a statistically significant correlation between Actual cooling time and average extrusion pressure. As an independent variable, the weak positive relationship (r = 0.07904), with negligible **positive** influence of 0.01358 on pressure for the dependent variable with every minute increase in time for the independent, confirms a weak predictor. Table 8.4 for the MR model shows this variable as significant with a negligible **positive** influence of 0.0113 on pressure for the dependent variable with every minute increase in time for the independent variable. Therefore, variable 3 is a weak significant predictor for the MR model and as an independent variable.



Graph 8.10: Scatter plot estimated period: (Actual dump temperature (x_4) – Raw)

8.5.4 Hypothesis test first period – Actual dump temperature (x_4)

Single variable null hypothesis for Actual dump temperature

There is no correlation between Actual dump temperature and Average extrusion pressure.

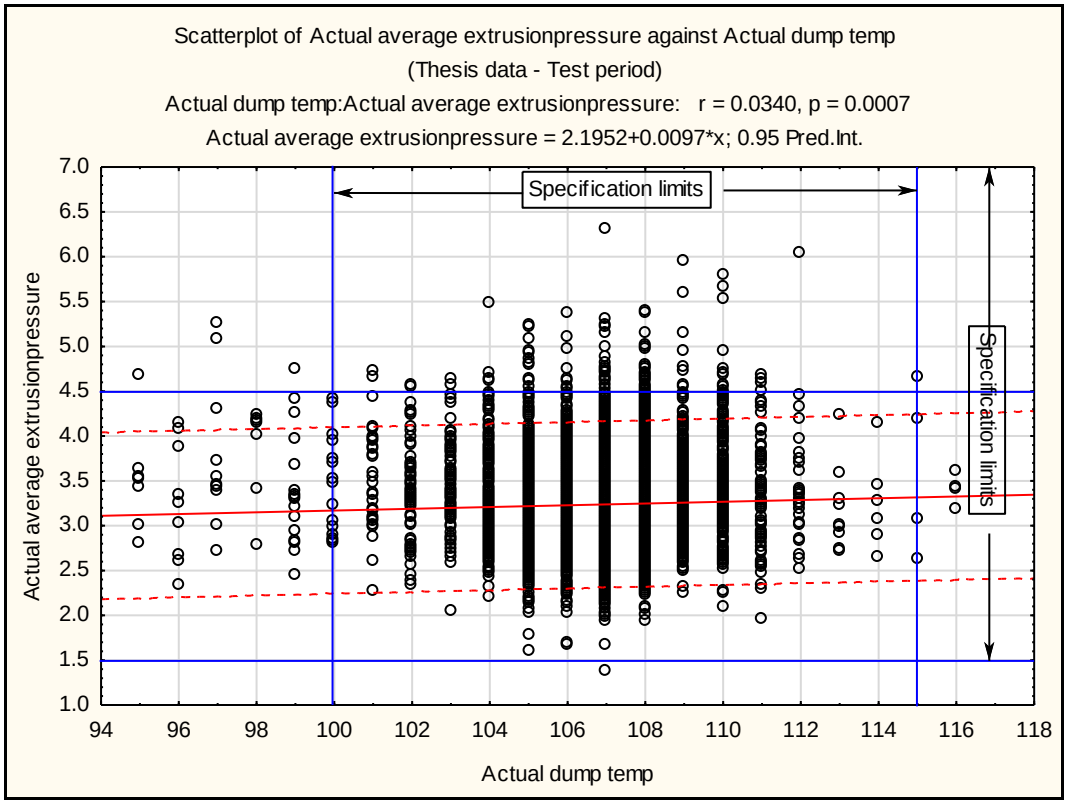
Single variable alternative hypothesis for Actual dump temperature

There is a correlation between Actual dump temperature and Average extrusion pressure.

Correlation between Actual dump temperature and Average extrusion pressure is 6.91%. (See Graph 8.11). The associated p-value of no correlation between Actual dump temperature and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

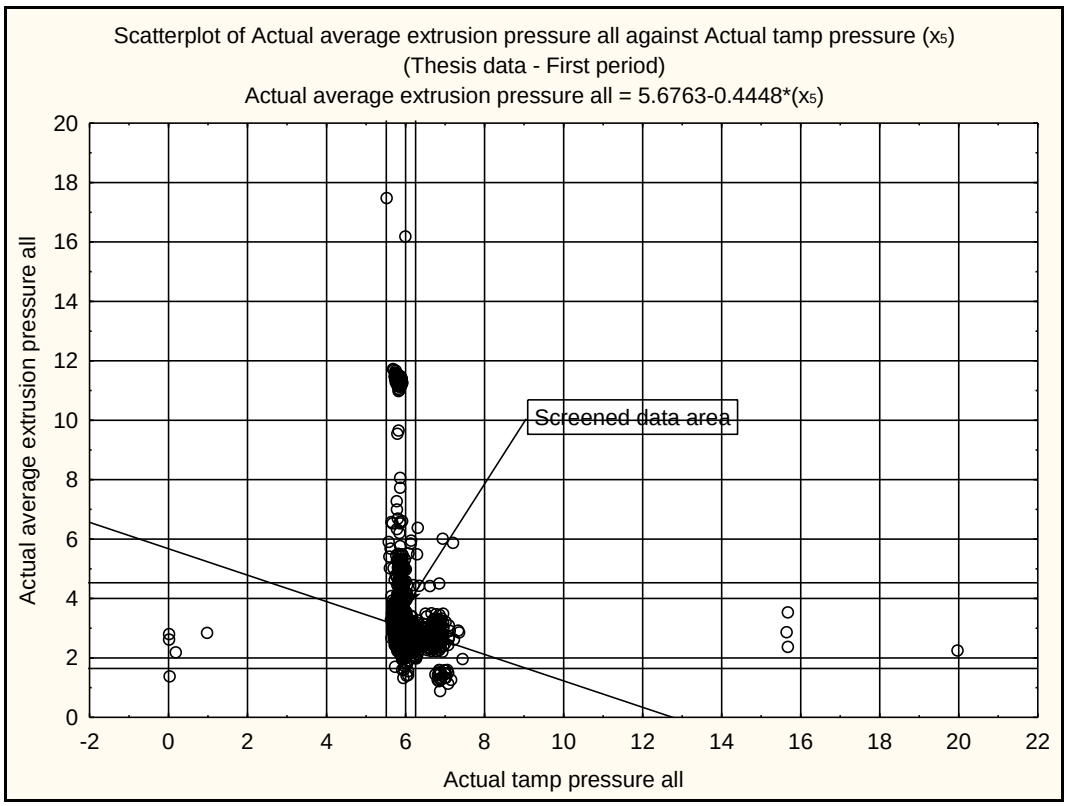
Conclusion

There is a statistically significant correlation between Actual dump temperature and average extrusion pressure.



Graph 8.11: Scatter plot estimated period: (Actual dump temperature - Screened)

For variable 4, the hypothesis shows a statistically significant correlation between Actual dump temperature and average extrusion pressure. As an independent variable, the weak positive relationship (r = 0.06910), with negligible positive influence of 0.01179 on pressure for the dependent variable with every degree increase of temperature for the independent variable, confirms a weak predictor. Table 8.4 for the MR model shows this variable as significant with a negligible positive influence of 0.02793 on pressure for the dependent variable with every degree increase of temperature for the independent variable. Therefore, variable 4 is a weak significant predictor for the regression model, and as an independent variable.



Graph 8.12: Scatter plot estimated period: (Actual tamp pressure (x_5) - Raw)

8.5.5 Hypothesis test first period– Actual tamp pressure (x_5)

Single variable null hypothesis for Actual tamp pressure

There is no correlation between Actual tamp pressure and Average extrusion pressure.

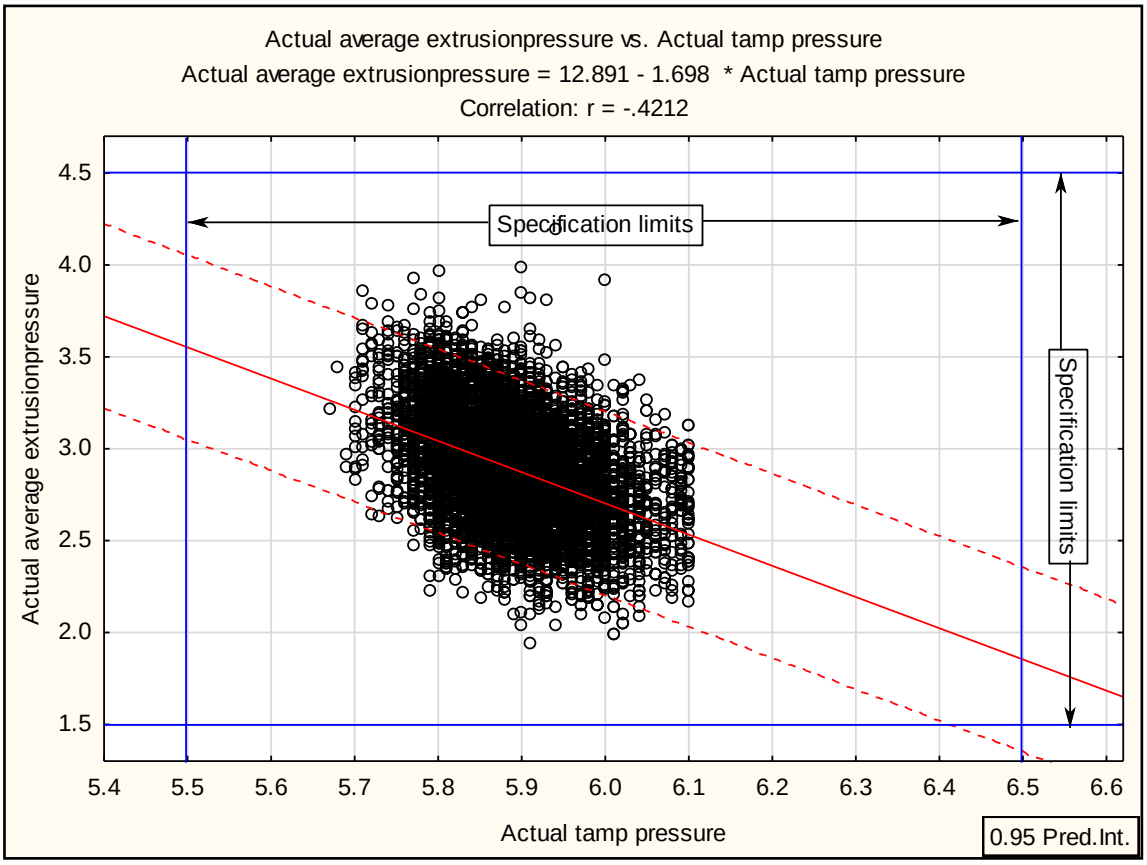
Single variable alternative hypothesis for Actual tamp pressure

There is a correlation between Actual tamp pressure and Average extrusion pressure.

Correlation between Actual tamp pressure and Average extrusion pressure is -42.12%. (See Graph 8.13). The associated p-value of no correlation between Actual tamp pressure and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

Conclusion

There is a statistically significant correlation between Actual tamp pressure and average extrusion pressure.

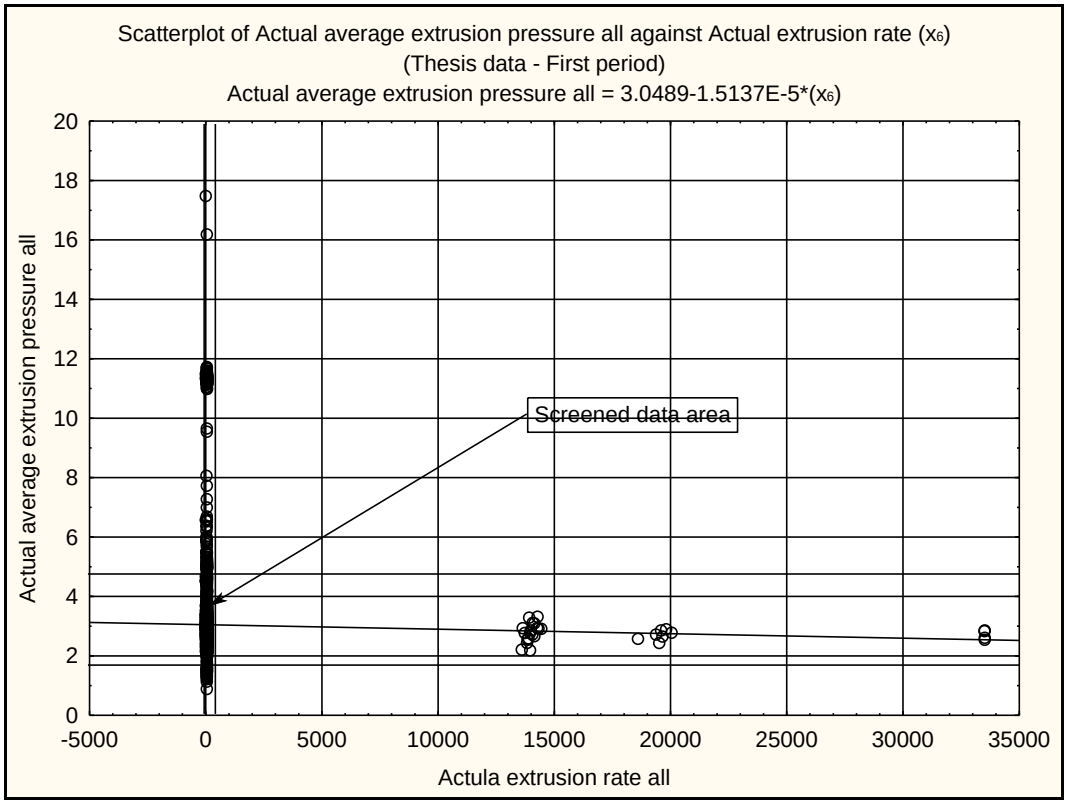


Graph 8.13: Scatter plot estimated period: (Actual tamp pressure - Screened)

For variable 5, the hypothesis shows a statistically significant correlation between Actual tamp pressure and average extrusion pressure. As an independent variable, the strong negative relationship (r = -0.4212), with a strong **negative** influence of -1.698 on pressure for the dependent variable with every unit increase of pressure for the independent variable, confirms a strong predictor. Table 8.4 for the MR model shows this variable as significant with a strong **negative** influence of -1.72077 on pressure for the dependent variable with every unit increase of pressure for the independent variable. Therefore, variable 5 is a strong and a significant predictor for the regression model, and as an independent variable.

Variable 5 shows the strongest regression relationship for both the regression model and as an independent variable. Seeing that no collinearity exists, this variable seems to be the largest single contributor to changes for the dependent variable in the regression model. For this reason, the prediction results for the regression equation for variable 5 will be compared to the regression model evaluating prediction accuracy between MR and single regression.

The regression formula is: Predicted average pressure = 12.891-1.698*x_5.



Graph 8.14: Scatter plot estimated period: (Actual extrusion rate (x_6) - Raw)

8.5.6 Hypothesis test first period– Actual extrusion rate (x_6)

Single variable null hypothesis for Actual extrusion rate

There is no correlation between Actual extrusion rate and Average extrusion pressure.

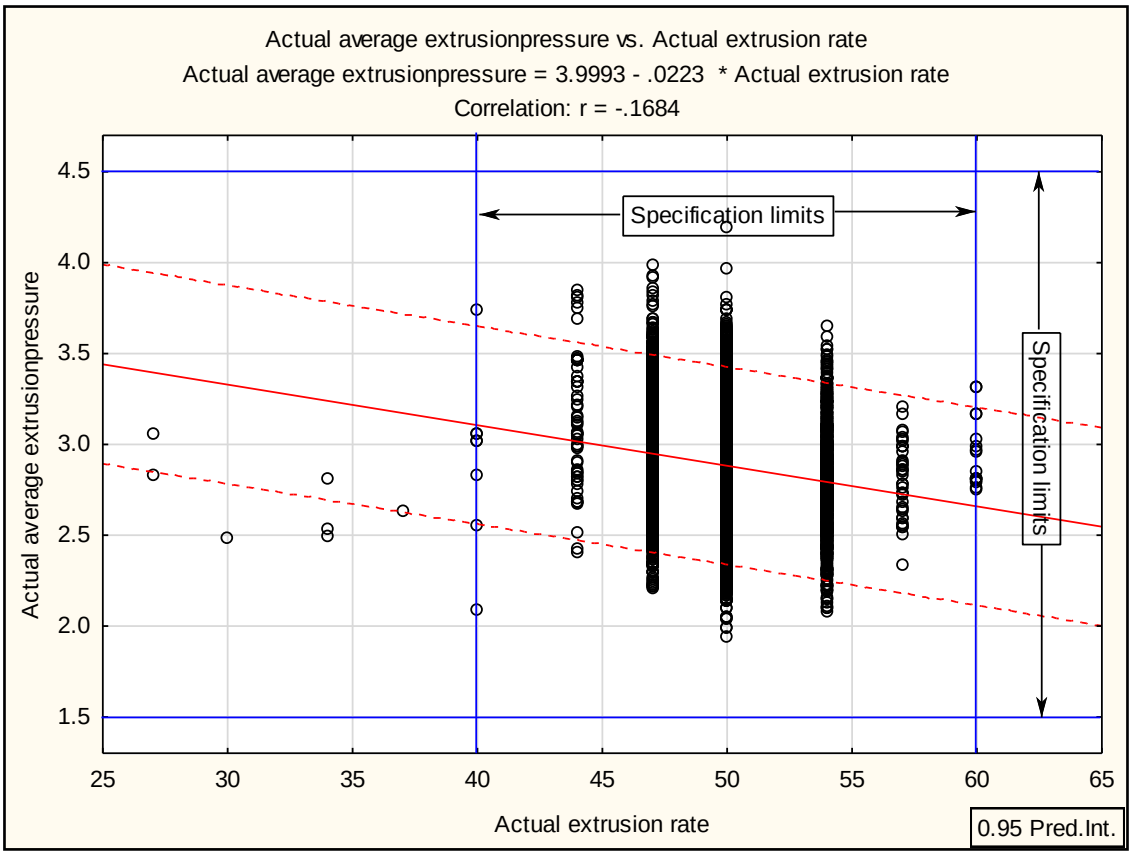
Single variable alternative hypothesis for Actual extrusion rate

There is a correlation between Actual extrusion rate and Average extrusion pressure.

Correlation between Actual extrusion rate and Average extrusion pressure is -16.84%. (See Graph 8.15). The associated p-value of no correlation between Actual extrusion rate and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

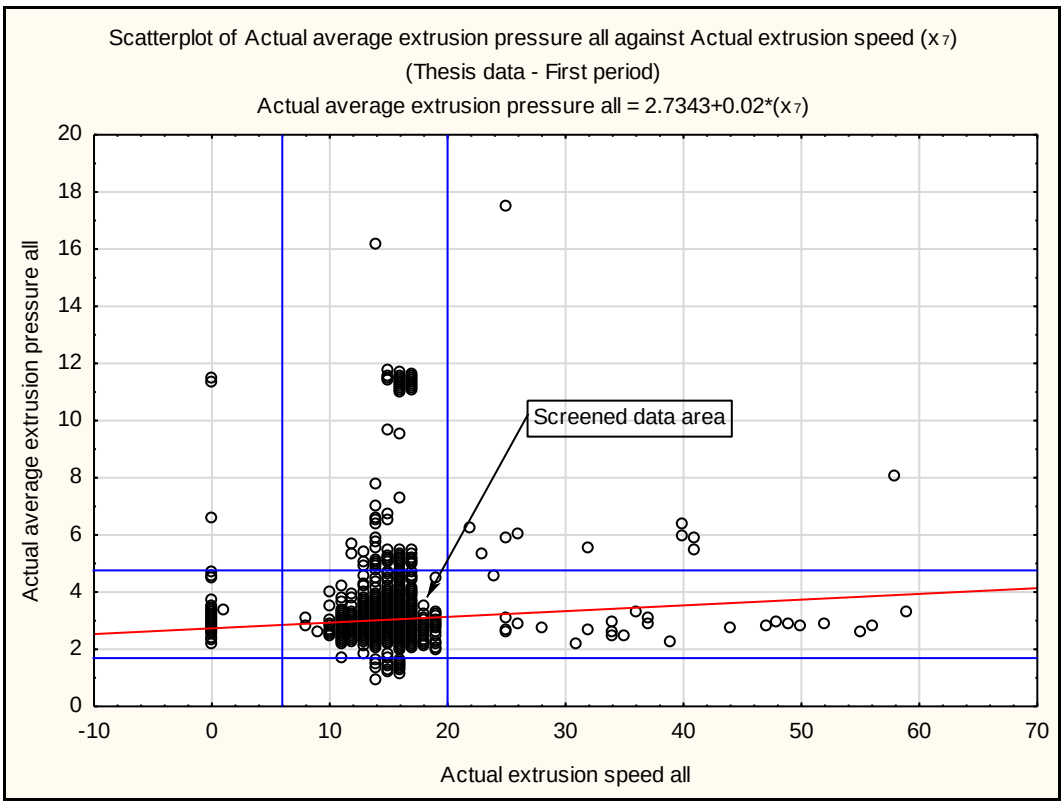
Conclusion

There is a statistically significant correlation between Actual extrusion rate and average extrusion pressure.



Graph 8.15: Scatter plot estimated period: (Actual extrusion rate - Screened)

For variable 6, the hypothesis shows a statistically significant correlation between Actual extrusion rate and average extrusion pressure. As an independent variable, the negative relationship (r = -0.1684), with negative influence of -0.0223 on pressure for the dependent variable with every unit increase for the independent variable, confirms a weak predictor. Table 8.4 for the MR model shows this variable as significant with a negative influence of -0.01775 on pressure for the dependent variable with every unit increase of temperature for the independent variable. Therefore, variable 6 is a weak significant predictor for the regression model and as an independent variable.



Graph 8.16: Scatter plot estimated period: (Actual extrusion speed (x_7) - Raw)

8.5.7 Hypothesis test first period – Actual extrusion speed (x_7)

Single variable null hypothesis for Actual extrusion speed

There is no correlation between Actual extrusion speed and Average extrusion pressure.

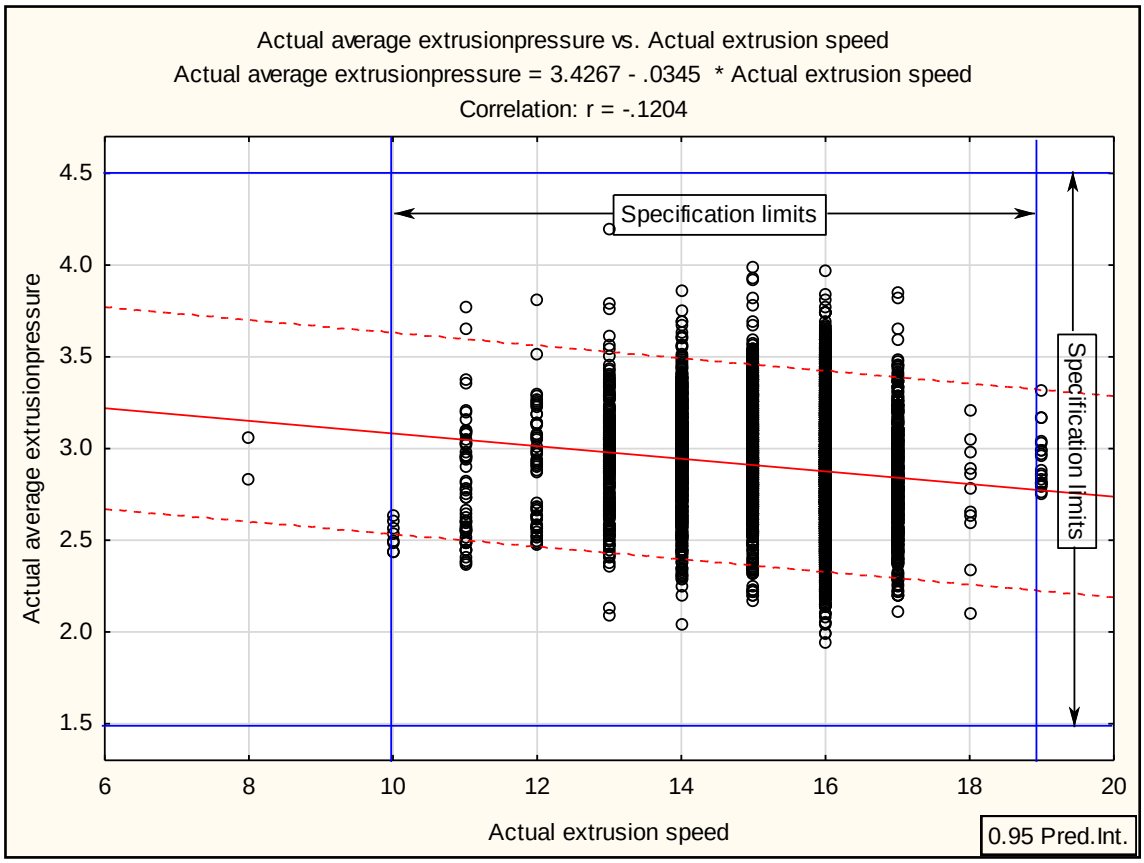
Single variable alternative hypothesis for Actual extrusion speed

There is a correlation between Actual extrusion speed and Average extrusion pressure.

Correlation between Actual extrusion speed and Average extrusion pressure is -12.04%. (See Graph 8.17). The associated p-value of no correlation between Actual extrusion speed and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

Conclusion

There is a statistically significant correlation between Actual extrusion speed and average extrusion pressure.



Graph 8.17: Scatter plot estimated period: (Actual extrusion speed - Screened)

For variable 7, the hypothesis shows a statistically significant correlation between Actual extrusion speed and average extrusion pressure. As an independent variable the negative relationship (r = -0.1204), with negative influence of -0.0345 on pressure for the dependent variable with every unit increase of speed for the independent variable, confirms a weak predictor. Table 8.4 for the regression model shows this variable as not significant with a positive influence of 0.00312 on pressure for the dependent variable with every unit increase of speed for the independent variable. Therefore, variable 7 is a weak predictor for the regression model, and as an independent variable, but only a significant predictor as an independent variable.

N=10048	Regression Summary for Dependent Variable: Actual average extrusionpressure (y) - Test period R= .54762173 R²= .29988956 Adjusted R²= .29940144 F(7,10040)=614.37 p<0.0000 Std.Error of estimate: .3967					
	b*	Std.Err. of b*	b	Std.Err. of b	t(10040)	p-value
Intercept			11.9624	0.74641	16.027	0.00000
1. Mix discharge temp (x)	-0.03390	0.00838	-0.0161	0.00398	-4.045	0.00005
2. Cool begin temp (x)	0.15954	0.00882	0.0160	0.00088	18.094	0.00000
3. Actual cool time (x)	0.01978	0.00903	0.0049	0.00223	2.190	0.02857
4. Actual dump temp (x)	0.03991	0.00889	0.0115	0.00255	4.488	0.00001
5. Actual tamp pressure (x)	-0.41084	0.00859	-1.3304	0.02782	-47.820	0.00000
6. Actual extrusion rate (x)	-0.13591	0.00907	-0.0269	0.00179	-14.989	0.00000
7. Actual extrusion speed (x)	-0.12690	0.00899	-0.0447	0.00317	-14.117	0.00000

Table 8.5: Individual regression summary validation period – Average pressure

Table 8.5 represents MR summary statistics for the validation period of data. With the F statistic significant, (p>0.000), the independent variables are considered to be useful in predicting the dependent variable. All seven variables in red (Mix discharge temperature, Cool begin temp, Actual cool time, Actual dump temp, Actual tamp pressure, Actual extrusion rate and Actual extrusion speed) are statistically significant, with p values of close to 0.000 for all of the seven variables respectively, calculated for p = 0.05. The R = 0.54762 indicates a good relationship of the independent variables to the dependent variable, but not ideal. An r = 1.0 is a perfect or ideal relationship. These seven variables are significant for the MR model that statistically influences the output of this model. The formula for this MR model, read for Table 8.5, is:

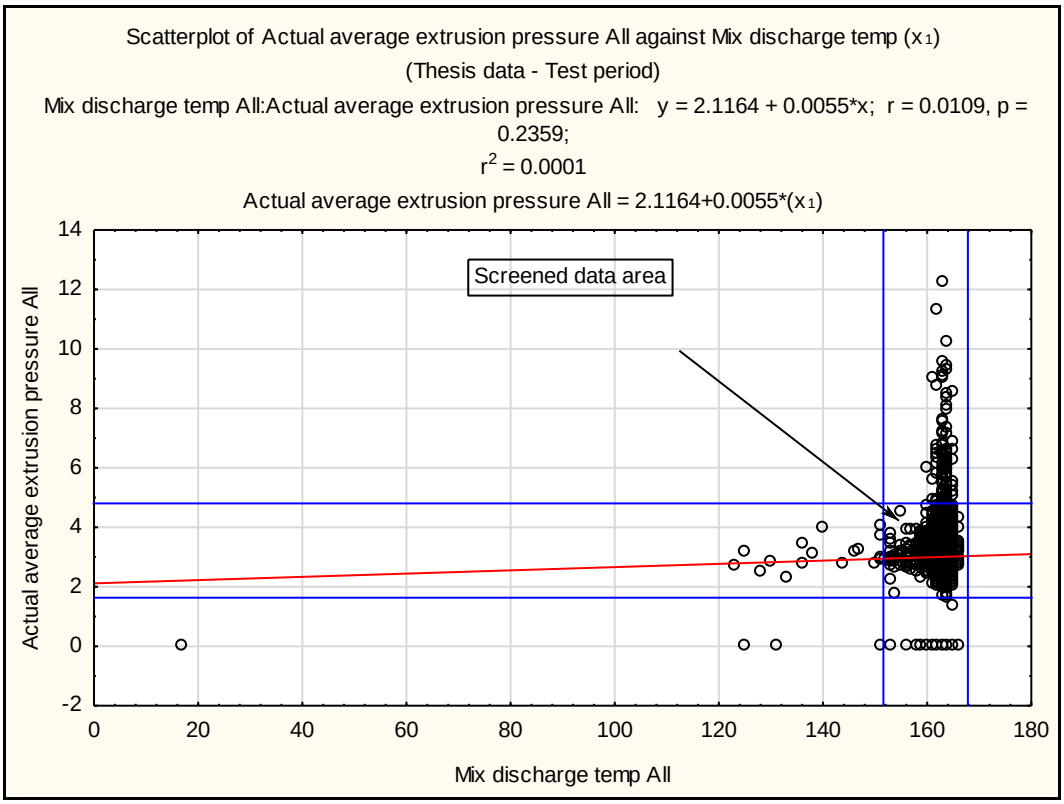
$y = 11.9624 - 0.01611*x_1 + 0.01596*x_2 + 0.00487x_3 + 0.01145*x_4 - 1.33038*x_5 - 0.02689*x_6 - 0.04474*x_7$.

Multicollinearity tests are not applicable because for this study, the validation period is for comparison purposes only and will not be used for prediction.

In addition to the MR summary for the **validation period**, Table 8.24, scatter plots for each independent variable represented in the summary show the graphical relationship between each independent variable and dependent variable. ***These scatter plots are for comparison purposes to the estimated period to evaluate prediction consistency between the two periods***. Showing the effect of screening data following the proposed screening phases in chapter 4, section 4.4 was **firstly** to draw scatterplots for each independent variable showing raw unscreened data for each independent

variable compared to raw dependent variable data. **Secondly**, to draw scatterplots for each independent variable showing screened independent data for each independent variable compared to dependent variable data.

The graphical effect follows below for the validation period, with regression analysis and hypothesis tests done on screened data. In addition, raw data scatter plots for each independent variable were done, showing the screened portion form the raw data.



Graph 8.18: Scatter plot validation period: (Mix discharge temperature (x_1) - Raw)

8.5.8 Hypothesis test second period – Mix discharge temperature (x_1)

Single variable null hypothesis for Mix discharge temperature

There is no correlation between Mix discharge temperature and Average extrusion pressure.

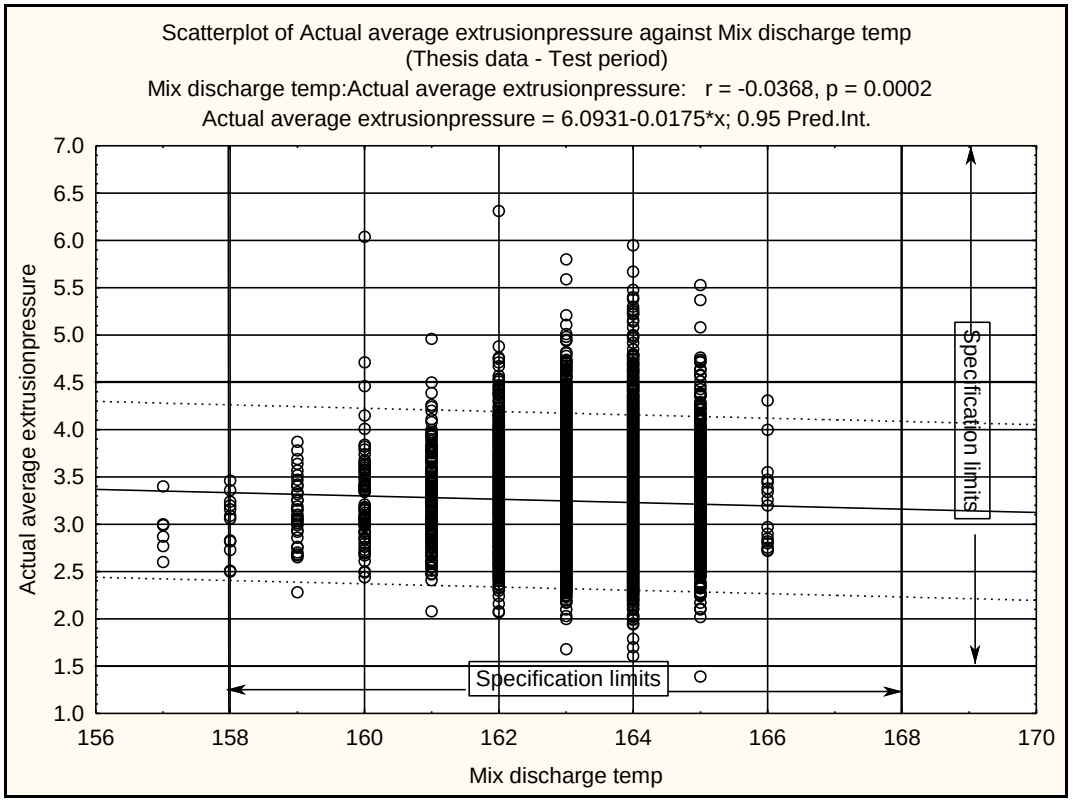
Single variable alternative hypothesis for Mix discharge temperature

There is a correlation between Mix discharge temperature and Average extrusion pressure.

Correlation between Mix discharge temperature and Average extrusion pressure is -3.68%. (See Graph 8.19). The associated p-value of no correlation between Mix discharge temperature and Average extrusion pressure is 0.00022, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

Conclusion

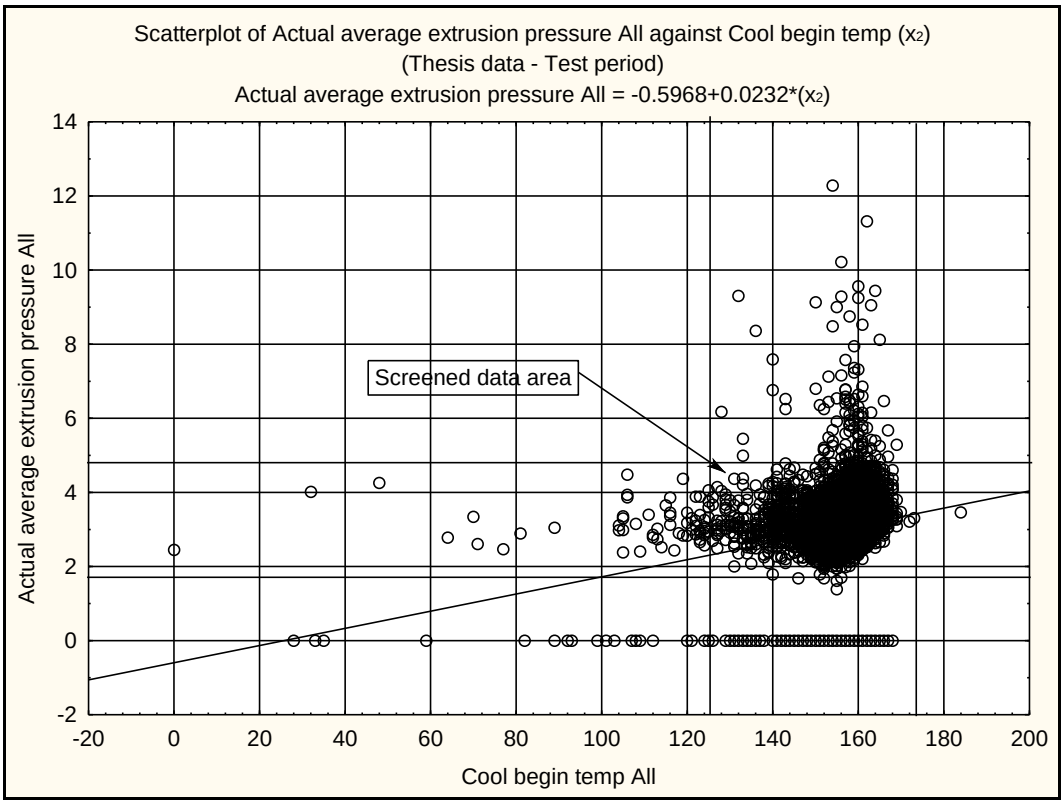
There is a statistically significant correlation between Mix discharge temperature and average extrusion pressure.



Graph 8.19: Scatter plot validation period: (Mix discharge temperature - Screened)

For variable 1, the hypothesis shows a statistically significant correlation between Mix discharge temperature and average extrusion pressure. As an independent variable the weak negative relationship (r = -0.0368), with negligible negative influence of 0.0175 on pressure for the dependent variable with every degree increase of temperature for the independent variable confirms a weak predictor. Table 8.5 shows this variable significant with a negligible negative influence of -0.01611 on pressure for the dependent variable with every degree increase of temperature for the independent

variable. Therefore, variable 1 is a weak significant predictor for the MR model, and as an independent variable.



Graph 8.20: Scatter plot validation period: (Cool begin temperature (x_2) - Raw)

8.5.9 Hypothesis test second period– Cool begin temperature (x_2)

Single variable null hypothesis for Cool begin temperature

There is no correlation between Cool begin temperature and Average extrusion pressure.

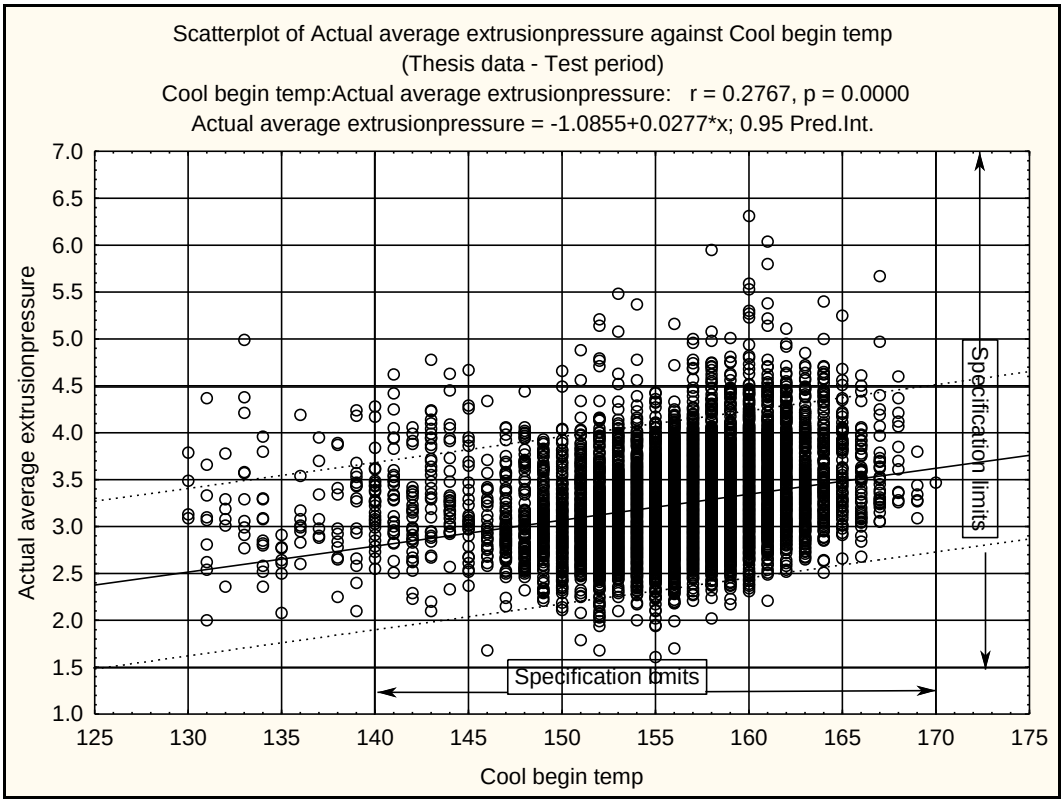
Single variable alternative hypothesis for Cool begin temperature

There is a correlation between Cool begin temperature and Average extrusion pressure.

Correlation between Cool begin temperature and Average extrusion pressure is 27.670%. (See Graph 8.21). The associated p-value of no correlation between Cool begin temperature and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

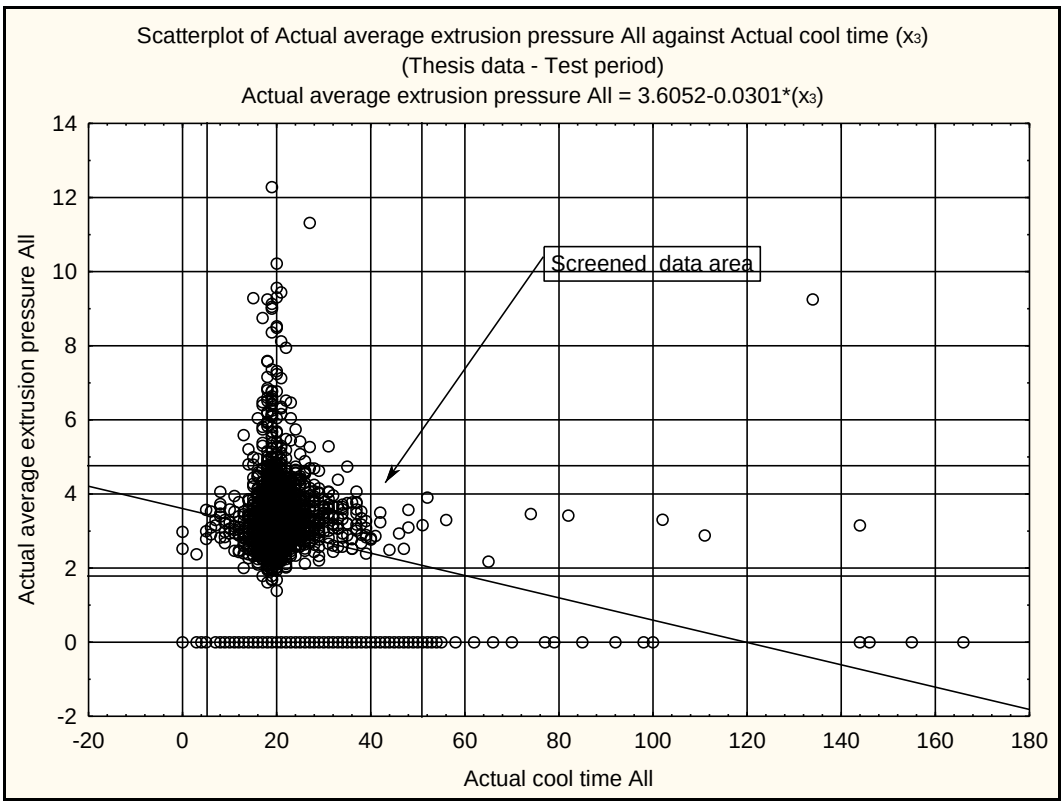
Conclusion

There is a statistically significant correlation between Cool begin temperature and average extrusion pressure.



Graph 8.21: Scatter plot validation period: (Cool begin temperature - Screened)

For variable 2 the hypothesis shows a statistically significant correlation between Cool begin temperature and average extrusion pressure. As an independent variable the fair positive relationship (r = 0.27670), with fair **positive** influence of 0.02769 on pressure for the dependent variable with every degree increase of temperature for the independent variable confirms a fair predictor. Table 8.5 shows this variable as significant with a negligible **positive** influence of 0.01596 on pressure for the dependent variable with every degree increase of temperature for the independent variable. Therefore, variable 2 is a fair significant predictor for the MR model, and as an independent variable.



Graph 8.22: Scatter plot validation period: (Actual cooling time(x_3) - Raw)

8.5.10 Hypothesis test second period – Actual cooling time (x_3)

Single variable null hypothesis for Actual cooling time

There is no correlation between Actual cooling time and Average extrusion pressure.

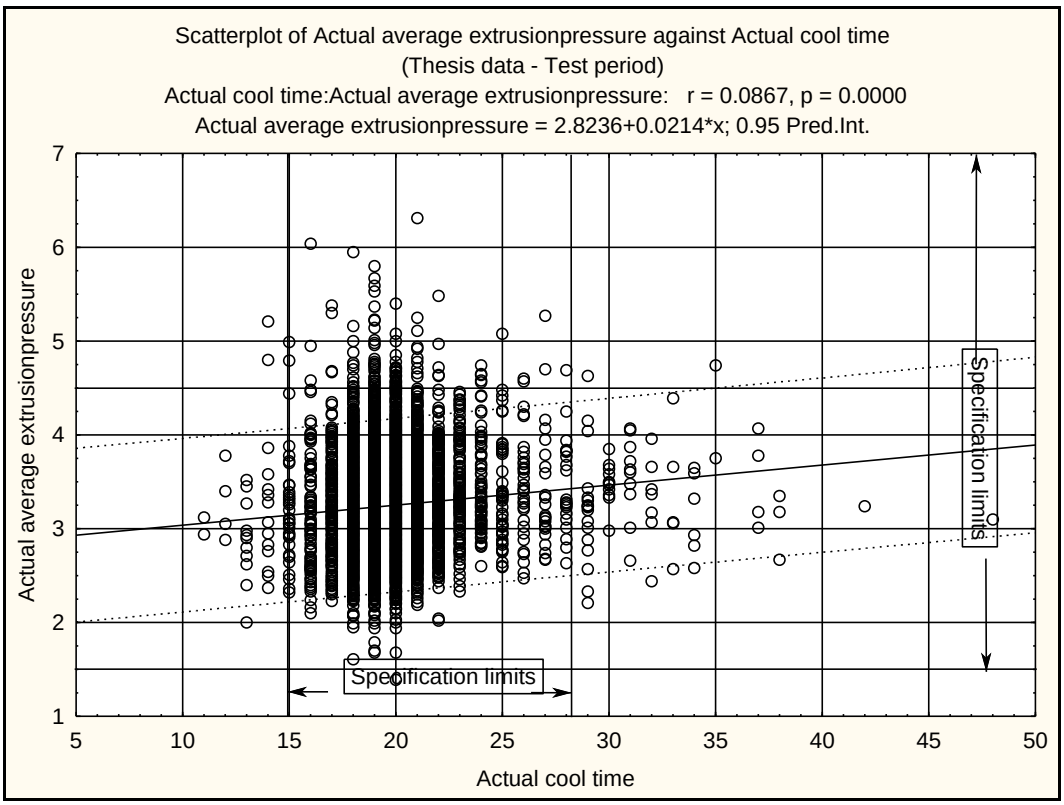
Single variable alternative hypothesis for Actual cooling time

There is a correlation between Actual cooling time and Average extrusion pressure.

Correlation between Actual cooling time and Average extrusion pressure is 8.674%. (See Graph 8.23). The associated p-value of no correlation between Actual cooling time and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

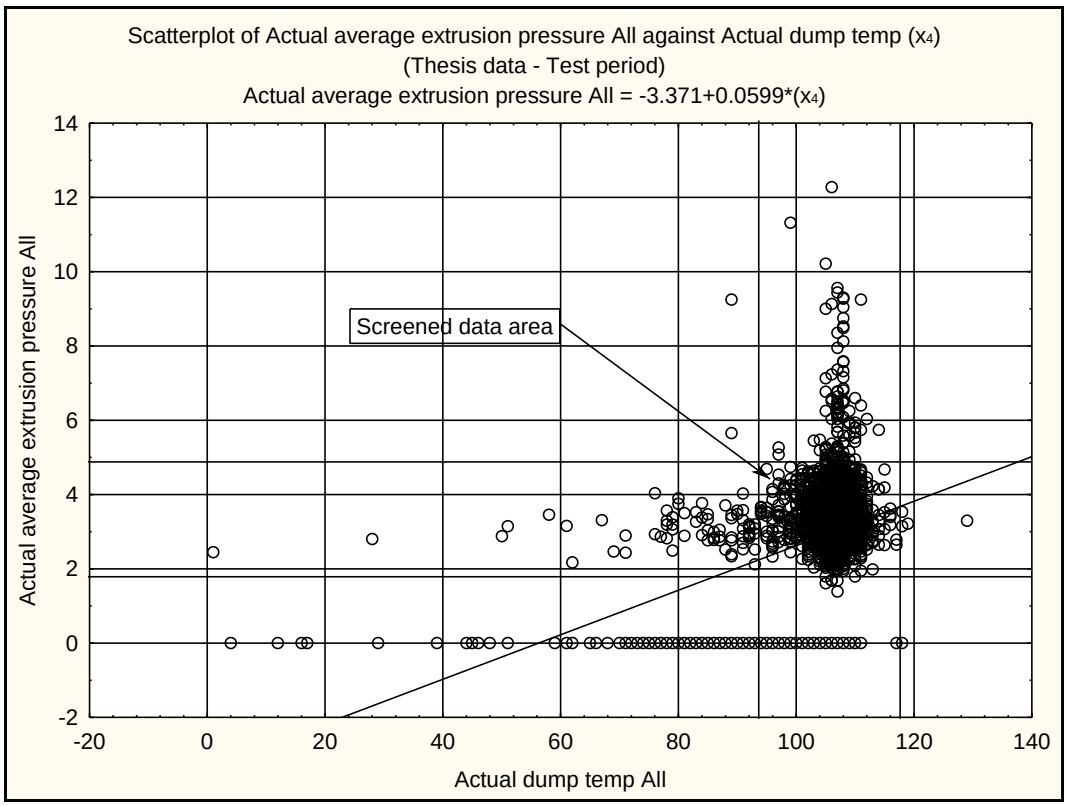
Conclusion

There is a statistically significant correlation between Actual cooling time and average extrusion pressure.



Graph 8.23: Scatter plot validation period: (Actual cooling time - Screened)

For variable 3, the hypothesis shows a statistically significant correlation between Actual cooling time and average extrusion pressure. As an independent variable the weak positive relationship (r = 0.08674), with negligible **positive** influence of 0.02136 on pressure for the dependent variable with every minute increase in time for the independent confirms a weak predictor. Table 8.5 shows this variable as significant with a negligible **positive** influence of 0.00487 on pressure for the dependent variable with every minute increase in time for the independent variable. Therefore, variable 3 is a weak significant predictor for the MR model, and as an independent variable.



Graph 8.24: Scatter plot validation period: (Actual dump temperature (x_4)- Raw)

8.5.11 Hypothesis test second period – Actual dump temperature (x_4)

Single variable null hypothesis for Actual dump temperature

There is no correlation between Actual dump temperature and Average extrusion pressure.

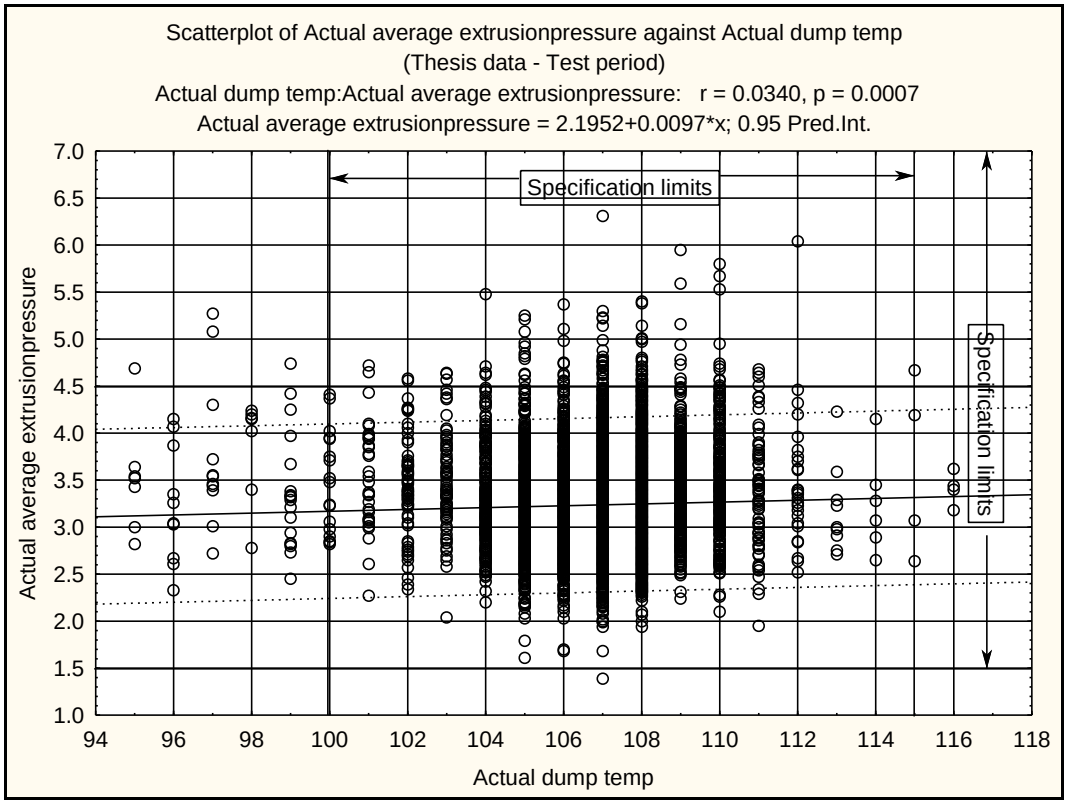
Single variable alternative hypothesis for Actual dump temperature

There is a correlation between Actual dump temperature and Average extrusion pressure.

Correlation between Actual dump temperature and Average extrusion pressure is 3.396%. (See Graph 8.25). The associated p-value of no correlation between Actual dump temperature and Average extrusion pressure is 0.0007, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

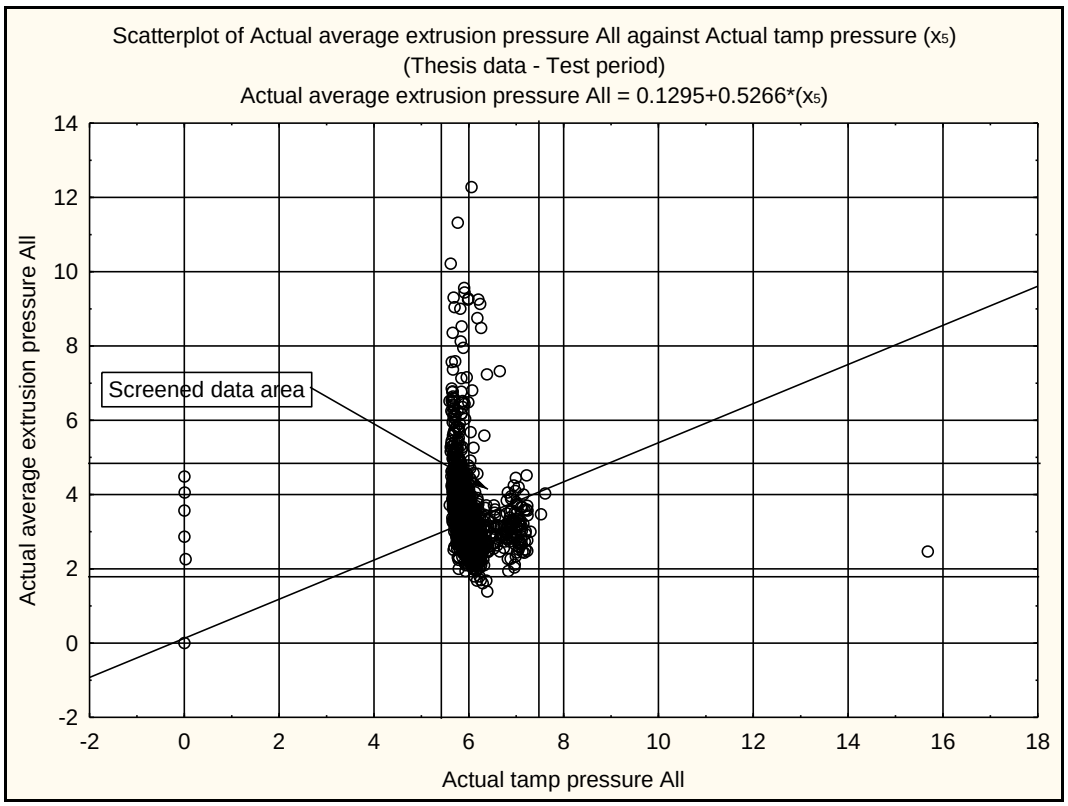
Conclusion

There is a statistically significant correlation between Actual dump temperature and average extrusion pressure.



Graph 8.25: Scatter plot validation period: (Actual dump temperature - Screened)

For variable 4, the hypothesis shows a statistically significant correlation between Actual dump temperature and average extrusion pressure. As an independent variable the weak positive relationship (r = 0.03396), with negligible **positive** influence of 0.00974 on pressure for the dependent variable with every degree increase of temperature for the independent variable confirms a weak predictor. Table 8.5 shows this variable as significant with a negligible **positive** influence of 0.01145 on pressure for the dependent variable with every degree increase of temperature for the independent variable. Therefore, variable 4 is a weak significant predictor for the MR model, and as an independent variable.



Graph 8.26: Scatter plot validation period: (Actual tamp pressure (x_5) - Raw)

8.5.12 Hypothesis test second period – Actual tamp pressure (x_5)

Single variable null hypothesis for Actual tamp pressure

There is no correlation between Actual tamp pressure and Average extrusion pressure.

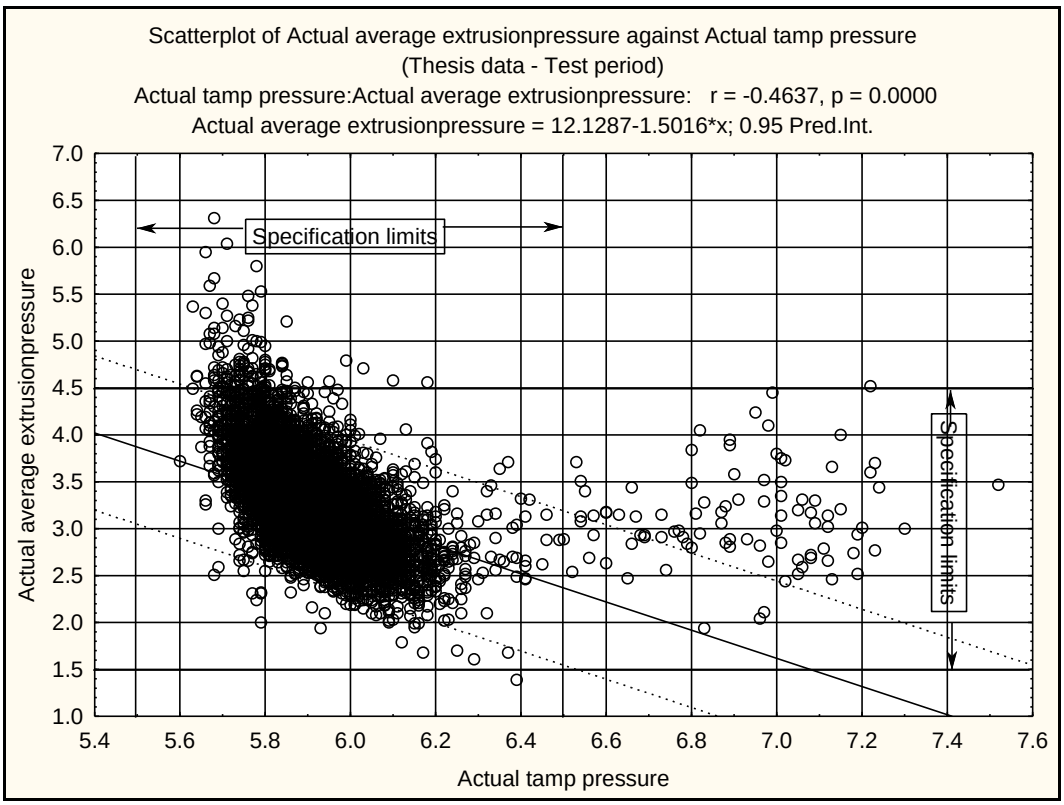
Single variable alternative hypothesis for Actual tamp pressure

There is a correlation between Actual tamp pressure and Average extrusion pressure.

Correlation between Actual tamp pressure and Average extrusion pressure is -46.37%. (See Graph 8.27). The associated p-value of no correlation between Actual tamp pressure and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

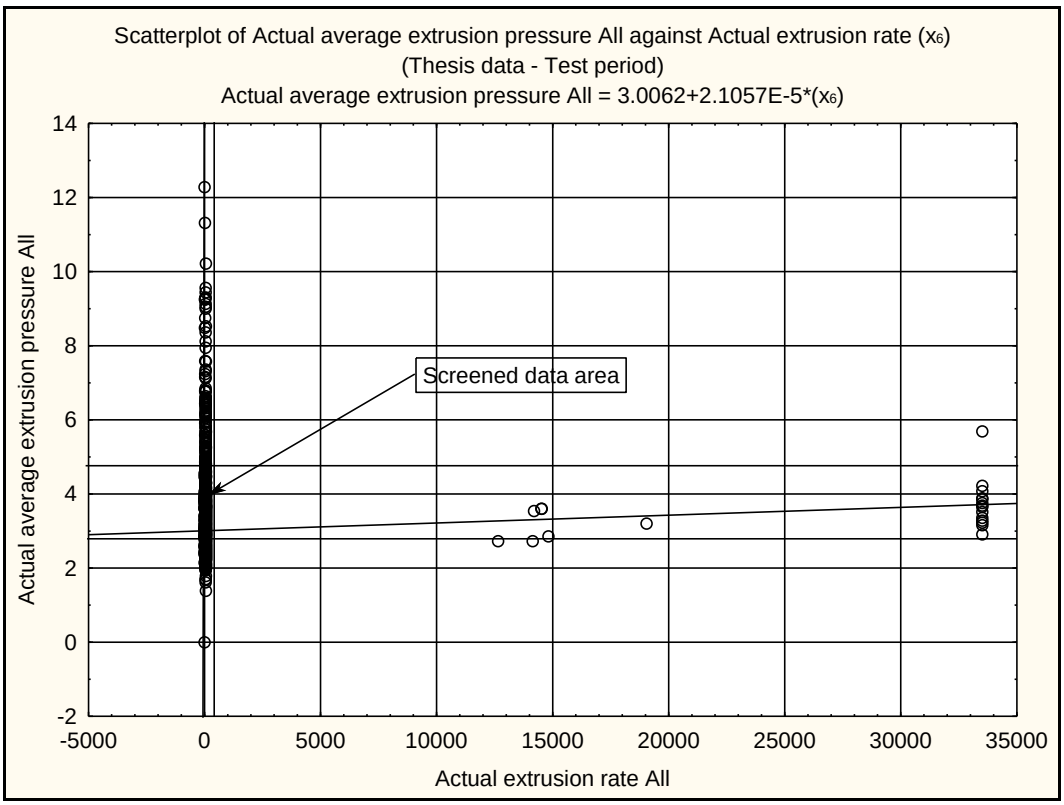
Conclusion

There is a statistically significant correlation between Actual tamp pressure and average extrusion pressure.



Graph 8.27: Scatter plot validation period: (Actual tamp pressure - Screened)

For variable 5, the hypothesis shows a statistically significant correlation between Actual tamp pressure and average extrusion pressure. As an independent variable the strong negative relationship (r = -0.4637), with a strong negative influence of -1.502 on pressure for the dependent variable with every unit increase of pressure for the independent variable, confirms a strong predictor. Table 8.5 shows this variable as significant with a strong negative influence of -1.33038 on pressure for the dependent variable with every unit increase of pressure for the independent variable. Therefore, variable 5 is a strong and a significant predictor for MR model and as an independent variable. Variable 5 shows the strongest regression relationship for both the MR model and as an independent variable.



Graph 8.28: Scatter plot validation period: (Actual extrusion rate (x_6) - Raw)

8.5.13 Hypothesis test second period – Actual extrusion rate (x_6)

Single variable null hypothesis for Actual extrusion rate

There is no correlation between Actual extrusion rate and Average extrusion pressure.

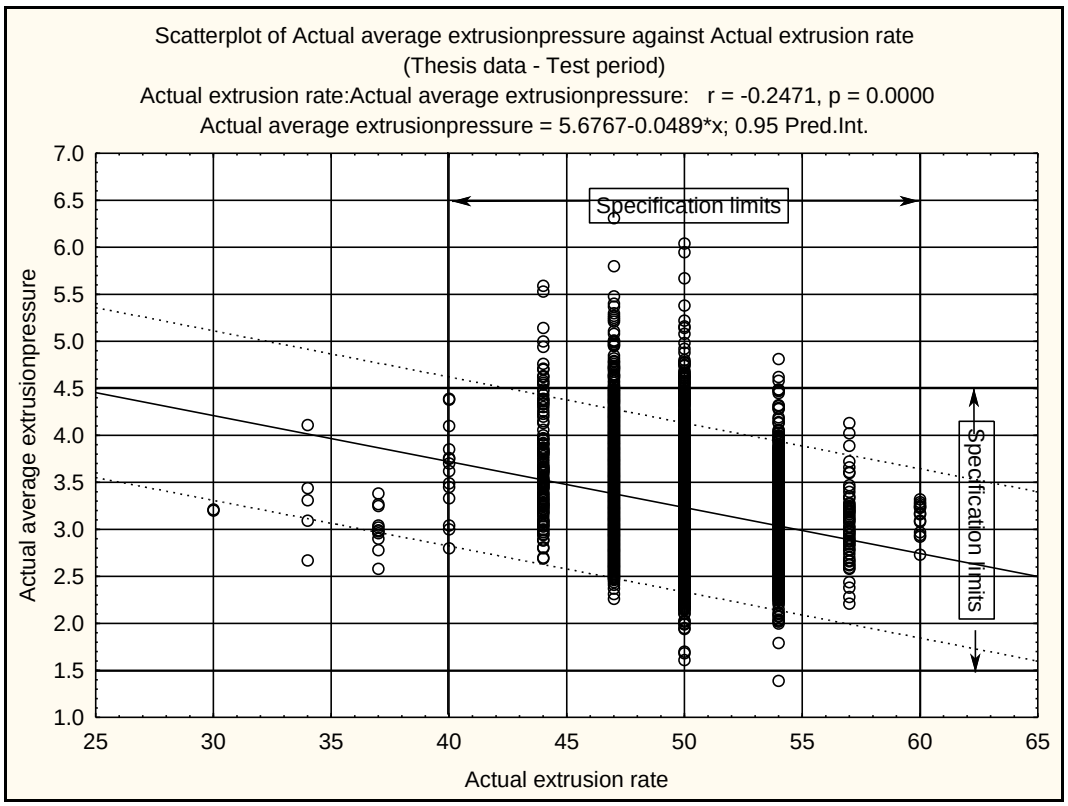
Single variable alternative hypothesis for Actual extrusion rate

There is a correlation between Actual extrusion rate and Average extrusion pressure.

Correlation between Actual extrusion rate and Average extrusion pressure is -24.71%. (See Graph 8.29). The associated p-value of no correlation between Actual extrusion rate and Average extrusion pressure is 0.0001, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

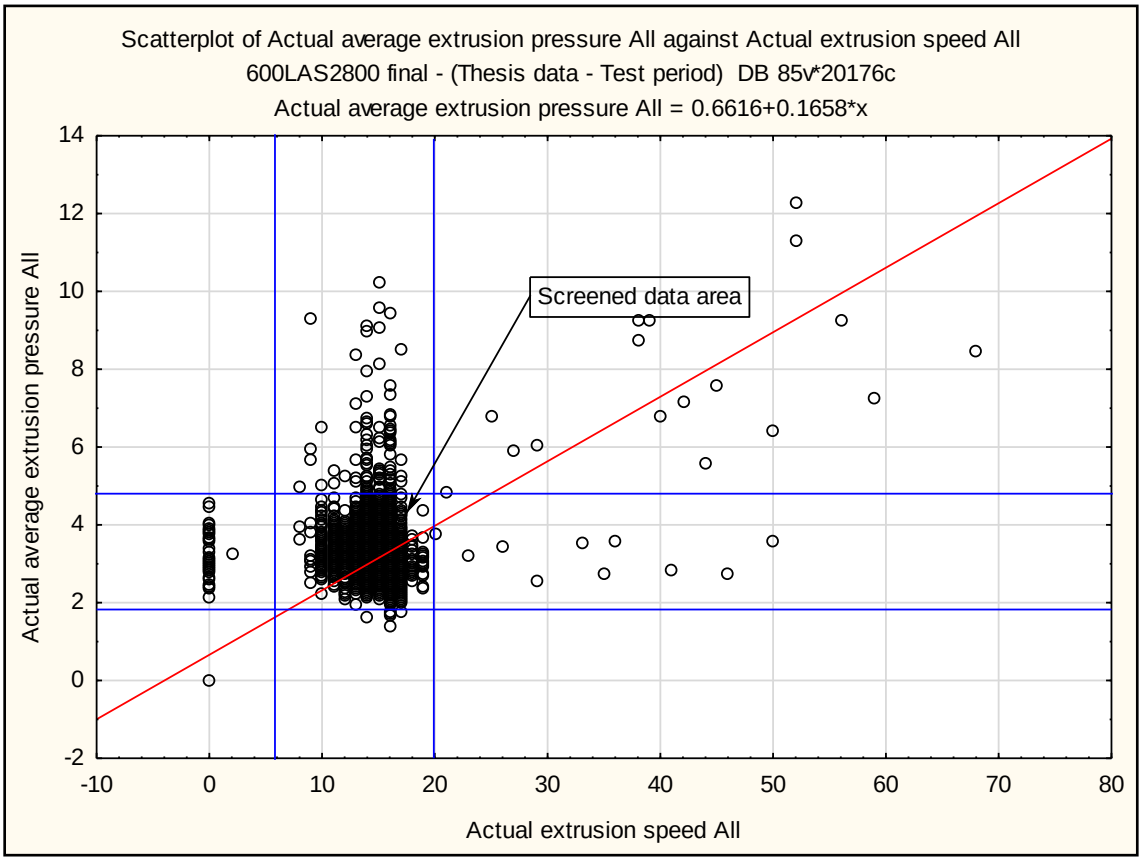
Conclusion

There is a statistically significant correlation between Actual extrusion rate and average extrusion pressure.



Graph 8.29: Scatter plot validation period: (Actual extrusion rate - Screened)

For variable 6 the hypothesis shows a statistically significant correlation between Actual extrusion rate and average extrusion pressure. As an independent variable the negative relationship (r = -0.2471), with **negative** influence of -0.0489 on pressure for the dependent variable with every unit increase for the independent variable shows a fair predictor. Table 8.5 shows this variable as significant with a **negative** influence of -0.02689 on pressure for the dependent variable with every unit increase of temperature for the independent variable. Therefore, variable 6 is a weak significant predictor for the MR model and as an independent variable.



Graph 8.30: Scatter plot validation period: (Actual extrusion speed (x_7) - Raw)

8.5.14 Hypothesis test second period – Actual extrusion speed (x_7)

Single variable null hypothesis for Actual extrusion speed

There is no correlation between Actual extrusion speed and Average extrusion pressure.

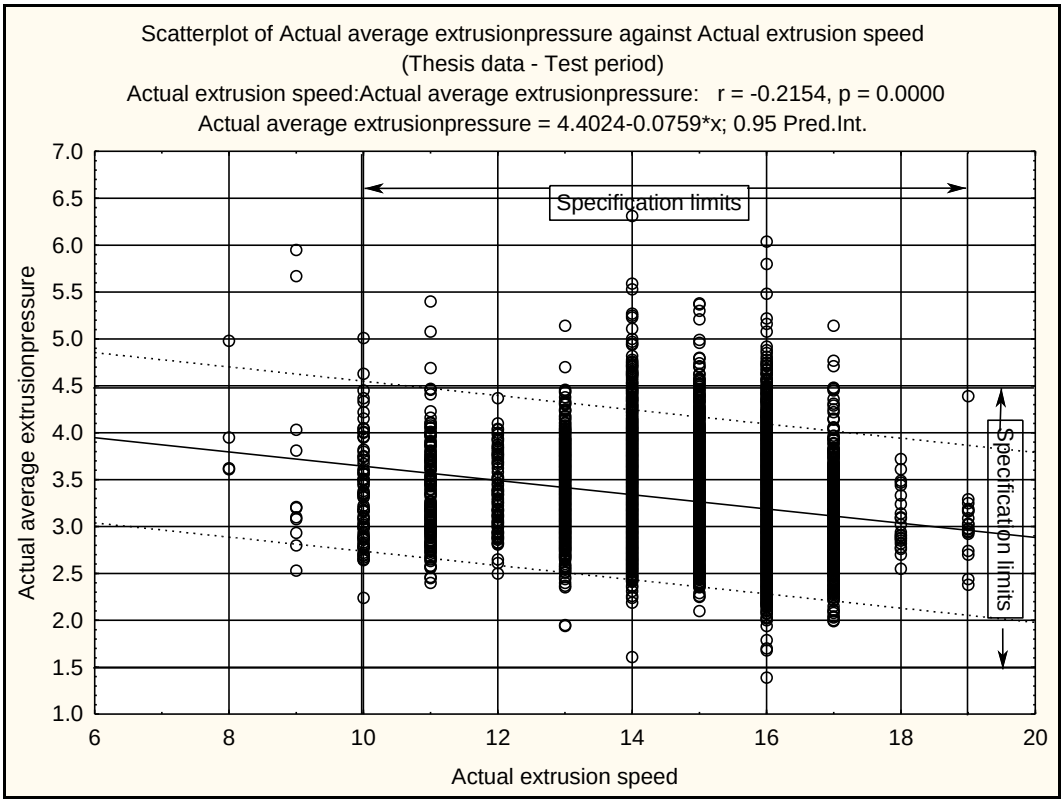
Single variable alternative hypothesis for Actual extrusion speed

There is a correlation between Actual extrusion speed and Average extrusion pressure.

Correlation between Actual extrusion speed and Average extrusion pressure is -21.54%. (See Graph 8.31). The associated p-value of no correlation between Actual extrusion speed and Average extrusion pressure is 0.00022, which is smaller than the significance level of α=0.05. The null hypothesis will be rejected and therefore the correlation is statistically significant for a significance level of 0.05.

Conclusion

There is a statistically significant correlation between Actual extrusion speed and average extrusion pressure.



Graph 8.31: Scatter plot validation period: (Actual extrusion speed - Screened)

For variable 7, the hypothesis shows a statistically significant correlation between Actual extrusion speed and average extrusion pressure. As an independent variable the negative relationship (r = -0.2154), with **negative** influence of -0.0759 on pressure for the dependent variable with every unit increase of speed for the independent variable shows a fair predictor. Table 8.5 shows this variable as significant with a **negative** influence of -0.0759 on pressure for the dependent variable with every unit increase of speed for the independent variable. Therefore, variable 7 is a fair significant predictor for the MR model and as an independent variable

8.6 CONCLUSIONS

The analytical approach initially started with a top down approach using MR analysis to evaluate the significance of independent variable contribution within a MR model. During the analytical process the need for a bottom up approach became evident for

individually evaluating the significance of each independent variable to the process output. By doing both approaches, a holistic analytical approach was applied.

By analysing individual independent variables, weak collinearity exists that suggests little interaction amongst variables. For this reason, the proposed prediction model was based on one independent variable that individually had the largest impact on process output. For this study, it was identified as variable 5.

A progressive reduction process was followed to find the “best” fit for the screen data for the estimation period to predict the validation period. This process started with a MR model, then individual linear regression, then a polynomial fit, then a combination of polynomial and linear regression. Visually the combined model fits the best, especially if the validation period is used as the model base. Refer to model in 8.4.

Additional models to evaluate screened data, which is part of future, work and not included for this study are:

- A fit for inverse of x compared to y
- A third order polynomial fit
- A log transformation model
- Fit proposed and existing models on screened data for the validation period that only include data applicable for the changed raw material.

Only two variables were not significant for both periods using the same dependent variable. This shows process repeatability and operating consistency for both periods.

The sensitivity of effects (b values) between the two periods for each independent variable differs but is not applicable for this study because we were only interested if the same independent variables are significant between the two periods. The reasons for different slope values will add little value to this study and may be evaluated at a later stage.

A comparison between the MR and DOE model regression results for individual independent variables was necessary for evaluating for collinearity by showing the significance and slope for each independent variable. The proposed operating levels for each independent variable by DOE regression may still include some collinearity when

slope swapping occurs between MR and DOE. This is discussed in chapter 10 with reference to Table 10.3.

DOE regression is the main prediction technique, because it is the main focus of this study and is based on a multivariable platform that incorporates different independent variables on different processing levels. Normal multiple regressions were used as a comparative technique to evaluate the similarities and differences between these two techniques.

The excel database was folded into 32 different DOE runs with the seven selected independent variables, for both periods, at determined min and max levels. Both DOE regression and multiple regression were applied on the individual data in the excel database for both periods, then grouped into the same 32 DOE runs. The average of the predicted individual data was then calculated. This will be the comparison basis between normal and DOE regression for the first and second processing period.

Prediction comparisons between DOE and normal multiple regression for the two periods were done to determine the prediction accuracy, see Table 8.6.

Variables	Normality Test	SPC	DOE (50/50) Median split	Regressions				DOE model
	Kolmogorov -Smirnov (p>0.1) (Estimated period)			MR Estimated period	MR Validation period	DOE Estimated period	DOE Validation period	
				t (P<0.5)	t (P<0.5)	t (P<0.5)	t (P<0.5)	
Mix discharge temp (x_1)	0.27273	Yes	Yes	-0.5451	**-4.0452**	1.3969	-1.6931	**Yes**
Cool begin temp (x_2)	0.11551	Yes	Yes	**-6.8485**	**18.0936**	-0.7802	5.1754	**Yes**
Cool time (x_3)	0.24063	Yes	Yes	**5.8970**	**2.1897**	-0.2851	**-1.6070**	**Yes**
Dump temp (x_4)	0.24000	Yes	Yes	**15.2337**	**4.4883**	**2.8022**	0.2409	**Yes**
Tamp pressure (x_5)	0.07029	Yes	Yes	**-43.3532**	**-47.8198**	**-7.9890**	**-11.1409**	**Yes**
Extrusion rate (x_6)	0.34537	Yes	Yes	**-13.6042**	**-14.9890**	**-3.5531**	**-4.4687**	**Yes**
Extrusion speed ($_7$)	0.38480	Yes	Yes	1.0875	**-14.1170**	**-2.6350**	**-5.7512**	**Yes**

Table 8.6: Regression and normality test summary for independent variables

Table 8.6 represents a summary of all independent and dependent variables showing the measuring characteristics for individual and DOE regression, normality tests, SPC, median split, and independent variable selection for model. Significance was measured on $\alpha < 0.05$. A summary of Table 8.6 is as follows:

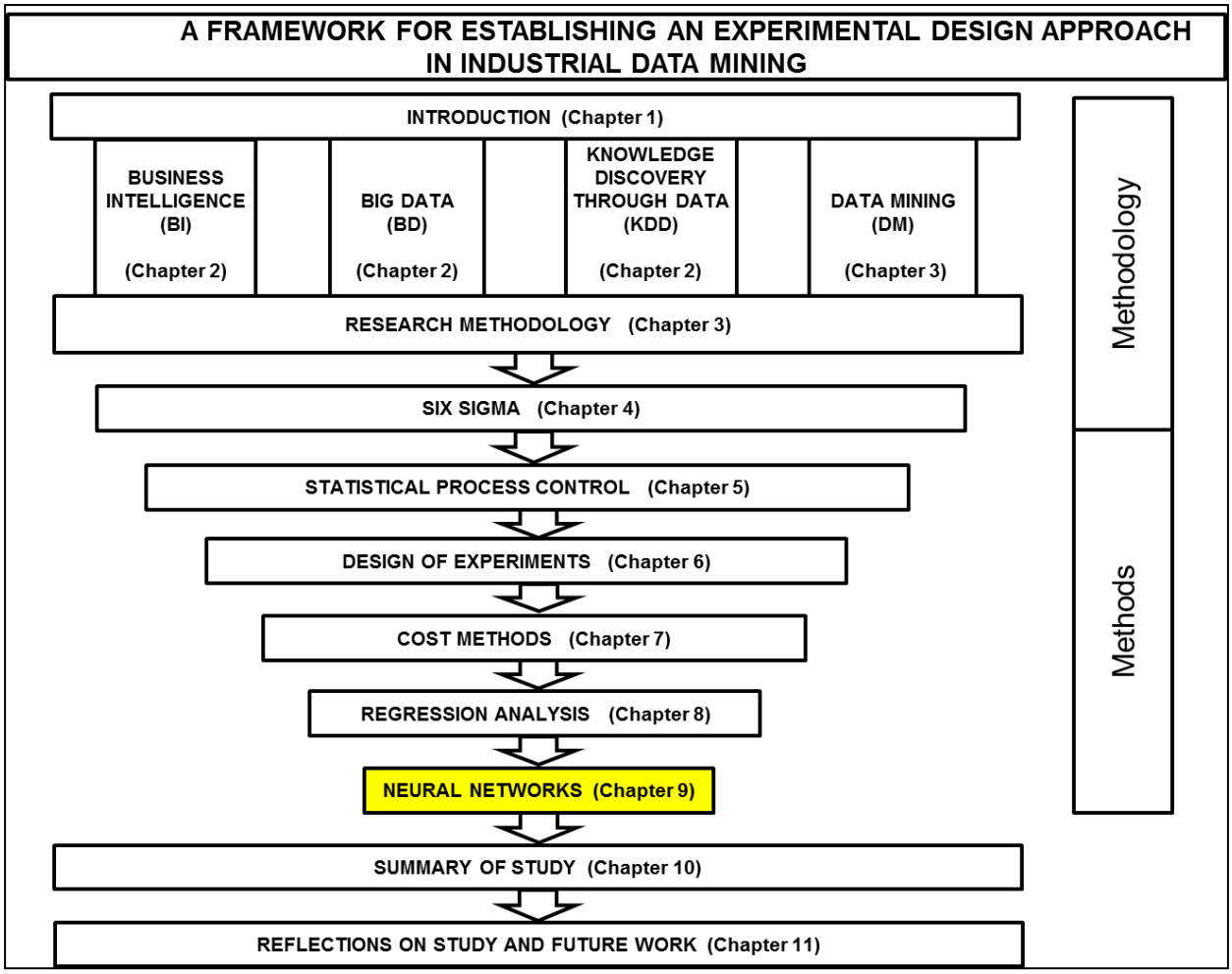
None of the variables passes the normality test. This is also evident in the SPC charts (x-bar and s) in chapter 5. To pass the normality test is not a prerequisite for any variable for a prediction model. It is, however, a guide in flagging the risk of having missing values when a database collapses into DOE experimental runs when a full factorial design is selected. Because none of the variables passed, a full factorial was not practical, therefore a standard 2** (7-2) resolution IV design was used because of the missing value limitation of the database. A resolution IV provided data for 32 runs instead of 128. Also keep in mind that the database was split in two based on the median for each independent variable.

All variables that could be split with the median, and gave close to 50-50 percent data points, irrespective of regression significance, were chosen for the DOE model. The significance for the independent variables with DOE regression for different operation levels will be established later.

The proposed model for process development could not be tested because the company has shut its operations in South Africa. The concept for the proposed DOE methodology proved to be representative for the period upon which the model was developed and tested, based on all the different the comparative results between the predictive model and the validation period. For the methodology to be effective for any process improvement, it is critical that the independent and dependent variables remain within statistical process control and process specification. Stability of the production process in the long-term is critical for consistent high prediction accuracy.

CHAPTER 9

NEURAL NETWORKS



9.1 INTRODUCTION

The aim of this chapter is to compare the results of Neural networks applications to the same database as the conventional statistical methods applied, using the proposed framework as a guide. Comparisons are only to illustrate the compatibility between the two data analytical approaches and not to choose one approach above the other.

Data Mining applications through Neural networks is becoming increasingly popular as a management tool to explore knowledge that can guide strategic decisions. For this reason the comparison is necessary to complete the process analytics for this research.

9.2 DESCRIPTION OF DATA MAINING TECHNIQUE

The ultimate goal of data mining is prediction, also called predictive DM. Predictive DM is the most common type of data mining and has the most direct business applications.

Data mining is an analytical process designed to explore large data sets, quantitative or qualitative, to search for patterns and/or systematic relationships between variables. The findings are then validated by applying the detected patterns to new subsets of data.

The main difference between data mining and traditional Exploratory Data Analysis (EDA) is that Data Mining is more applications orientated and does not identify specific relations between variables. DM focuses on producing solutions that can generate predictions for business. The DM analytical modelling is done through a "black box" approach for data exploration or knowledge discovery, and uses not only the traditional Exploratory Data Analysis techniques, but also techniques such as neural networks.

The ability of neural networks to learn by examples is one of the many features that enable the analyst to model data and establish accurate rules governing the underlying relationship between various data attributes. Neural network uses training algorithms, which can automatically learn the structure of the data presented by the analyst. This unique analytical feature of neural networks makes it a popular DM technique for analysts as a predictive model. Although users need to have knowledge of how to select and prepare data, how to select the appropriate neural network, and how to interpret the results, the level of user knowledge needed to successfully apply neural networks is much lower than those needed in most traditional statistical tools and techniques. This is because neural network algorithms are hidden in a “black box” within computer programs. During the analytical process, Neural networks derive and extract meaning, rules, and trends from complicated data sets. Neural networks use complicated mathematical functions that are too difficult, if not impossible, to model using analytic or parametric techniques.

Because of the broad applicability, neural networks are suitable for applications of real world problems in research and science, business, and industry. Areas where neural networks have been successfully applied are Signal processing, Process control, Robotics, Classification, Data pre-processing, Pattern recognition, Image and speech

analysis, Medical diagnostics and monitoring, Stock market and forecasting and Loan or credit solicitations.

Neural networks consists of three basic stages:

Stage 1: Exploration. This stage usually starts with data preparation that may involve cleaning data, data transformations, selecting subsets of records, and, in case of data sets with large numbers of variables, screening of these variables to work only with those variables that add value to the process analysed.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one, based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples).

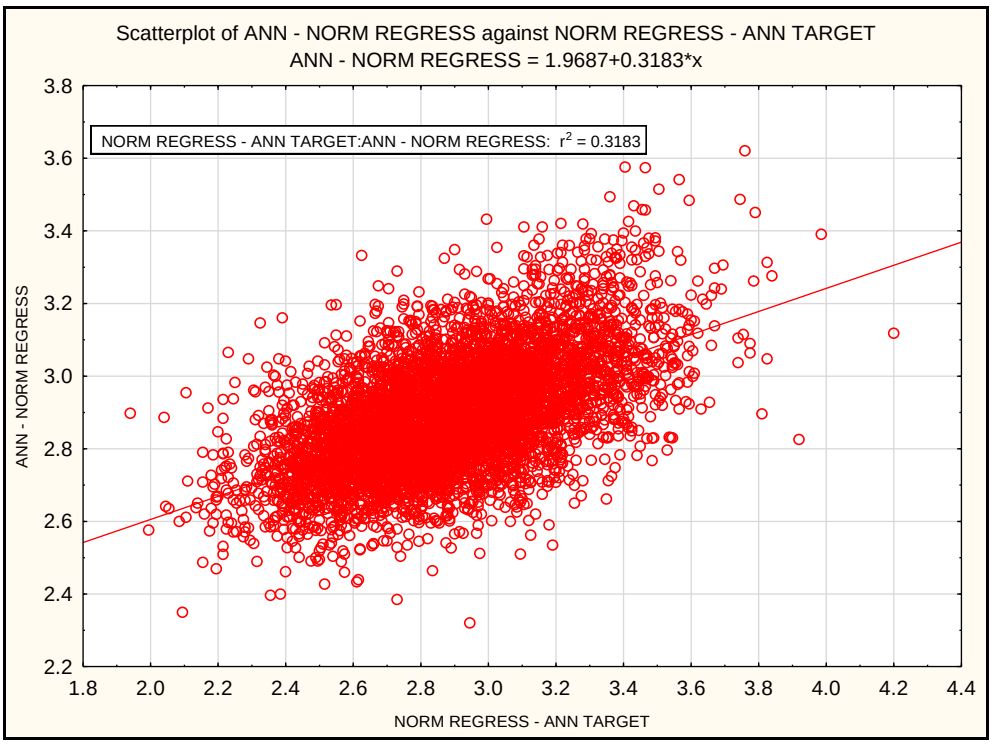
Stage 3: Deployment. This final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

9.3 DATA MINING ANALYSIS

Neural networks were used for the DM application because of their predictive qualities; they are closely related to regression analysis. A comparison is made between the multiple regression and the neural networks (NN) results applied to the same database. The goal of these comparisons is to evaluate if neural networks (NN) provide similar results as normal statistical exploratory data analysis for multiple regression analysis.

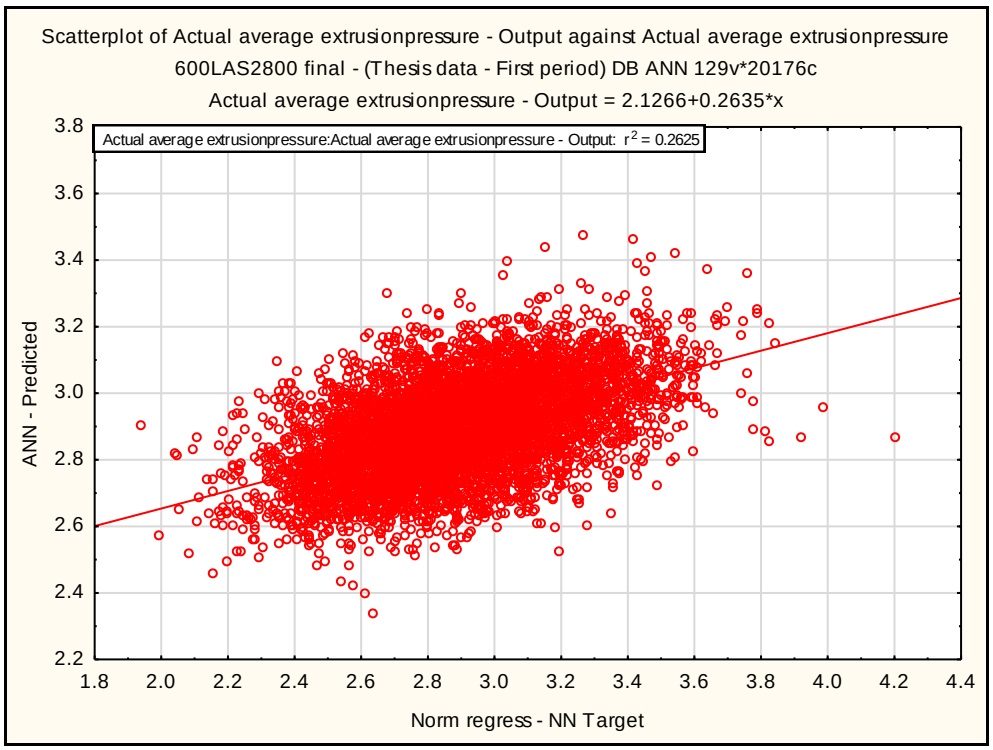
The prediction model for NN was used for screening independent variables for DOE selection. Statistical process control (SPC) was used for screening independent variables for DOE selection with normal exploratory data analysis. Doing variable screening through SPC gives the analyst a graphical representation of how data for each independent variable are distributed.

These comparisons are for illustrations only and not to decide which application is better than the other.



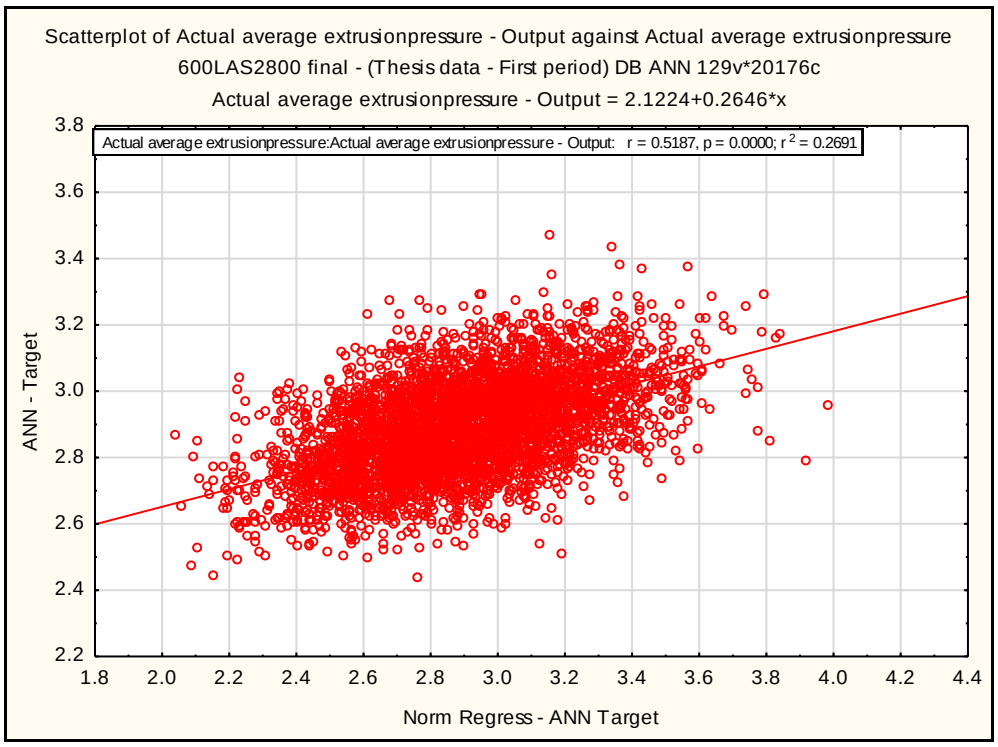
Graph 9.1: Neural network (NN) regression – Phase 1

Graph 9.1 represents the NN regression analysis for the seven independent variables and one dependent variable selected in chapter 4, section 4.4 used for normal multiple regression analysis, that were based on the data set in phase 7. A coefficient of determination of $r^2 = 0.3183$ was achieved for NN on the complete data set. A reduction of 30% in data points occurred for reaching $r^2 = 0.3183$; this is a result through the training algorithm to reach an optimal model accuracy.



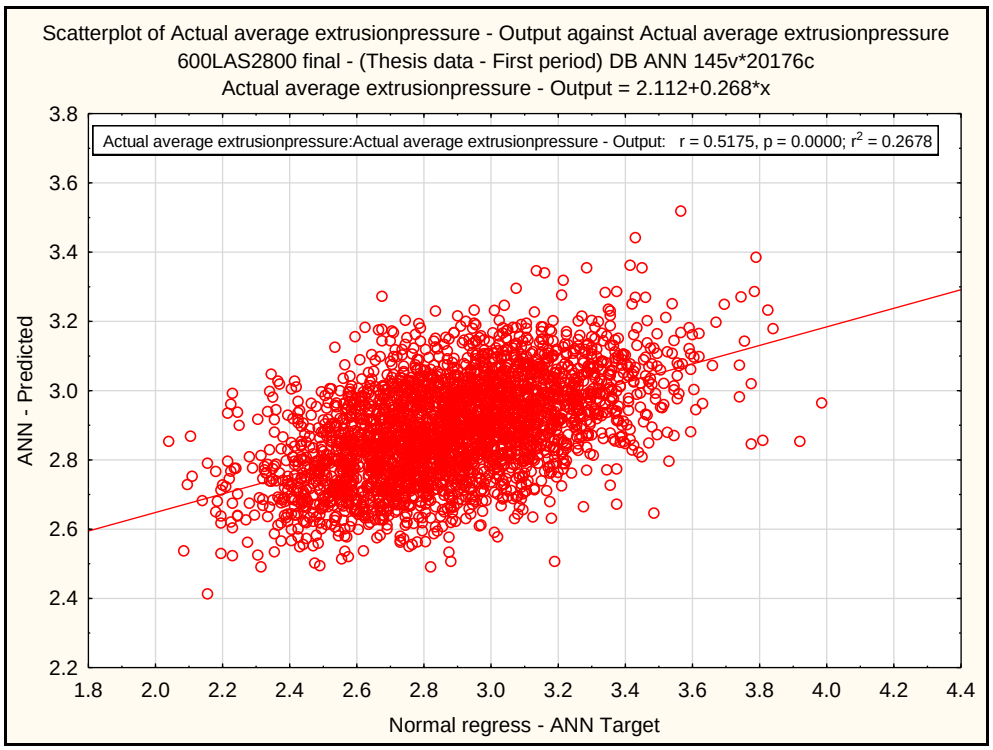
Graph 9.2: Neural network (NN) regression – Phase 2

Graph 9.2 represents the NN regression analysis for the same seven independent variables and the one dependent variable used for Graph 9.1, but is **based on the reduced data base after the coefficient of determination of $r^2 = 0.3183$ was calculated for Graph 9.1**. A coefficient of determination of $r^2 = 0.2625$ was achieved for NN based on a further 19% reduction of the data set because of the training algorithm to reach an optimal model accuracy.



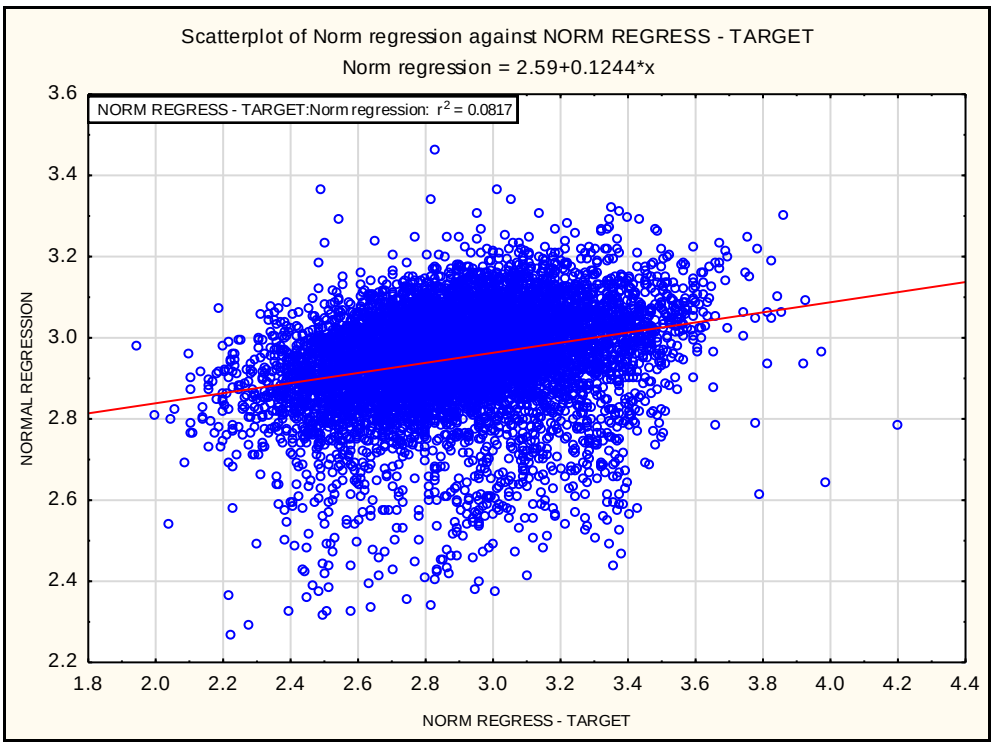
Graph 9.3: Neural network (NN) regression – Phase 3

Graph 9.3 represents the NN regression analysis for the same seven independent variables and the one dependent variable used for Graph 9.2, but is **based on the reduced data base after the coefficient of determination of $r^2 = 0.2625$ was calculated for Graph 9.2**. A coefficient of determination of $r^2 = 0.2691$ was achieved for NN based on a further 16% reduction of the data set because of the training algorithm to reach an optimal model accuracy.



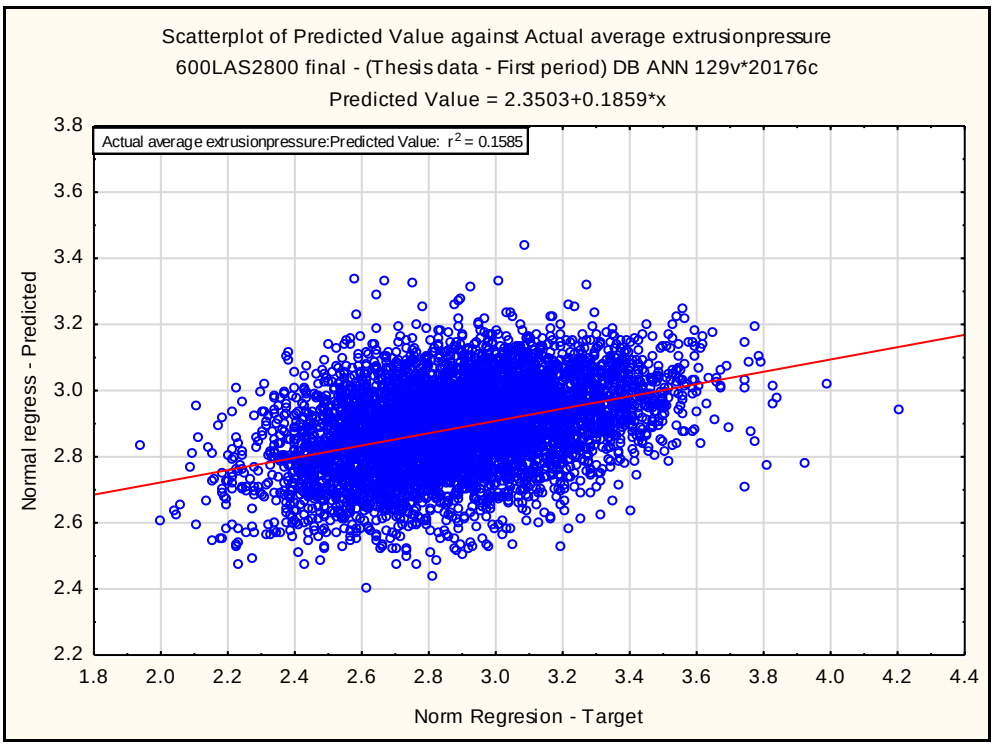
Graph 9.4: Neural network (NN) regression – Phase 4

Graph 9.4 represents the NN regression analysis for the same seven independent variables and the one dependent variable used for Graph 9.3 but is **based on the reduced data base after the coefficient of determination of r² = 0.2691 was calculated for Graph 9.3**. A coefficient of determination of r² = 0.2678 was achieved for NN based on 16% reduction of the data set because of the training algorithm to reach an optimal model accuracy.



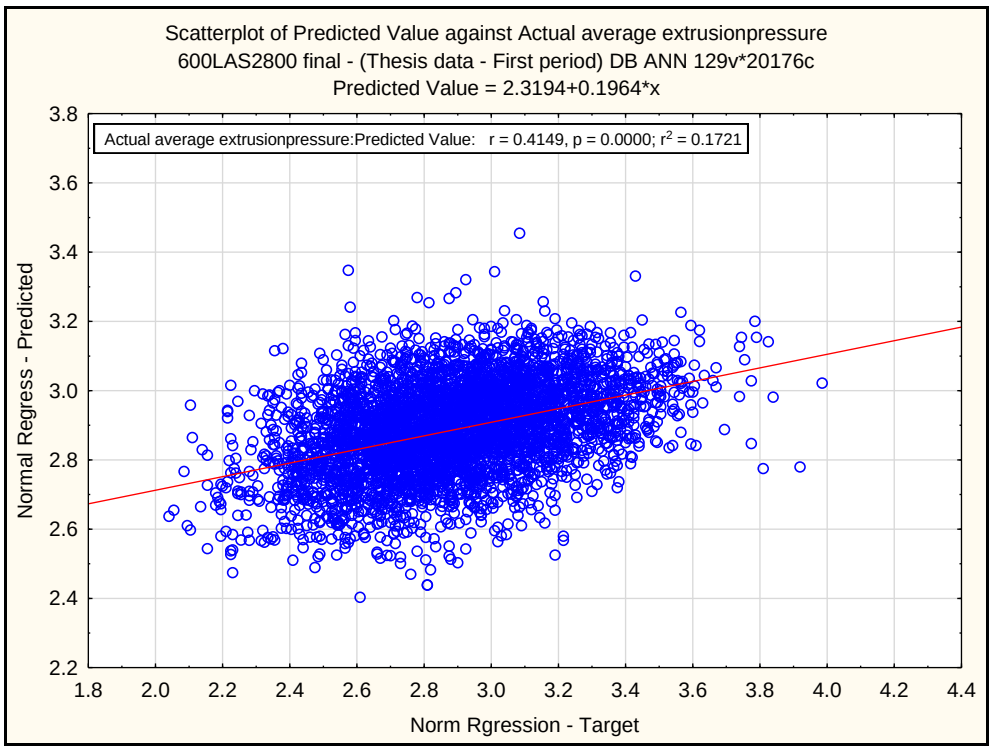
Graph 9.5: Normal multiple regression – Phase 1

Graph 9.5 represents the normal regression analysis for the seven independent variables and one dependent variable selected in chapter 4, section 4.4 on the data set in phase 7. **A coefficient of determination of r² = 0.0817 was achieved for normal regression on the complete data set.** No reduction of data points occurred for reaching r² = 0.0817. The data points in Graph 9.5 are more scattered than Graph 9.1. This is evident in the difference in the coefficient of determination of r² = 0.3183 for NN regression compared to r² = 0.0817 for normal regression respectively. From this comparison NN networks give similar data spread and slope but is more clustered around the slope with much less variation. The higher coefficient of determination for NN networks is a result of a reduced data set due to the exclusion of fliers that influence the prediction error negatively compared to multiple regression analysis that uses the complete data set.



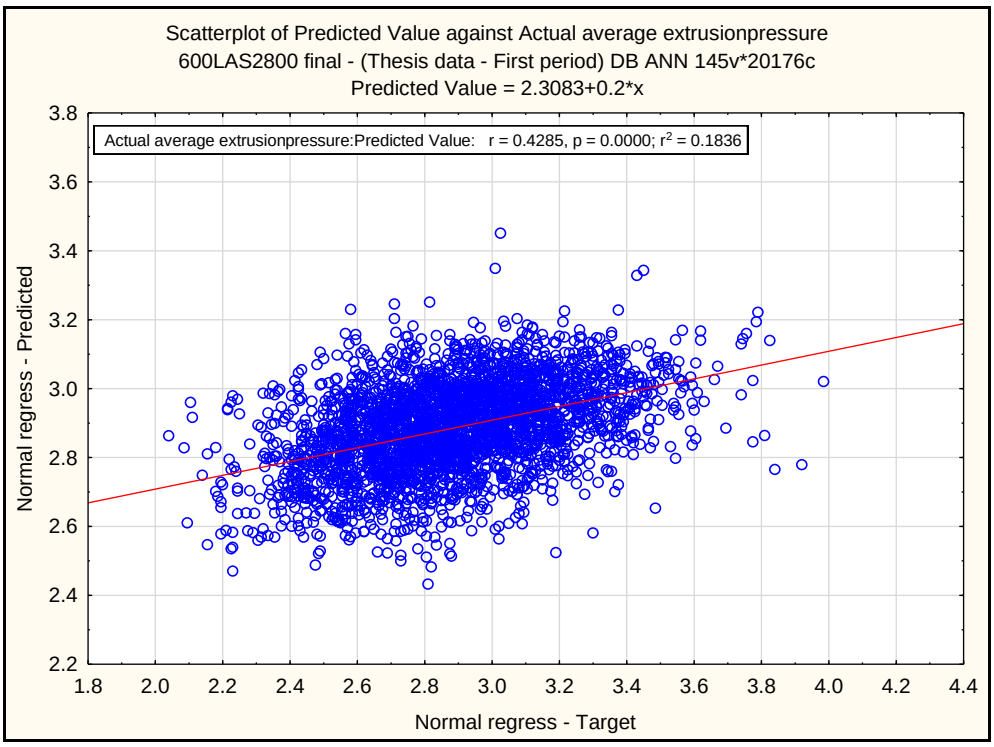
Graph 9.6: Normal multiple regression – Phase 2

Graph 9.6 represents the normal regression analysis for the seven independent variables and one dependent variable selected in chapter 4, section 4.4 on the data set in phase 7. **A coefficient of determination of r² = 0.1585 was achieved for normal regression on the 30% reduced data set used for NN regression.** The data points in Graph 9.6 are more clustered and similar than Graph 9.2. Therefore multiple regression gives similar data spread and slope, and is similarly clustered around the slope with much less variation than Graph 9.5. The higher coefficient of determination for multiple regression is a result of **using the same reduced data set when NN regression achieved r² = 0.2625.**



Graph 9.7: Normal multiple regression – Phase 3

Graph 9.7 represents the normal regression analysis for the seven independent variables and one dependent variable selected in chapter 5, section 5.4 on the data set in phase 7. A coefficient of determination of **r² = 0.1721 was achieved for normal regression on the 16% reduced data set used for NN regression.** The data points in Graph 9.7 are more clustered and similar than Graph 9.3. Therefore multiple regression gives similar data spread and slope, and is similarly clustered around the slope with much less variation than Graph 9.6. The higher coefficient of determination for multiple regression is a result of **using the same reduced data set when NN regression achieved r² = 0.2691.**



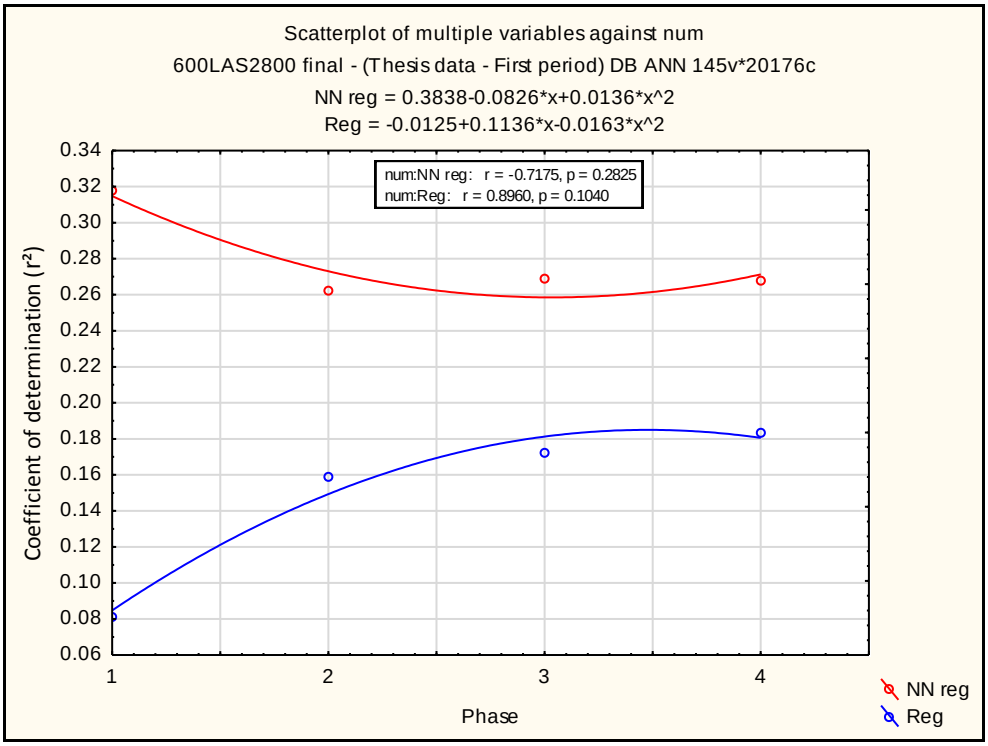
Graph 9.8: Normal multiple regression – Phase 4

Graph 9.8 represents the normal regression analysis for the seven independent variables and one dependent variable selected in chapter 4, section 4.4 on the data set in phase 7. **A coefficient of determination of r² = 0.1836 was achieved for normal regression on the 16% reduced data set used for NN regression.** The data points in Graph 9.8 are more clustered and similar than Graph 9.4. Therefore multiple regression gives similar data spread and slope, and is similarly clustered around the slope with much less variation than Graph 9.7. The higher coefficient of determination for multiple regression is a result of **using the same reduced data set when NN regression achieved r² = 0.2678.**

Phase	Neural Network (NN) r²	Regression r²	Data points used for - NN	Number used for – Regress	Reduction in data for NN	% data Reduction for NN
1	0.3183	0.0817	8933	8933	-	-
2	0.2625	0.1585	6253	6253	2680	30%
3	0.2691	0.1721	4378	4378	4555	16%
4	0.2678	0.1836	3065	3065	5868	16%

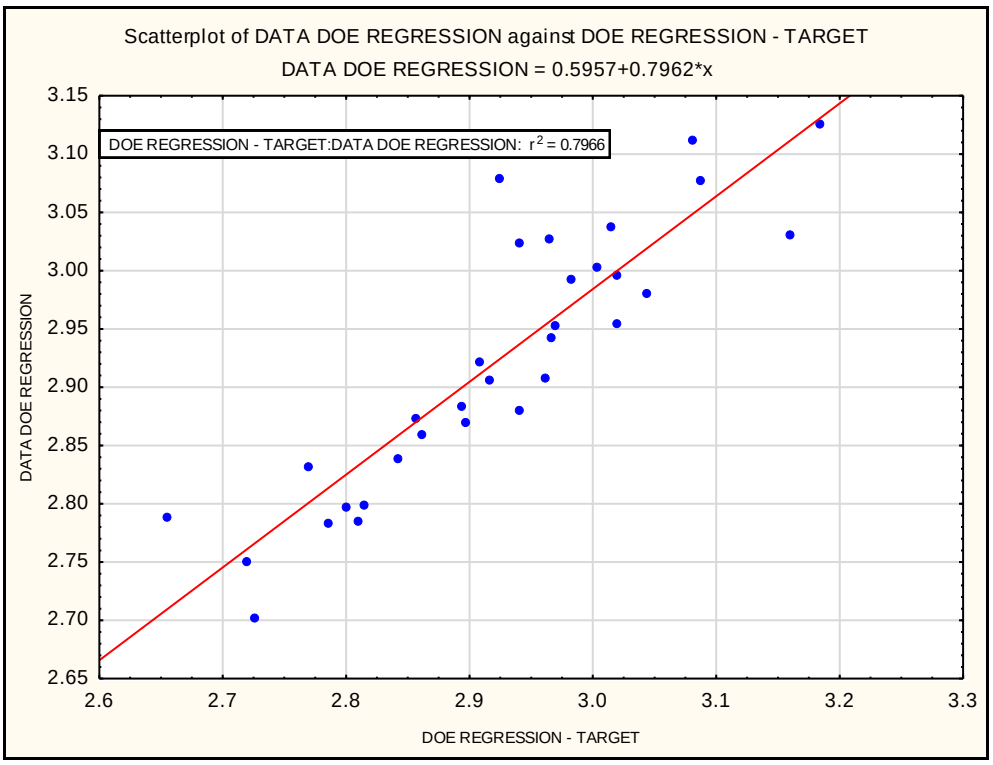
Table 9.1: Comparison summary of Neural network (NN) and multiple regression

Table 9.1 represents a summary of sequential phases comparing NN regression to multiple regression. It shows the data reduction for each phase for running NN regression based on the residual data set of the previous NN regression run. For comparison, multiple regression was run on the same reduced data set. The summary shows that the coefficient of determination improves for multiple regression with each reduced data set compared to NN regression, where NN regression coefficient of determination reduced and then stabilized even with a data set reduced by 62%.



Graph 9.9: Normal multiple regression

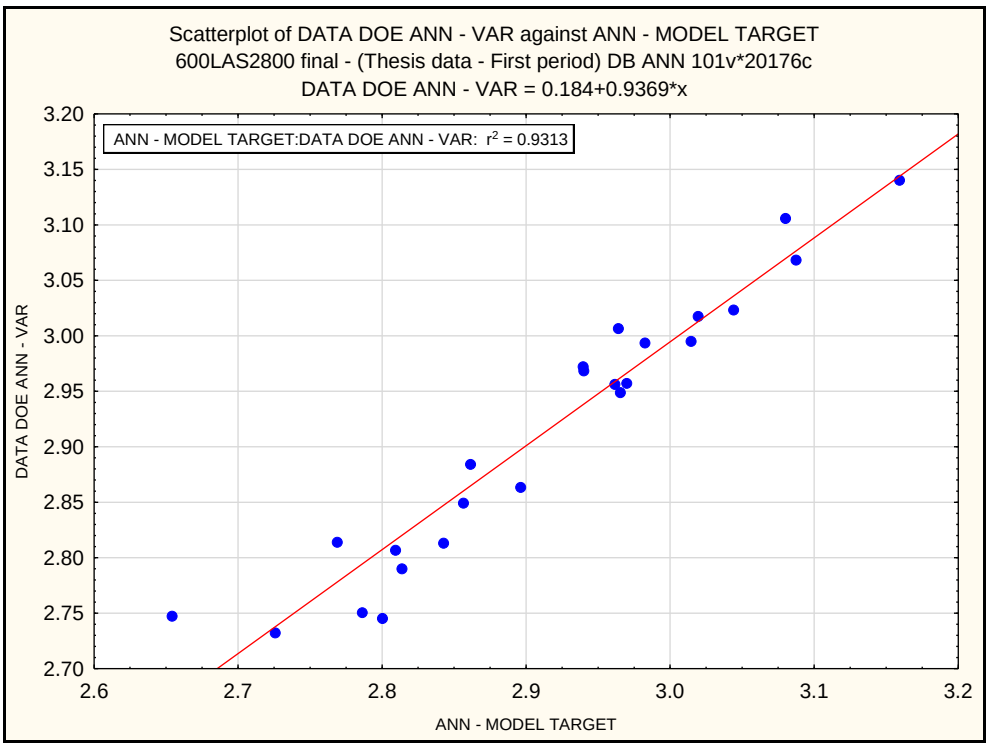
Graph 9.9 shows the change in **r²** for the comparison between NN regression and multiple regression based on the same data set size. For this application, the coefficient of determination for NN regression reduces then stabilizes compared to the coefficient of determination for multiple regression that progressively increases.



Graph 9.10: DOE regression

Graph 9.10 represents the DOE regression analysis for the seven independent variables and one dependent variable selected in chapter 6, based on the 32 DOE runs.

An $r^2 = 0.7966$ was achieved for normal DOE regression with 32 runs.



Graph 9.11: Neural network DOE regression

Graph 9.11 represents the NN regression analysis for the seven independent variables and one dependent variable selected in chapter 6, based on the 32 DOE runs.

An $r^2 = 0.9313$ was achieved for NN regression with the DOE 32 runs.

The data points in Graph 9.10 are more scattered than Graph 9.11. This is evident in the difference in the coefficient of determination $r^2 = 0.7966$ for multiple regression compared to $r^2 = 0.9313$ for NN regression respectively. From the comparison NN networks give similar data spread and slope as multiple regression but has less variation. Both coefficients of determination are relatively high, therefore for this application they are similar.

Step No.	Model Building Summary			
	1 Model Variables	2 df	3 F	4 F p-value
1	Mix discharge temp	1	0.19416(	0.66358(
	Cool begin temp	1	1.56229(	0.22389(
	Actual cool time	1	3.83443(	0.06244(
	Actual dump temp	1	3.51309(	0.07364(
	Actual tamp pressure	1	0.21011(	0.65097(
	Actual extrusion rate	1	0.38126(	0.54299(
	Actual extrusion speed	1	0.96646(	0.33579(
	Actual plug temp	1	0.00100(	0.97501(
	Actual mix time	0		
	Die top temp	0		
	Actual mud cyclinder temp	0		
	Actual mud extrution temp	0		
	Actual ram temp	0		
	Ram set temp	0		

Table 9.2: Summary: Neural network regression for 14 independent variables

Effect	Parameter Estimates (600LAS2800 final - (Thesis data (*Zeroed predictors failed tolerance check) Sigma-restricted parameterization				
	Comment (B/Z/P)	Actual average extrusion pressure Param.	Actual average extrusion pressure Std.Err	Actual average extrusion pressure t	Actual average extrusion pressure p
Intercept		1.436140	7.662160	0.18743	0.852965
Mix discharge temp		-0.016324	0.037047	-0.44064	0.663586
Cool begin temp		-0.014239	0.011392	-1.24992	0.223899
Actual cool time		0.068097	0.034776	1.95817	0.062447
Actual dump temp		0.057796	0.030835	1.87433	0.073643
Actual tamp pressure		-0.248487	0.542093	-0.45838	0.650978
Actual extrusion rate		-0.010268	0.016629	-0.61746	0.542995
Actual extrusion speed		0.035486	0.036097	0.98309	0.335793
Actual plug temp		-0.000445	0.014044	-0.03166	0.975013
Actual mix time	Zeroed*	0.000000			
Die top temp	Zeroed*	0.000000			
Actual mud cyclinder temp	Zeroed*	0.000000			
Actual mud extrution temp	Zeroed*	0.000000			
Actual ram temp	Zeroed*	0.000000			
Ram set temp	Zeroed*	0.000000			

Table 9.3: Summary: Neural network regression for 14 independent variables

Tables 9.2 and 9.3 represent the NN stepwise regression showing the significant independent variables. The results show that eight variables were significant of the fourteen selected. These fourteen independent variable selected are the same variables selected for NN regression and normal regression analysis. The six variables that were

excluded by NN in the model have zero variation, which coincides with the variable selection process in chapter 6 by using SPC.

Independent variable selection comparison between SPC and DM (Stepwise regression)		
Independent Variables	**SPC selection**	**DM (stepwise regression)**
Mix discharge temp	**yes**	**Yes**
Cool begin temp	**yes**	**Yes**
Actual cool time	**Yes**	**Yes**
Actual dump temp	**Yes**	**Yes**
Actual tamp pressure	**Yes**	**Yes**
Actual extrusion rate	**Yes**	**Yes**
Actual extrusion speed	**Yes**	**Yes**
Actual plug temp	**No**	**Yes**
Actual mix time	**No**	**No**
Die top temp	**No**	**No**
Actual mud cylinder temp	**No**	**No**
Actual mud extrusion temp	**No**	**No**
Actual ram temp	**No**	**No**
Ram set temp	**No**	**No**

Table 9.4: Independent variable selection summary comparison (SPC Vs NN)

Table 9.4 represents a summary of independent variable selection for DOE comparing SPC and NN. Only one variable was not significant for both NN and SPC. This is interesting in that SPC is a visual analytical technique and NN is a black box algorithm driven technique. From this comparison, it seems that either technique is effective but from my perspective, SPC is more appropriate because it allows the analyst to visually evaluate variables based on their spread, shape and variability over time.

9.4 CONCLUSION

For this study NN regression performs better than multiple regression based on the calculated coefficient of determination but is not a true comparison seeing that NN

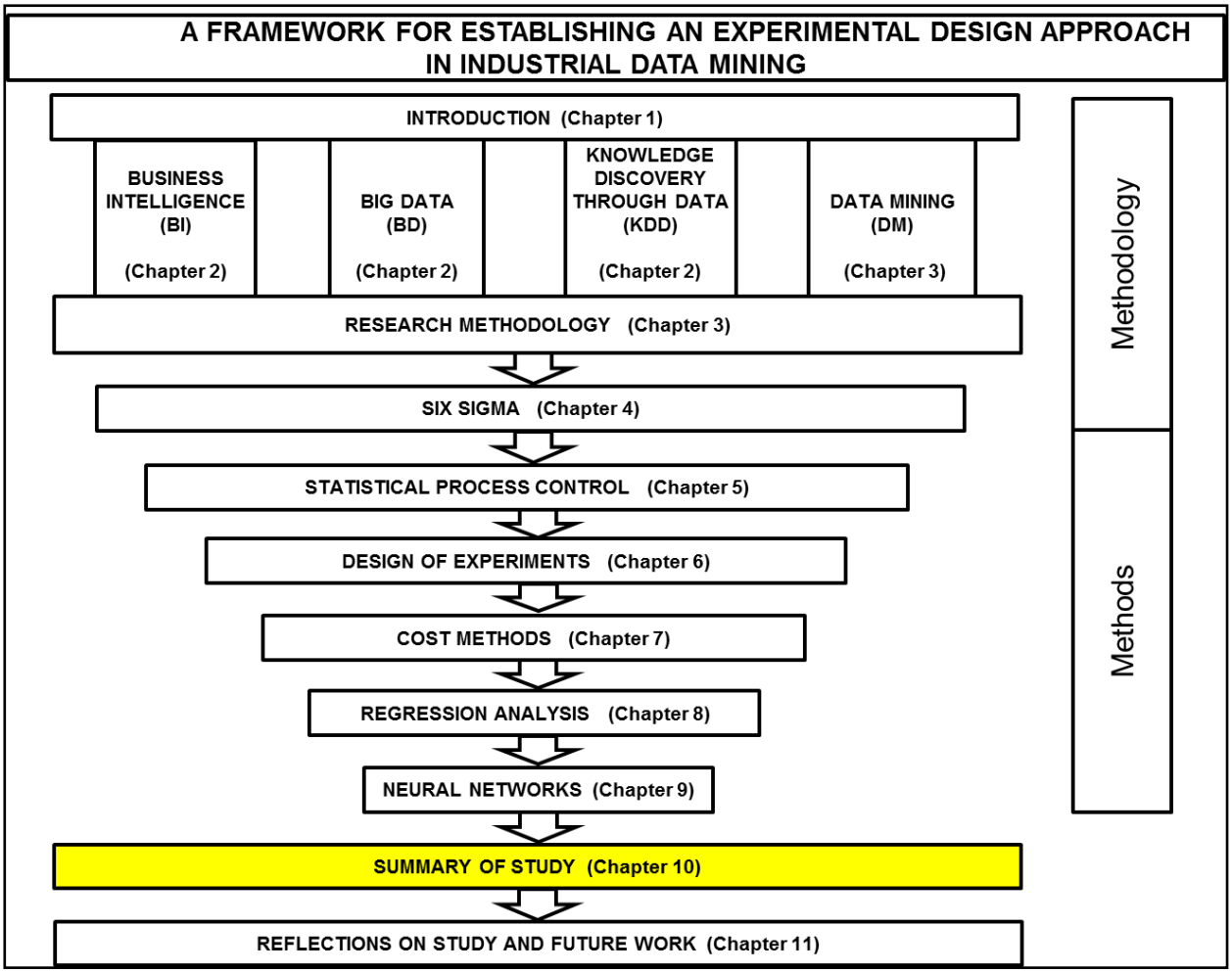
regression reduces the data set for each calculation to achieve a low prediction error by eliminating fliers.

No fliers are removed during the multiple regression calculation, all data are used. Because of the difference in coefficient of determination calculation by both techniques, NN regression will be superior, but multiple regression will catch up as the residual database stabilises, see Graph 9.9. A data loss of 60% after four phases for NN regression stabilizes the coefficient of determination but does not represent all data for residual analysis. For this reason, a comparison between these two techniques is not recommended.

Neural Networks show that eight variables of the fourteen independent variables selected were significant. These variables coincide with the variable selection process in chapter 5 by using SPC. Reaching the same variable selection shows that there is not only one way of selecting significant independent variables as a screening process.

CHAPTER 10

SUMMARY OF STUDY



10.1 SUMMARY OF GOALS

Chapter 1 reflects the proposal for the study that included five primary goals. Below are each of the five goals linking them to the appropriate chapter or chapters for a discussion how each of these goals were achieved. A detailed analytical discussion that includes the sample illustrations and the approach for each goal is available within each associated chapter.

This research met all goals with detailed discussions of how the process evolved in reaching these goals through detailed discussions and illustrations within the referenced chapters.

Goal 1: **To accommodate DOE as a Data Mining Technique in an Industrial Data Mining environment**

Refer to chapter 3 that discusses the DM methodology as a roadmap for data analysis which integrates with DOE, DMAIC, BI, KDD and BD for this study to form a framework within which the study is done. DOE is discussed as an analytical method applied to the case study in chapter 6, this supports the application of DOE within the proposed framework.The research philosophy focuses on the research approach and the research methodology focuses on Data Mining (DM) used in the study as well as the company with its related database for statistical analysis. The goal was not to exhaust all possible theories and philosophies but only those few that relate to the research framework for this study set out in the research approach.

Goal 2: **To enhance the awareness of expanding DOE as a statistical approach to complement existing methods and methodologies used for Data Mining**

Refer to chapters 4 (Six Sigma (SS)), 8 (Regression analysis (RA)) and 9 (Neural Networks (NN)) that focus on the Six Sigma (SS) methodology, Regression (RA) and Neural Networks (NN) as multivariate statistical techniques respectively. These three chapters discuss how SS, RA and NN complement DOE as a statistical analytical approach. Six Sigma follows the DMAIC methodology that includes DOE during the measuring and analysing phase; Regression Analysis is complimentary as a statistical multivariate control technique, and NN is a data mining technique for finding “nuggets” but also to complement regression analysis in finding significant variables for regression analysis. All three complement DOE to support a balanced DOE design.

For this research, only selected techniques from the DMAIC process, for fitting, cleaning of data, designing of experiments, predicting process behaviour and model building are used. The technique selection process was done to show the power of using basic statistical techniques in a complex analytical environment. The analytical process however shows a progression started with univariate analysis (Histograms) then progressed to bi-variate analysis (Statistical Process Control) then to Multivariate analysis (Multi variable Regression, DOE, Regression).

Goal 3: **To validate the integrity of captured data through the refining process to determine upper and lower operating conditions required by DOE, any abnormal data points will be exposed**

Refer to chapter 5 that introduce Statistical Process Control (SPC) as a data refining process through as a variable reduction technique through validating data integrity, process variability and variable process capability comparing process specifications to process variability. The upper and lower operating conditions required for DOE is enhanced by the application of SPC prior to determining these operating levels. Chapter 6 discuss the use of the core data by DOE after the refining process by SPC.

From a SPC perspective, none of the independent or dependent variables are in statistical control showing many cyclical patterns, shifts, trends, high and low fliers. Even though these patterns exist, specification limits for each of these variables are wider than the three-sigma variation limits used for SPC. For this reason the opportunity for process improvement does exist for all variables, with independent variable 5 showing the most room for improvement. From a process capability perspective, all capability indices were greater than 1, but only two variables show a Cpk > 2 which provides the most process improvement opportunities by reducing specifications **or** adjusting the process mean towards a higher or lower operating level.

In addition to SPC and capability analysis, the normality assumption that sample averages and sample standard deviations should be approximately normally distributed was used, and scatter plots were constructed for each independent variable showing the relationship between sample average and sample standard deviation. The expectation was that no trends should be evident, only random scattered data points. Most variables have noticeable trends with only two randomly distributed. Only one variable shows a normal like distribution for both sample average and sample standard deviation distributions, a high CPk and close to random scattered data points for scatter plots representing sample average compared to sample standard deviation. This variable turned out to be the “red X” in the regression analysis.

Goal 4: To focus on Industrial Data Mining, and concentrate on process data, applying DOE rather than generic, traditional Data Mining techniques

Refer to chapter 6 that shows how experimental design analysis was applied in developing a DOE model to identify significant variables to process improvement. Using a validation process, of the seven independent variables, four support the DOE proposed model for process improvement and three are either not significant or need

more evaluation for understanding. A detailed discussion follows in chapter 6 on experimental design analysis, which complements goal 4.

Three models were tested,

- 2** (7-0) resolution FULL (128 RUNS) DOE factorial design
- 2** (7-1) resolution VII (64 RUNS) DOE factorial design
- 2** (7-2) resolution IV (32 RUNS) DOE factorial design.

Comparing significant main effects for the three models, the 2** (7-2) resolution IV (32 RUNS) DOE factorial design were chosen because this design has no missing values and therefore the significant independent variables may be more credible and reduce the risk of masked factor effects due to missing values.

Goal 5: To develop a methodology to accommodate the use of DOE as a Data Mining technique to determine impacts of variables on process outcomes through experimenting with data within current databases

Refer to chapter 7, showing the cost impact of DOE runs in finding an optimum region to operate within by reducing the risk of producing non-conforming products.

Cost impact analysis is an important factor when analysing data bases. To find an optimal process solution provided by the cost impact analysis is not enough. The associated costs to a proposed solution provide a benchmark for financial evaluations and can also serve as a control element in the analytical phases of DMAIC.

10.2 POINTS TO REMEMBER FOR THE CASE STUDY

When reflecting back to the analytical process followed during the study, a few generic points are summarized in form of a flow chart to keep in mind if an analyst or management need to embark on a study of a similar nature. These points are expressed in sequential “steps” because they follow a generic roadmap of the study for this research. This roadmap was specific in keeping experimental design as the main theme which also complements the title of this research.

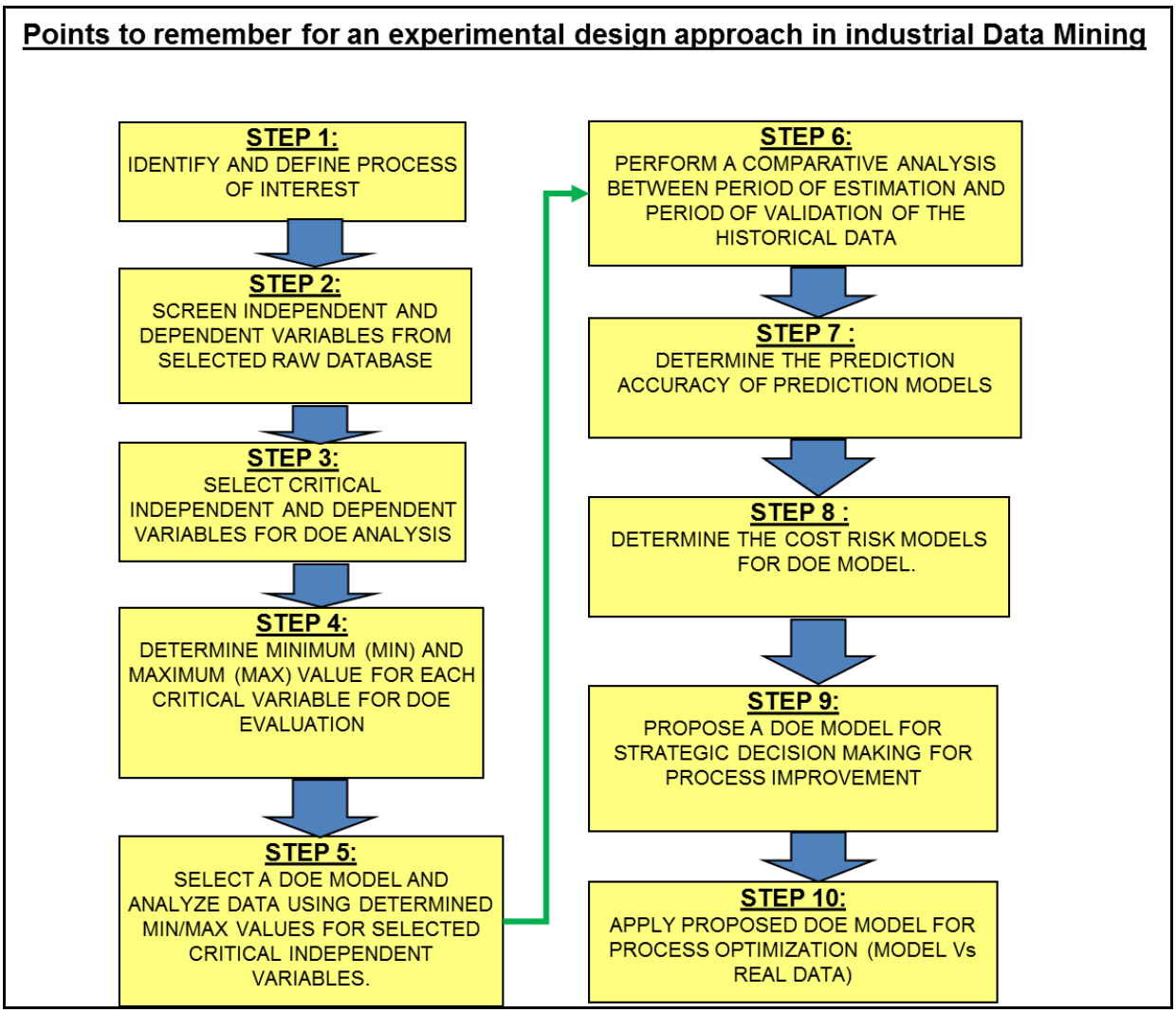


Figure 10.1: Points to remember for an industrial data mining analysis

Referencing each of the ten steps in Figure 10.1 to appropriate chapters and sections in this study, gives substance to the integration of these points amongst all methods and techniques in this study.

10.2.1 Step 1: Identify and define process of interest

Refer to chapter 3 (Research Approach) section 3.2 and chapter 4 (Six Sigma) section 4.3. These refer to the research methodology and delimitation of research and the selected database respectively. Here the company agrees to the specific process, variables and database for analysis purposes.

In the first step, identifying the process or system to be analysed is the main goal. There are many approaches to the endeavour of selecting a process or system to analyse; we will follow three basic approaches selecting a project for a specific process:

Blatantly obvious. Things that clearly occur on a repetitive basis and present problems in delivering service(s) or product(s).

Brainstorming. Identifies projects based on individuals' experience and tribal knowledge of areas that may be creating problems in delivering service(s)/product(s), and hopefully, ties these to bottom-line business impact

Structured approach. Identifies projects based on organisational data; provides a direct plan to affect core business metrics that have bottom-line impact.

In addition to the three basic approaches used in selecting a project or process, **triggers** also influence, identifying a process to be analysed.

Some of these **triggers** are:

- A process that produces defective products;
- Feedback from a customer to inform your company that substandard products are processed;
- A new market that requires tighter specifications than what are currently produced;
- Streamlining current product range;
- Reduction in process variability;
- Random excessive variability that renders processes not to be statistically capable;
- A deeper understanding of the process of how different variables influence the final product;
- The need for components of variation that impact on the financial value chain becomes critical;
- Reduction of process cycle times.

By the end of **Step 1,** the process or system is identified for further analysis as agreed by management.

10.2.2 Step 2: Screen independent and dependent variables form selected raw database

Refer to chapter 4 (Six Sigma (SS)). Section 4.4.2 describes the seven phases used to reduce the original 44 variables to 17 variables (dependent and independent). See Table 4.1, showing these 17 variables corresponding to phase 5. These are not the critical variables, but the phase before evaluating these 17 variables to select the final critical variables.

After the process identification phase, the challenge is to only work with the critical few variables that drives the process. The goal during this stage is not determine the critical variables but to arrive at a first tier database with variables that represent the most accepted variables after going through an initial screening process.

During the initial high-level screening process, a general guideline to follow in determining which variables will remain and which will be removed is determined by three basic types of input variables, namely **procedural inputs**, **controllable inputs** and **noise inputs.** To distinguish between these types of variables is not easy, but will assist efforts to identify those variables that are key to the process. General guidelines to distinguish between the different variables are:

Controllable variables. Input variables that can be adjusted or controlled **while** the process is running; for this study, actual extrusion speed, actual extrusion rate, actual dump temperature. These variables usually are continuous by nature and form the core operating basis of any process.

Procedural variables. A specific procedure is followed prior to and during adjusting or controlling an input variable. These variables have proven to have interaction effects that could cause major defects and/or disasters. They are discharge temperature, cool begin temperature, and actual dump temperature.

Noise variables. Process conditions we do not think we can control, we are unaware of or do not see, or are too expensive or too difficult to control like ambient temperature and humidity. Pragmatically these variables are controllable, but at a cost.

In addition to the selection process, a combination of experience, process knowledge and scientific measurements could also assist with the final selection. During this initial

screening stage, only variables (dependent and Independent) that have survived the initial screening phase for process improvement should be identified for further analysis.

10.2.3 Step 3: Select critical independent and dependent variables for DOE analysis

Refer to chapter 5 (Statistical Process Control (SPC)). Sections 5.3 and 5.4 describe the critical variable selection process using SPC and capability analysis. Subsequent to this analysis, seven critical independent variables and one dependent variable were selected for DOE analysis.

After the most accepted variables are selected, a further screening process is needed to find the critical variables(Xs) for the function Y=F(x). It is imperative to find the critical (core variables that directly affect a process, always part of the model) or significant (only statistically significant variables affecting outputs) variables that impact a process, but they also very difficult to find. This stage must be early in the Experimental design approach sequence because of its critical importance for effective data analysis going forward. Omitting, neglecting or poorly conducting **Step 3,** may cause the rest of the building blocks for process analysis to lead to poor decisions.

In my experience, too often organisations build complex data collection and information management systems without truly understanding how the data collected and metrics calculated actually benefit the organisation and the users of data on a daily basis.

Data are supposed to be accurate, reliable and organised, to such an extent that proper data analysis can be done. Data integrity starts at data feed points of the various processes, and a clear distinction has to be made between process data that are electronically accumulated, scanned data and manual inputs by business personnel. Each of these three data collection points has some degree of accuracy error, which could lead to inaccurate results.

Some statistical techniques that are useful during the data integrity evaluation phase are: 2D and 3D scatter plots, Regression analysis, Histogram, Distribution fitting, Box plots, Process capability charts, and statistical process control charts.

Each of the above statistical techniques, when used properly, will show abnormalities in data. The challenge is which technique or techniques to use. Fortunately, there is not a

prescribed way to select which method to use, but with modern statistical software, it is easy and quick to run the applicable data through different statistical techniques. Graphical data representation will show your data graphically so that you may ask questions about it.

A common **mistake** when analysing data for integrity is **not** to involve process experts during this phase. Involving process experts gives you information on how a process is operating in reality by accommodating all internal and external process influences. This will assist in **understanding the process**, not only through collected data. A failsafe test for process understanding is trying to operate a process when the usual operators are on strike. Trained and experienced persons will assist in distinguishing between reality and facts.

Patterns that do appear, that may seem not to be a part of the process, and are discarded due to non-value data, could be real and part of the natural variation of the specific process variable. Figure 10.2 illustrates the perceived differences between the process of collecting data that have a direct influence on the integrity of data amongst the analyst (1), the process expert (2), and management (3).

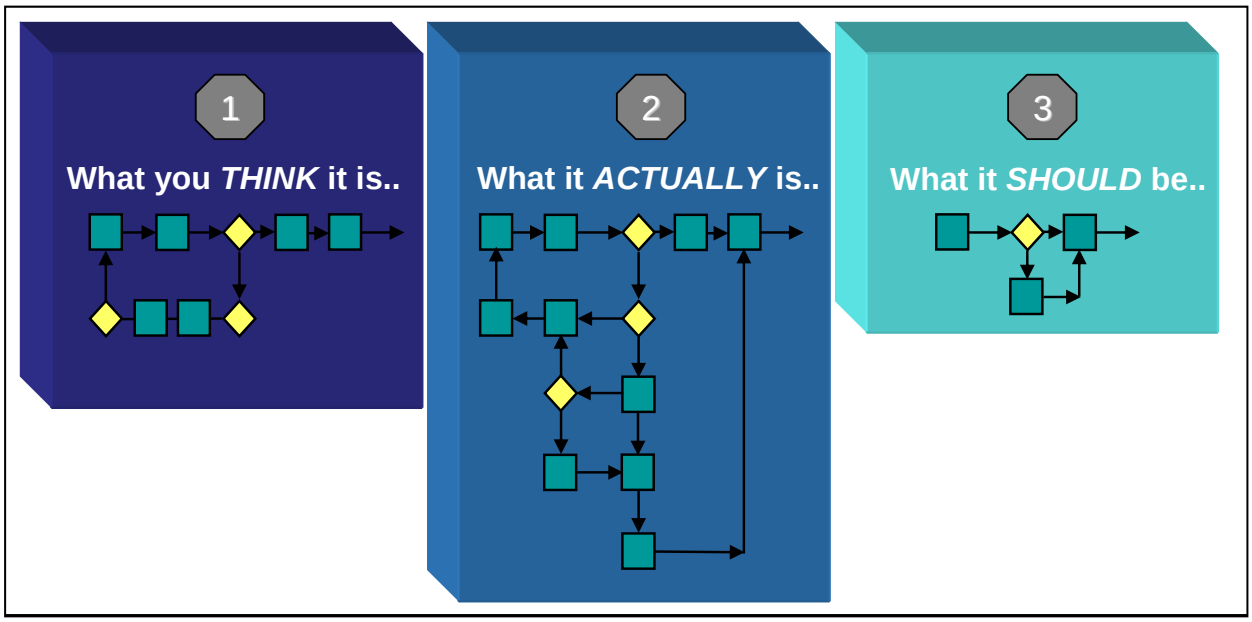


Figure 10.2: Differences that have a direct influence on data integrity

To distinguish factually between these realities in Figure 10.2 takes careful planning, understanding and contribution of each process variable to a measured output. All three perceptions of data are always present for the analyst during the data integrity

evaluation phase. Based on my experience, analysts focus on what management wants and not necessarily on what the data show, because the time constraint for solutions is a major strategic management driver. For this reason, perceptions of data play a major role during the analysis phase.

The focus of **Step 3** is to determine which critical dependent and independent variables contribute real added value to the process. The variables should be a combination between what the data actually show us, and what they should be excluding all the system noise.

10.2.4 Step 4: determine minimum (min) and maximum (max) value for each critical variable for DOE evaluation

Refer to chapter 6 (Design of Experiments (DOE)). Section 6.2 discusses how the MIN (-1) and MAX (+1) values and coding for the critical variables were approached and finally agreed upon.

A recommended process to follow in allocating MIN (-1) and MAX (+1) is as follows:

Recode data for each critical independent variable to represent the extreme (-1) and (+1) experimentation levels.

Divide the data for each independent variable in half (50/50) using the median statistic. Each half represents the min and max values, the lower half represents (-1) and the upper half represents (+1). The associated numerical value for each min and max represents the data for DOE experimentation.

Test if all permutations (-1) and (+1) for all experimental runs exist − No missing values.

During this analysis, the goal is to have at least one full factorial design. The type of DOE design resides with the data analyst. If this goal cannot be achieved due to limited database size then a fold over design may be considered with the highest resolution. Once this is achieved, the next step is to optimise the number of replicates within the database for each independent variable. There is a relationship between database size and the likelihood of achieving a full factorial design that represents all possible combinations of (+1) and (-1) for critical variables and the region that is operated in. If

there are combinations not operated in, these may indicate combinations to experiment in.

10.2.5 Step 5: Select a DOE model and analyse data using determined min/max values for selected critical independent variables

Refer to chapter 6 (Design of Experiments (DOE)). Section 6.3 discusses different DOE models for selection.

Select an appropriate DOE model for evaluating process. The goal is to have a full factorial design that provides all possible combinations to determine optimal and missing conditions from the experience region. The type of DOE design resides with the data analyst. If this goal cannot be achieved due to limited database size, then a foldover design may be considered. Once this is achieved, the next step is to optimise the number of replicates within the database for each independent variable. As the database increases the extreme values increase and therefore the likelihood of full factorial designs increases.

Being a simulation, and the model being database size dependent, a full factorial design is the first choice and should only be reduced if the available data are not sufficient in size when split on the median statistic. The larger the database, the lower the risk of having no data for a selected experimental run, because each time a complete experimental run level is selected, the database is halved in size. **However, experimental runs with no data may be opportunities for improvements or experimentation.**

AT **Stage 5** only the DOE model selection and DOE experimental analysis are done to determine optimal conditions on the selected database for the estimated period.

10.2.6 Step 6: perform a comparative analysis between period of estimated and period of validation of the historical data

Comparative analysis for study is based on the cost impact for different DOE runs compared to the optimal operating region. The terms target value (process outcome target for controlling) and target variable (dependent variable representing target value) need to be described for this step. From quality perspective, specific to this process, the goal is to control the process to operate as close to the target value as possible

because any deviation from the target value, up or down, has an associated risk of non-conforming products with additional cost for the process.

Chapter 7, section 7.4 discusses the DOE outcome comparison between the two periods. Table 10.1 (repeat of Table 7.1) below summarizes the DOE run target value comparison between the two periods. Four different colours represent different operating zones associated with the DOE ranked run target outcomes. Placing associated DOE outcomes within a colour zone informs the analyst how close the DOE predicted run is to the ideal target for both periods. The zones and target values are as follows: *Best = green* (3.25 – 3.75), *Good = yellow* (3.0 – 3.25, 3.75 – 4.0), *Fair = orange* (2.75 – 3.0, 4.0 – 4.25), *Poor = red* (2.5 – 2.75, 4.25 – 4.5).

Dependent variable - Average pressure										
	First period				Second period				Second period - Predicition	
	1 st Period DOE Base		1 st Period DOE regression	1 st Period regression	2 nd Period DOE Base		2 nd Period DOE regression	2 nd Period regression	2nd period prediciton DOE regression	2nd period prediciton normal regression
DOE run number	Outcome	S/N	Outcome	Outcome	Test outcome	S/N	Test outcome	Test outcome	Test outcome	Test outcome
1*	2.962	19.506	2.907	2.962	3.327	21.160	3.256	3.169	2.907	2.878
2*	2.924	20.481	3.079	3.003	3.487	17.857	3.504	3.406	3.079	3.000
3*	3.160	24.902	3.030	3.081	3.661	18.372	3.704	3.533	3.030	**3.011**
4*	2.908	23.780	2.921	2.926	3.316	19.632	3.357	3.195	2.921	**2.863**
5*	3.044	20.796	2.981	3.011	3.332	17.517	3.339	3.271	2.981	2.985
6*	2.983	20.316	2.992	2.958	3.398	23.434	3.327	3.254	2.992	2.919
7*	2.965	21.712	2.943	2.963	3.648	18.932	3.527	3.451	2.943	2.913
8*	3.020	23.627	2.995	3.026	3.518	18.904	3.440	3.451	2.995	2.997
9*	3.080	20.739	3.111	3.069	3.498	19.246	3.560	3.426	3.111	**3.040**
10	3.004	20.933	3.003	2.990	3.172	19.811	3.213	3.144	3.003	2.929
11*	2.970	22.527	2.953	2.972	3.351	18.823	3.413	3.227	2.953	2.901
12*	3.184	20.341	3.125	3.056	3.804	17.541	3.662	3.549	3.125	**3.062**
13*	2.940	18.801	3.024	3.015	3.274	18.773	3.383	3.159	3.024	2.981
14*	3.088	21.086	3.077	3.084	3.195	17.300	3.297	3.301	3.077	3.056
15*	2.964	20.662	3.027	3.036	3.510	18.122	3.497	3.483	3.027	**3.046**
16*	3.015	22.676	3.038	3.012	3.474	21.736	3.485	3.400	3.038	2.961
17	2.810	18.426	2.786	2.779	3.034	19.012	3.098	3.152	2.786	**2.698**
18	2.843	21.628	2.838	2.794	3.021	20.797	3.012	3.128	2.838	**2.760**
19	2.654	22.388	2.789	2.773	3.196	18.027	3.212	3.257	2.789	**2.785**
20	2.814	26.946	2.800	2.728	3.165	18.970	3.200	3.183	2.800	2.647
21	2.862	19.716	2.860	2.886	3.317	19.692	3.182	3.149	2.860	2.629
22	2.720	20.646	2.751	2.763	2.806	19.235	2.835	2.950	2.751	2.624
23	2.726	21.426	2.702	2.775	2.937	19.298	3.035	3.119	2.702	**2.661**
24	2.857	21.806	2.874	2.849	3.148	20.040	3.283	3.325	2.874	2.679
25	2.896	23.970	2.870	2.894	3.133	20.459	3.068	3.113	2.870	2.767
26	2.940	23.384	2.881	2.852	3.077	17.516	3.056	3.114	2.881	2.677
27	2.769	26.197	2.832	2.814	3.317	18.965	3.256	3.213	2.832	**2.677**
28*	2.893	22.841	2.884	2.856	3.183	19.483	3.170	3.251	2.884	2.793
29	2.786	21.132	2.783	2.783	2.955	18.550	2.891	3.014	2.783	**2.742**
30	3.019	18.981	2.955	2.922	3.134	17.515	3.140	3.233	2.955	2.854
31	2.917	21.980	2.906	2.865	3.273	19.413	3.340	3.365	2.906	2.735
32	2.802	22.699	2.797	2.801	3.074	18.352	2.992	3.155	2.797	**2.728**

Table 10.1: DOE ranking Average pressure

Divide the selected raw database into two periods, first and second (test) period. Group both periods into the selected DOE model experimental runs respectively. Calculate an average value for each DOE run (condition) per dependent variable, which will be the base for all data analysis going forward. Use the respective data for both periods to formulate DOE and normal regression equations to evaluate the predicted data for each period respectively.

Evaluate each period's database independently to compare the significant independent variables for both periods. This comparison shows the process stability irrespective of the processing operating level, and if the same independent variables irrespective of the period have the same significant effects on the process output for the dependent variable.

10.2.7 Step 7: Determine the prediction accuracy of control models

Refer to chapter 8 (Regression analysis (RA)). Section 8.2 describes the methodology of calculating the prediction error for the different models and section 8.2.3 provides a prediction error summary for the different models.

This is not a pre-requisite but was necessary for this study. Calculate the predicted DOE target level accuracy using the second period of the database as a comparative base to evaluate DOE and normal regression predictions calculated form the first period data. Step 7 provides a guideline in terms of prediction error to how accurately your selected models perform as well as for optimal settings.

10.2.8 Step 8: Determine the cost risk models for DOE model

Refer to chapter 7 (Cost methods). Section 7.3 describes the cost and analysis for the DOE and regression models for different periods.

Calculate the associated cost and signal to noise ratio by moving away from the process target value for each experimental run or process conditions. This step shows the financial impact of the DOE model for each experimental run as the results deviate from the process target value within the operating region.

10.2.9 Step 9: Propose a DOE model for strategic decision making for process improvement

Refer to chapter 6 (Design of Experiments (DOE)). Sections 6.4 and 6.5 discuss the proposed DOE model.

Do an analysis for the predictive accuracy of the model versus real data of a different period (second period) based on historical data. The assumption is that if the test model accurately predicts the outcomes of the second period of historical data, then this model can be used for process development in an unknown operational environment. This assumption is only valid if the historical data on which the test model was developed represent the second historic period (the “future”).

For large databases, full factorials may be possible for the initial design, but are not necessary for process development. The amount of variables, runs, levels and economic viability will dictate if a full factorial or a fold over design is used. This step formulates the test model to predict outputs for the second period that is based on the model formulated with the data for the first period of the historical data.

10.2.10 Step 10: Apply proposed DOE model for process optimization (model vs real data)

Once the analysts and management agree that the test model and optimal conditions represent the second historical period accurately, then the model is ready for process development based on current operational data. During this phase, process adjustments should be made based on the proposed DOE model in order to optimise the process. Step 10 is where the traditional application of DOE models for process development may start.

10.3 HOLISTIC ANALYTICAL SUMMARY OF STUDY

A holistic summary of the study combining MR, DOE and capability analysis of individual variables is necessary to determine how representative the proposed DOE model is. The validating process takes into consideration comparing the coefficient slopes for DOE and MR, the significance of the independent variable for both models, and a comparison between experimental runs that represent the lowest and highest cost to the proposed model. Also, evaluate the opportunity to shift operating levels of

variables based on the proposed DOE model within the allowable experimentation region, referenced by the process capability.

Row	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Description
1	+	-	-	+	-	-	-	Proposed DOE model coefficients
2	+	-	+	+	-	-	-	BOX graph slope moving low (-) to high (+)
3	+	-	+	+	-	-	-	Individual regression coefficients
4	+	-	-	+	-	-	-	DOE run combinations for min Cost
5	-	+	+	-	+	+	+	DOE run combinations for max Cost
6	X	X	X	✓	✓	✓	✓	Significant DOE variables
7	X	✓	✓	✓	✓	✓	X	Significant MR variables
8	X	X	✓	✓	✓	✓	✓	Significant single regression variables

Table 10.2: SPC, MR and DOE result summary

Explanations for Table 10.2 are:

Row 1: Operating levels by the proposed DOE model for each selected variable.

Row 2: Represents the slope for box plots between when moving from low (-) to high (+).

Row 3: Shows the coefficients for each variable represented by MR model. These coefficients represent the slope of each variable.

Row 4: Experimental (DOE) runs representing the lowest cost combination amongst the 32 runs.

Row 5: Experimental (DOE) runs representing the highest cost combination amongst the 32 runs.

Row 6: Represents which variable is significant within the DOE model.

Row 7: Represents which variable is significant within the MR model.

Row 8: Represents which variable is significant in a single regression model.

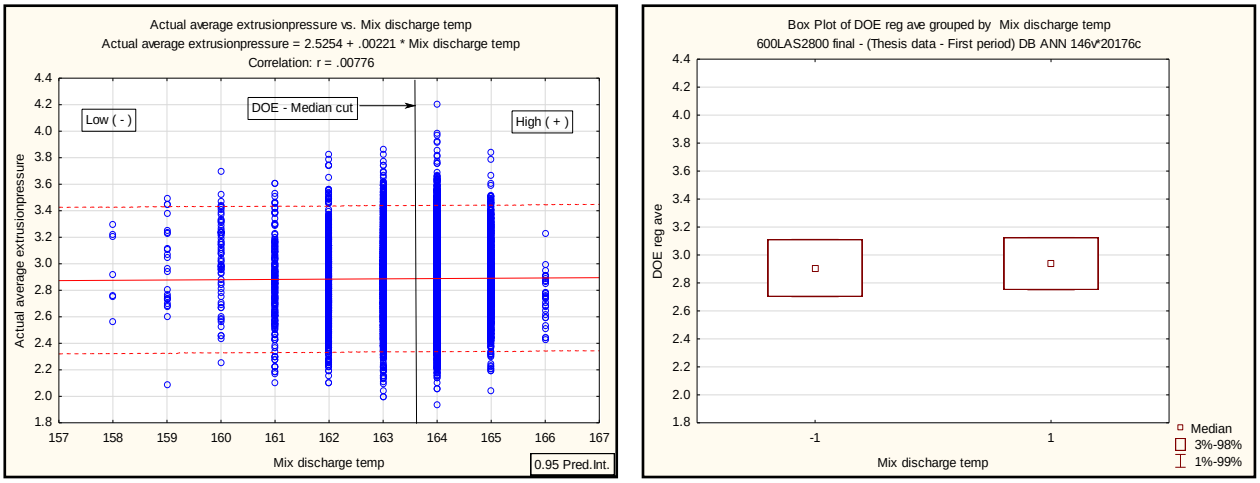
This summary shows the importance of combining analytical methods to reach balanced and factual conclusions. For this study, the identifying of the “red X” was substantial.

The criteria for validating each variable compared to the proposed DOE experimental combination by referring to Table 10.2, are:

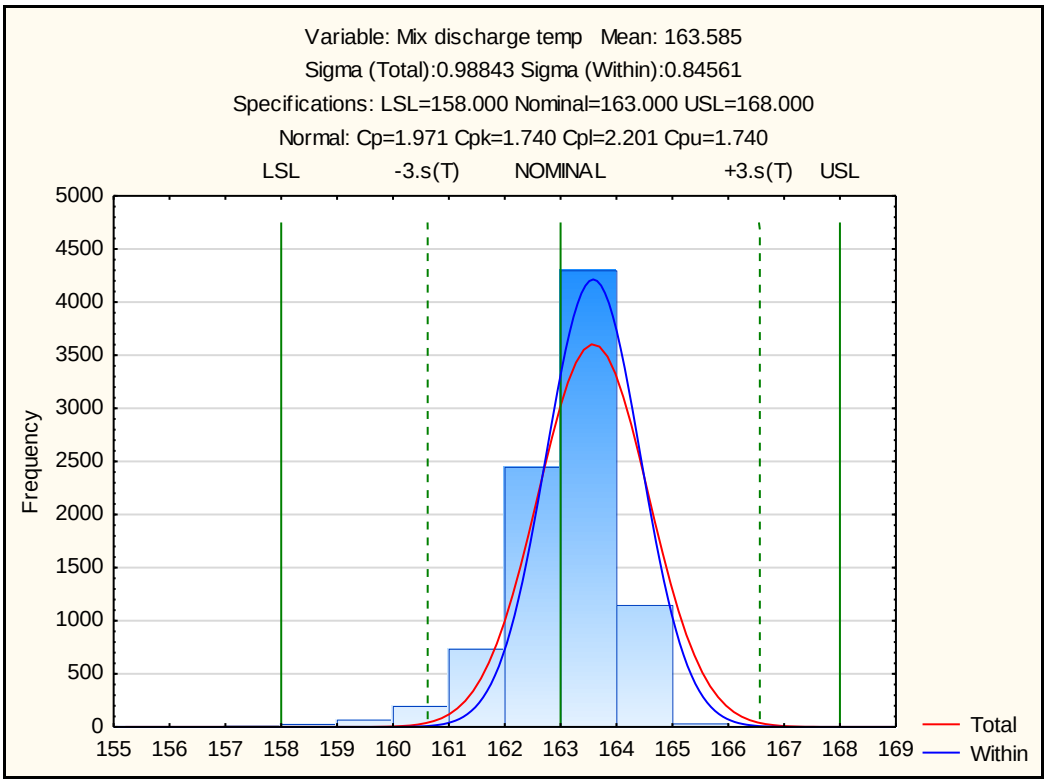
- an individual variable, which excludes possible collinearity, should be significant for both DOE and MR models; refer to Table 10.2, rows 6 and 7
- both DOE and MR should follow the same slope (coefficient sign) for both models; refer to Table 10.2, rows 2 and 3
- have comparative operating slope than the DOE proposed model; refer to Table 10.2, row 1, when moving from a low (-) to high (+) operating level represented by box plots
- an opportunity for individual variable process adjustment for process improvement based on capability analysis should exist, refer to graphs below, compared to proposed DOE model, refer to Table 10.2, row1
- proposed minimum cost model, refer to Table 10.2, row 4 should be in line with DOE proposed mode, refer to Table 10.4, row 1.

For each independent variable, the graphs below show the median split for screened data used for DOE analysis compared to the median regression value grouped by operating level. This comparison illustrates how the DOE operating level for an individual independent variable fits the screened database by grouping individual values into low or high operating levels. The added capability chart shows how each variable performed against set process specifications as well as how much room for process adjustment is available without exceeding process specifications.

10.3.1 ANALYSIS x_1



Graphs 10.1: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Mix discharge temperature (x_1)



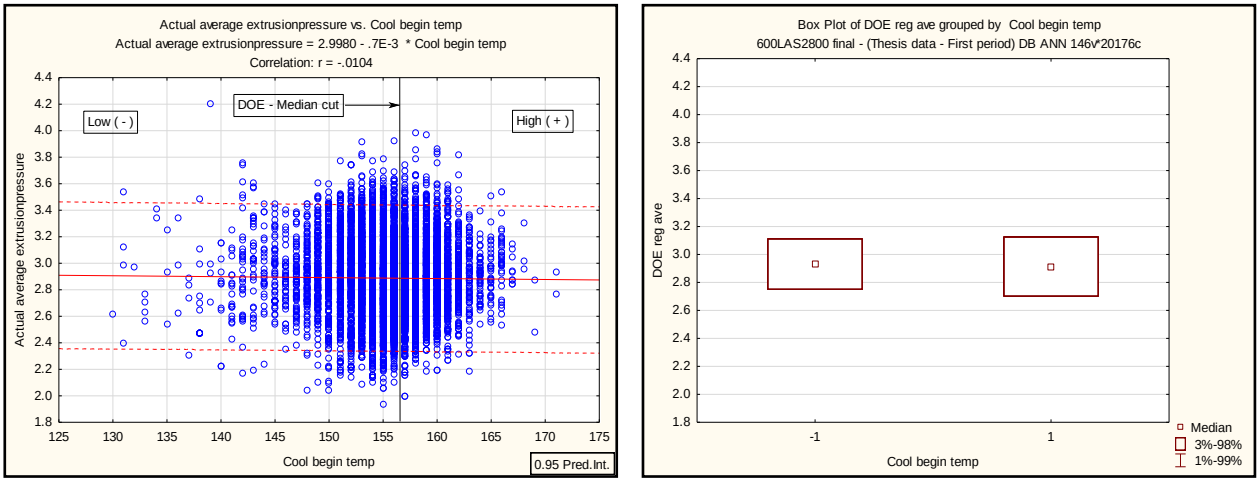
Graph 10.2: Mix discharge temperature (x_1): Process capability chart

Variable 1 is not significant for both DOE, MR and individual regression models; refer to Table 10.2, rows 6,7 and 8. Slopes, are positive for both MR and box plot, refer to Graphs 10.1. Process capability is greater than 1, refer to Graph 10.2, which provide an

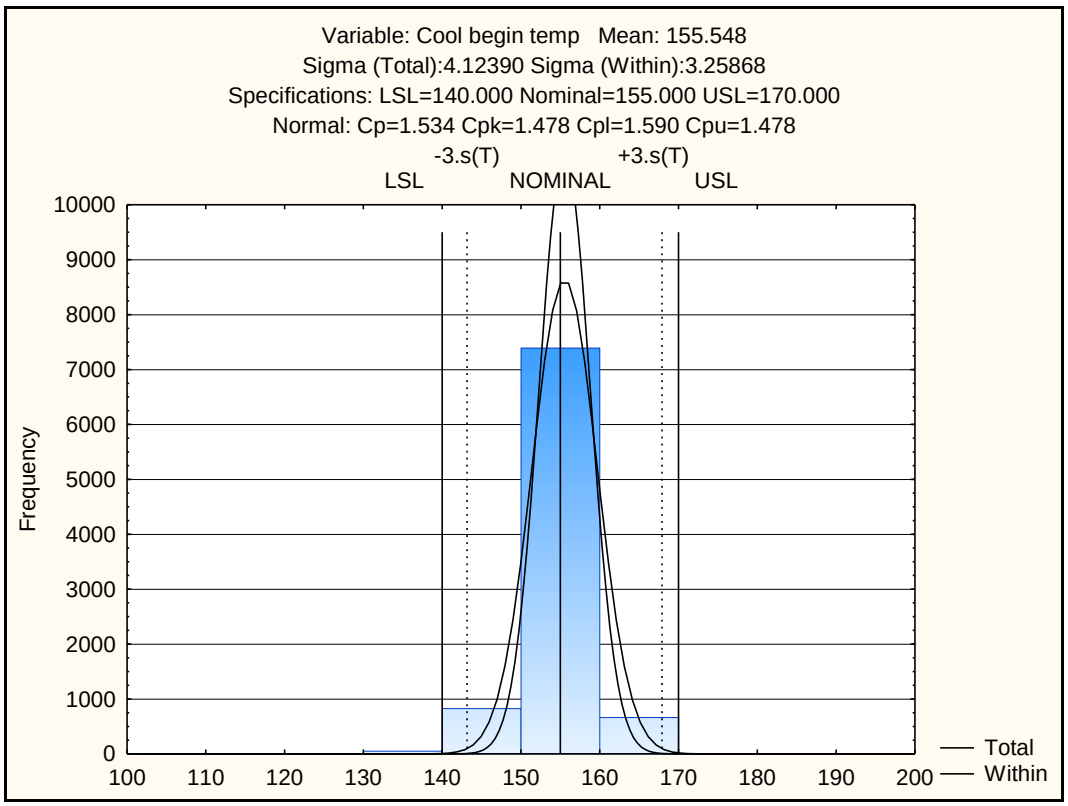
opportunity to move the operating level to a higher (+) operating level as proposed by the DOE model, refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a higher operating level; refer to Table 10.2, row 4 that is confirmed by the box plot.

From Table 10.2, because variable 1 is not statistically significant for all three models, see Table 10.2; rows 6-8, slope comparisons and process operating level adjustments are irrelevant and may have no impact on the process.

10.3.2 ANALYSIS x_2



Graphs 10.3: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Cool begin temp (x_2)



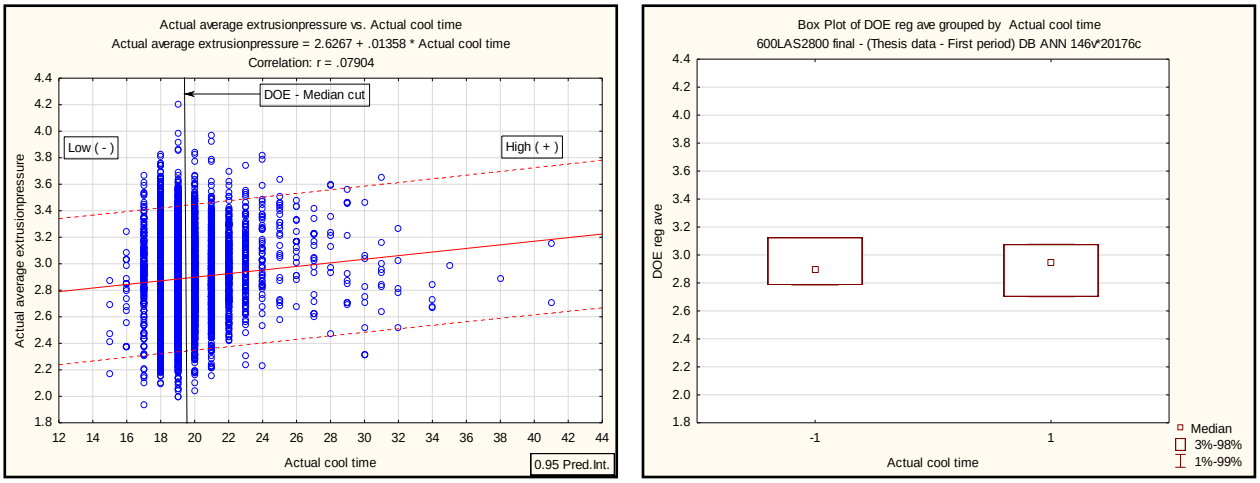
Graph 10.4: Cool begin temperature (x_2): Process capability chart

Variable 2 is not significant for DOE and individual regression models, but is significant for MR; refer to Table 10.2, rows 6, 7 and 8. Slopes are negative for both MR and box plot; refer to Graphs 10.3. Process capability is greater than 1; refer to Graph 10.4, which provides an opportunity to move the operating level to a lower (-) operating level as proposed by the DOE model, refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a lower operating level; refer to Table 10.2, row 4.

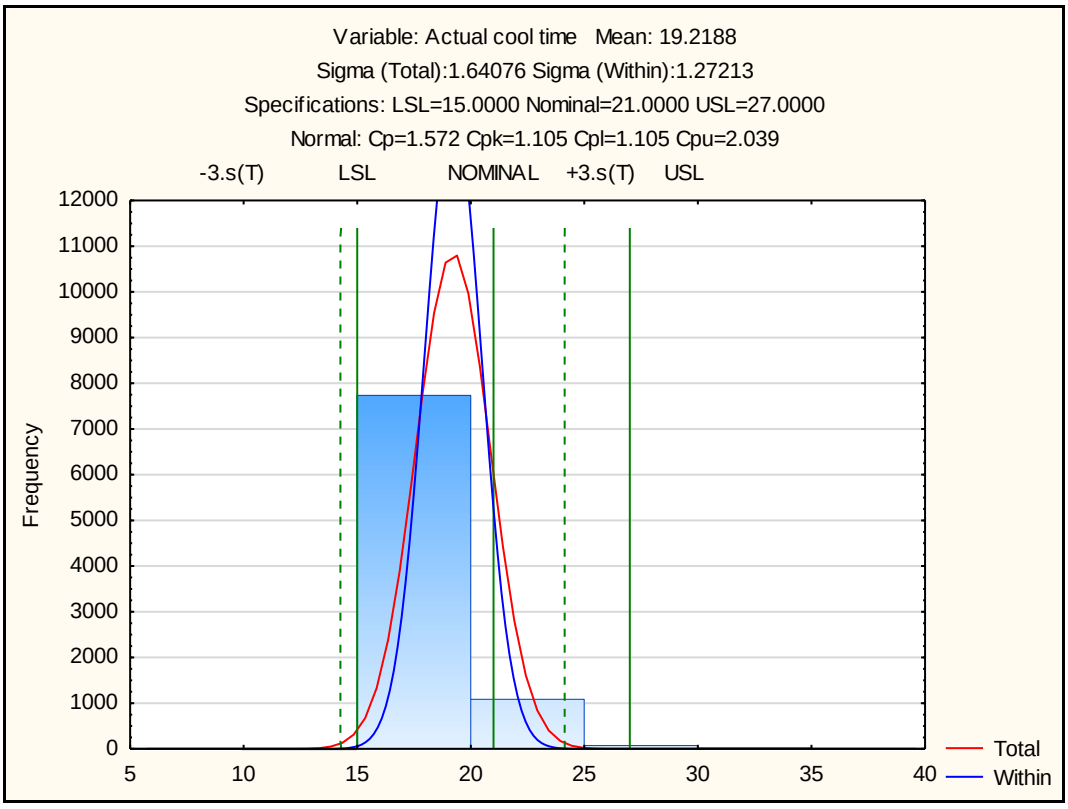
From the summary above, variable 2 is not significant for DOE and single regression models, but significant for MR. The negative slopes for both the MR and box plots follow the same slope as the DOE model representing the min cost, which increases the confidence of the box plot representing variable 2. Although the capability index is greater than 1, not much room is available for process operating level adjustment, proposed by the DOE model; refer to Graph 10.4.

Variable 2, not significant for the DOE and individual regression models, may have little impact on process improvement, based on box plot slope comparison; refer to Table 10.2, row 2, to DOE model operating level; refer to Table 10.2, row 1.Process operating level adjustments are irrelevant and may have no impact on the process.

10.3.3 ANALYSIS x_3



Graphs 10.5: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Actual cool time (x_3)



Graph 10.6: Actual cooling time (x_3): Process capability chart

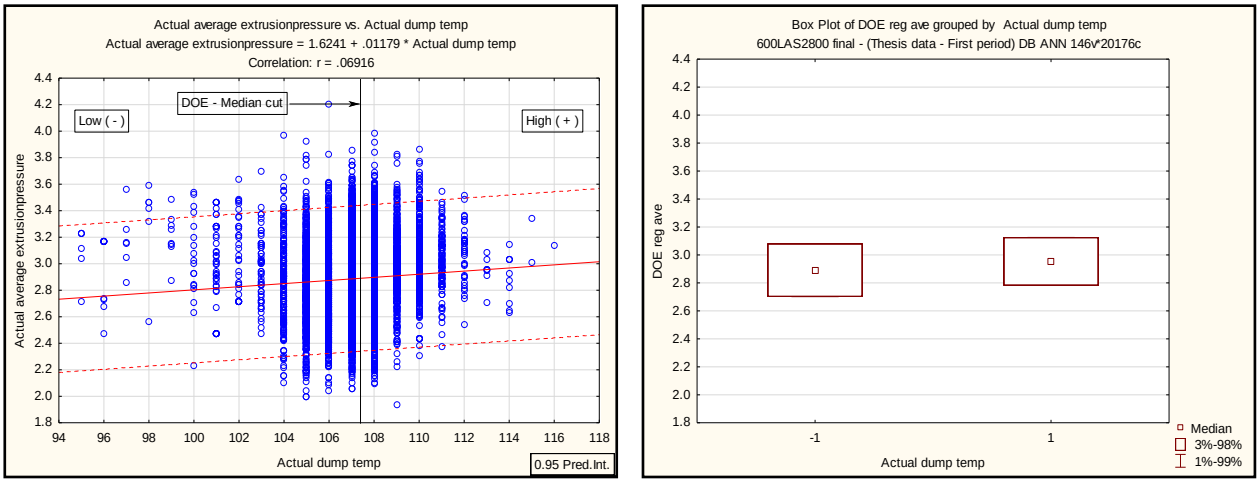
Variable 3 is not significant for DOE, but is significant for MR and single regression models; refer to Table 10.2, rows 6,7 and 8. Slopes are positive for both MR and box plot; refer to Graphs 10.5. Process capability is greater than 1; refer to Graph 10.6, ,

which provides an opportunity to move the operating level to a lower (-) operating level as proposed by the DOE model; refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a lower operating level; refer to Table 10.2, row 4.

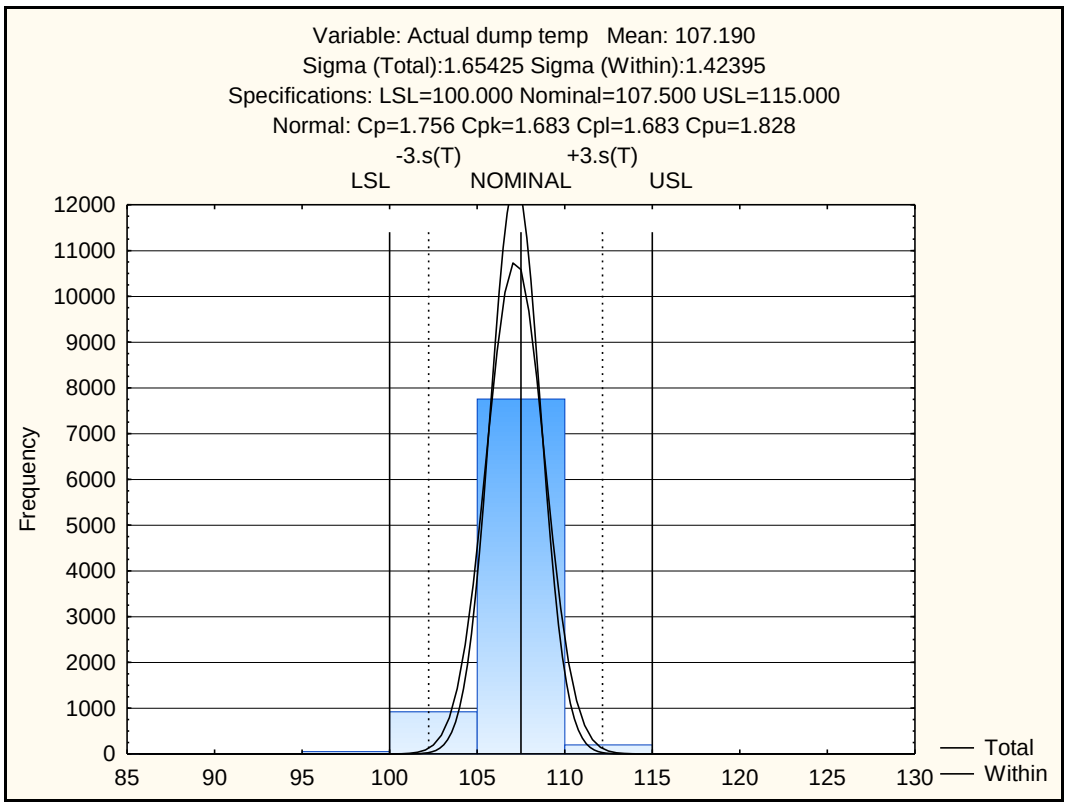
From the summary above, variable 3 is not significant for DOE, but significant for MR and single regression models. The positive slope for the box plots does not follow the same slope as the DOE model representing the min cost, which decreases the confidence of the box plot representing variable 3. Although the capability index is greater than 1, process adjustment is available for process operating level adjustment, but to the opposite level proposed by the DOE model; refer to Graph 10.6 which is not in line with the validation criteria.

Variable 3, not being significant for the DOE model, may not have an impact on process improvement, based on box plot slope comparison; refer to Table 10.4, row 2 to DOE model operating level; refer to Table 10.2, row 1. Process operating level adjustments will have a negligent effect and may have no impact on the process.

10.3.4 ANALYSIS x_4



Graphs 10.7: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Actual dump temp (x_4)



Graph 10.8: Actual dump temperature (x_4): Process capability chart

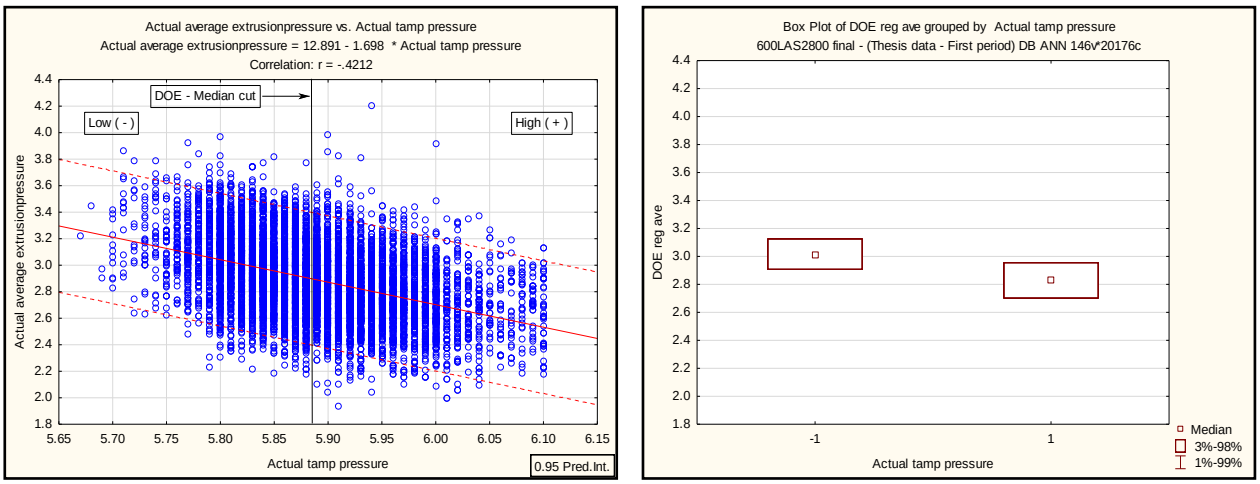
Variable 4 is significant for DOE, MR and single regression models; refer to Table 10.2, rows 6,7 and 8. Slopes are positive for MR and box plot, refer to Graphs 10.7. Process capability is greater than 1; refer to Graph 10.8, which provides an opportunity to move the operating level to a higher (+) operating level, as proposed by the DOE model; refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a higher operating level; refer to Table 10.2, row 4.

From the summary above, variable 4 is significant for all three models. The positive slope for the box plots follows the same slope as the DOE model representing the min cost which increases the confidence of the box plot representing variable 4. Although the capability index is greater than 1, process adjustment is restricted for process operating level adjustment as proposed by the DOE model; refer to Graph 10.8 which is in line with the validation criteria.

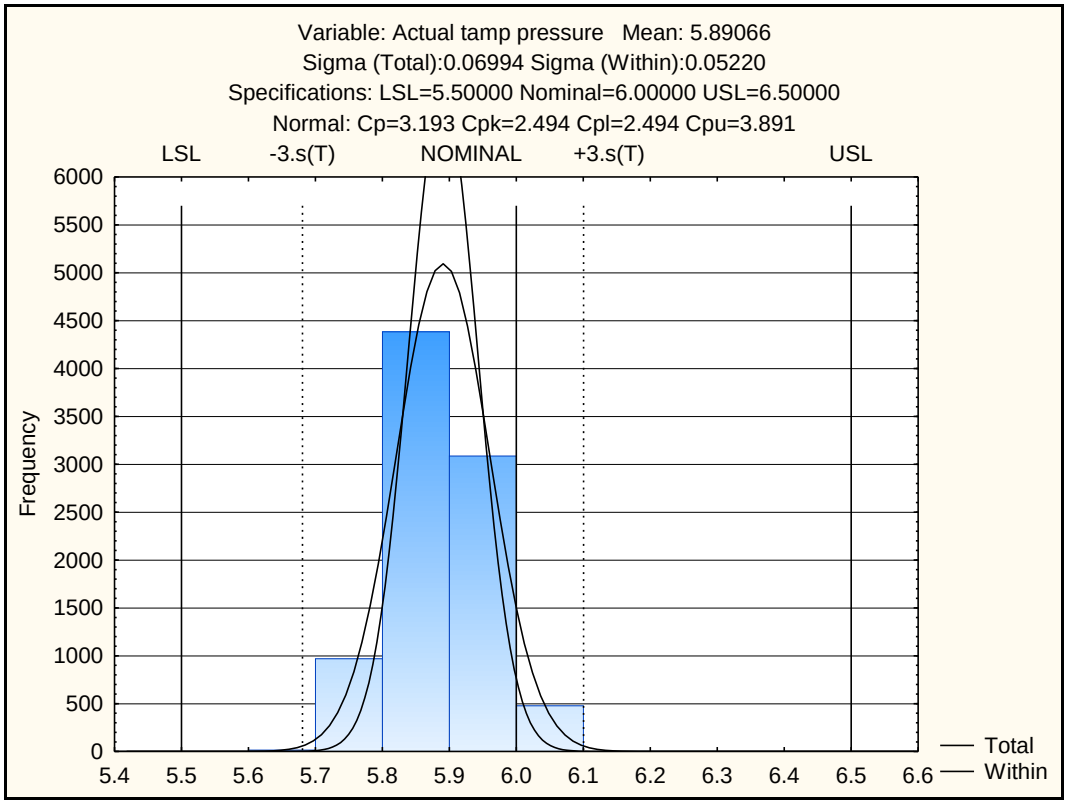
Variable 4, being significant for MR, DOE and single regression models, may have an impact on process improvement; refer to Graph 10.6. The compatibility for MR and box plot slopes comparison is also evident by Table 10.2, row 2 to DOE model operating level: refer to Table 10.2, row 1. The proposed DOE model, Table 10.2, row 1, also

follows the min cost slope; refer to Table 10.2, row 4. Variable 4 fulfils the validation criteria for selecting a variable for the prediction DOE model.

10.3.5 ANALYSIS x_5



Graphs 10.9: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Actual tamp pressure (x_5)



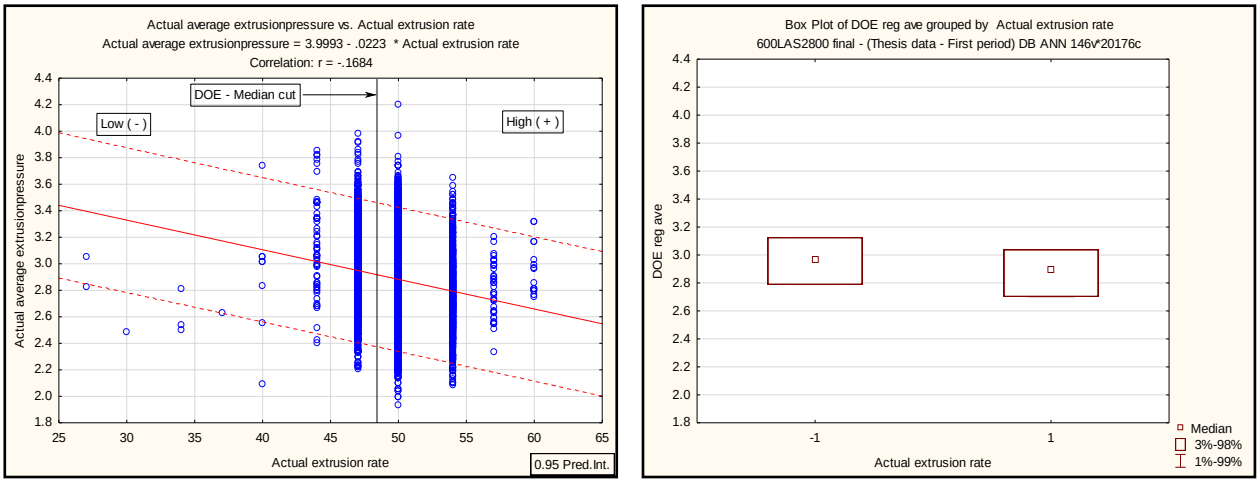
Graph 10.10: Actual tamp pressure (x_5): Process capability chart

Variable 5 is significant for DOE, MR and single regression models, refer to Table 10.2, rows 6,7 and 8. Slopes are negative for MR and box plot; refer to Graphs 10.9. Process capability is greater than 1; refer to Graph 10.10, which provides an opportunity to move the operating level to a lower (-) operating level as proposed by the DOE model; refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a lower operating level; refer to Table 10.2, row 4.

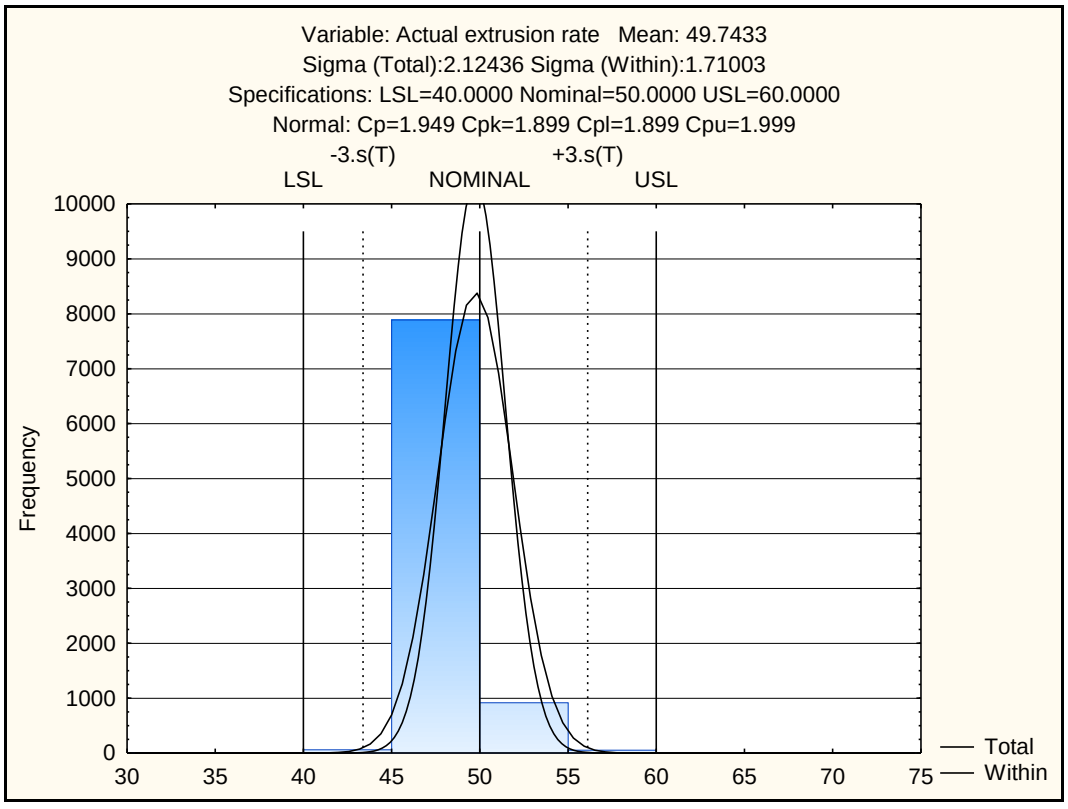
From the summary above, variable 5 is significant for all three models. The negative slope for the box plots follows the same slope as the DOE model representing the min cost, which increases the confidence of the box plot representing variable 5. For variable 5 the capability index is greater than 1, and therefore process adjustment is not restricted for process operating level adjustment to the lower level as proposed by the DOE model; refer to Graph 10.10 which is in line with the validation criteria.

Variable 5, being significant for all three models, may have an impact on process improvement, refer to Graph 10.10. The compatibility for MR and box plot slopes comparison is also evident by Table 10.2, row 2 to DOE model operating level; refer to Table 10.2, row 1. The proposed DOE model, Table 10.2, row 1 also follows the min cost slope; refer to Table 10.2, row 4. Variable 5 fulfils the validation criteria for selecting a variable for the prediction DOE model. This variable has the largest impact on process output based on effect for all three models. **This variable is also defined as the “red X” in the regression analysis chapter.**

10.3.6 ANALYSIS x_6



Graphs 10.11: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Actual extrusion rate (x_6)



Graph 10.12: Actual extrusion rate (x_6): Process capability chart

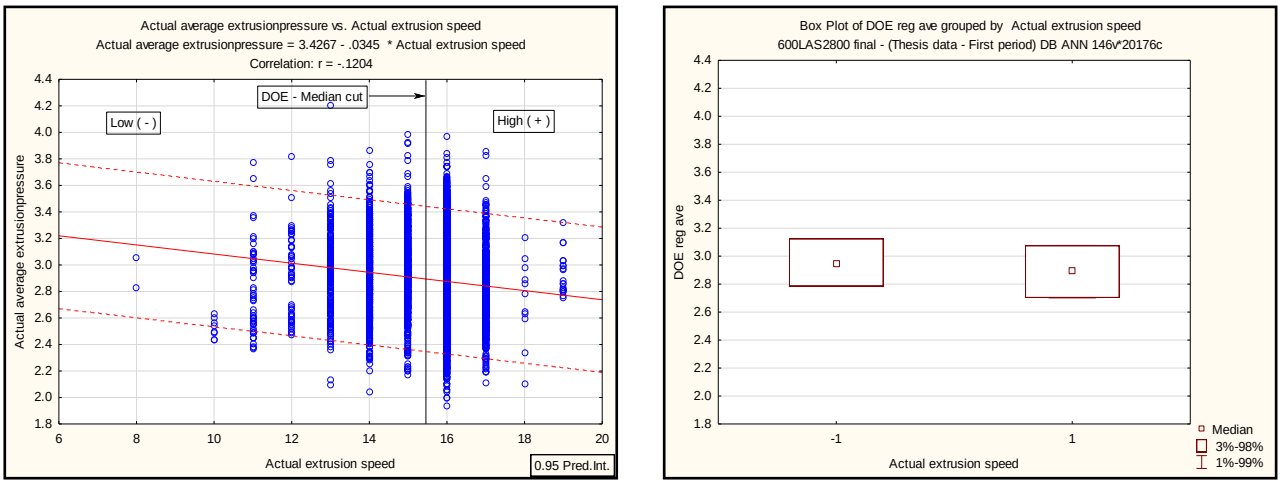
Variable 6 is significant for DOE, MR and single regression models; refer to Table 10.24, rows 6,7 and 8. Slopes are negative for MR and box plot, refer to Graphs 10.11. Process capability is greater than 1; refer to Graph 10.12, which provides an opportunity

to move the operating level to a lower (-) operating level, as proposed by the DOE model; refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost propose a lower operating level; refer to Table 10.2, row 4.

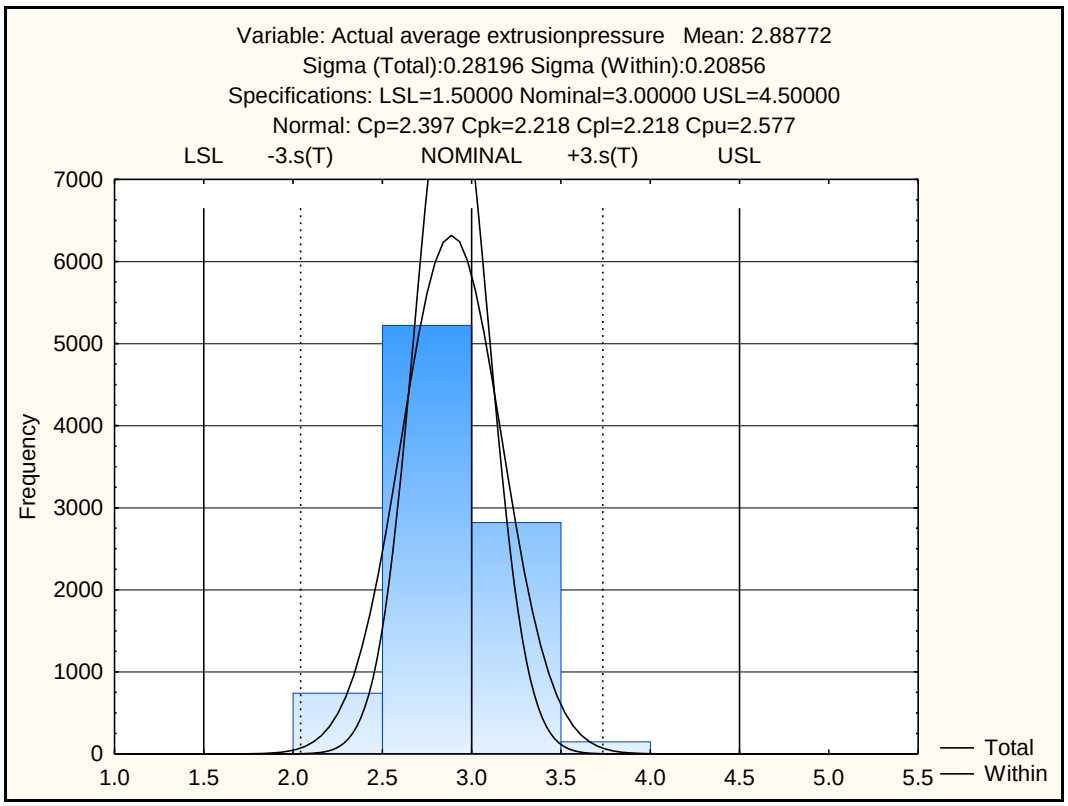
From the summary above, variable 6 is significant for all three models. The negative slope for the box plots follows the same slope as the DOE model representing the min cost, which increases the confidence of the box plot representing variable 6. For variable 6 the capability index is greater than 1, and therefore process adjustment is not restricted for process operating level adjustment to the lower level as proposed by the DOE model; refer to Graph 10.12 which is in line with the validation criteria.

Variable 6, being significant for all three models, may have an impact on process improvement; refer to Graph 10.12. The compatibility for MR and box plot slopes comparison is also evident by Table 10.2, row 2 to DOE model operating level; refer to Table 10.2, row 1. The proposed DOE model, Table 10.2, row 1, also follows the min cost slope; refer to Table 10.2, row 4. Variable 6 fulfils the validation criteria for selecting a variable for the prediction DOE model.

10.3.7 ANALYSIS x_7



Graphs 10.13: Comparative graphs: Regression (Scatter plot) and DOE regression (Box plot) for Independent variable (Actual extrusion speed (x_7)



Graph 10.14: Actual extrusion speed (x_7): Process capability chart

Variable 7 is significant for DOE and single regression model, but not for MR; refer to Table 10.2, rows 6, 7 and 8. Slopes are level for MR and negative for the box plot; refer to Graphs 10.13. Process capability is greater than 1; refer to, Graph 10.14, which provides an opportunity to move the operating level to a lower (-) operating level, as proposed by the DOE model, refer to Table 10.2, row 1. The DOE run combinations representing the lowest cost, propose a lower operating level; refer to Table 10.2, row 4.

From the summary above, variable 7 is significant for DOE and single regression models, but not for MR. The negative slope for the box plots follows the same slope as the DOE model representing the min cost which increases the confidence of the box plot representing variable 7. For variable 7, the capability index is greater than 1, and therefore process adjustment is not restricted for process operating level adjustment to the lower level as proposed by the DOE model; refer to Graph 10.14 which is in line with the validation criteria.

Variable 7, being significant for DOE and single regression models but not for MR, the DOE model may have an impact on process improvement; refer to Graph 10.14. The compatibility for the box plot slope comparison is also evident by Table 10.2, row 2 to

DOE model operating level, refer to Table 10.2, row 1. The proposed DOE model, Table 10.2, row 1, also follows the min cost slope; refer to Table 10.2, row 4. Variable 7 fulfils the validation criteria for selecting a variable for the prediction DOE model.

CHAPTER 11

REFLECTIONS ON STUDY AND FUTURE WORK

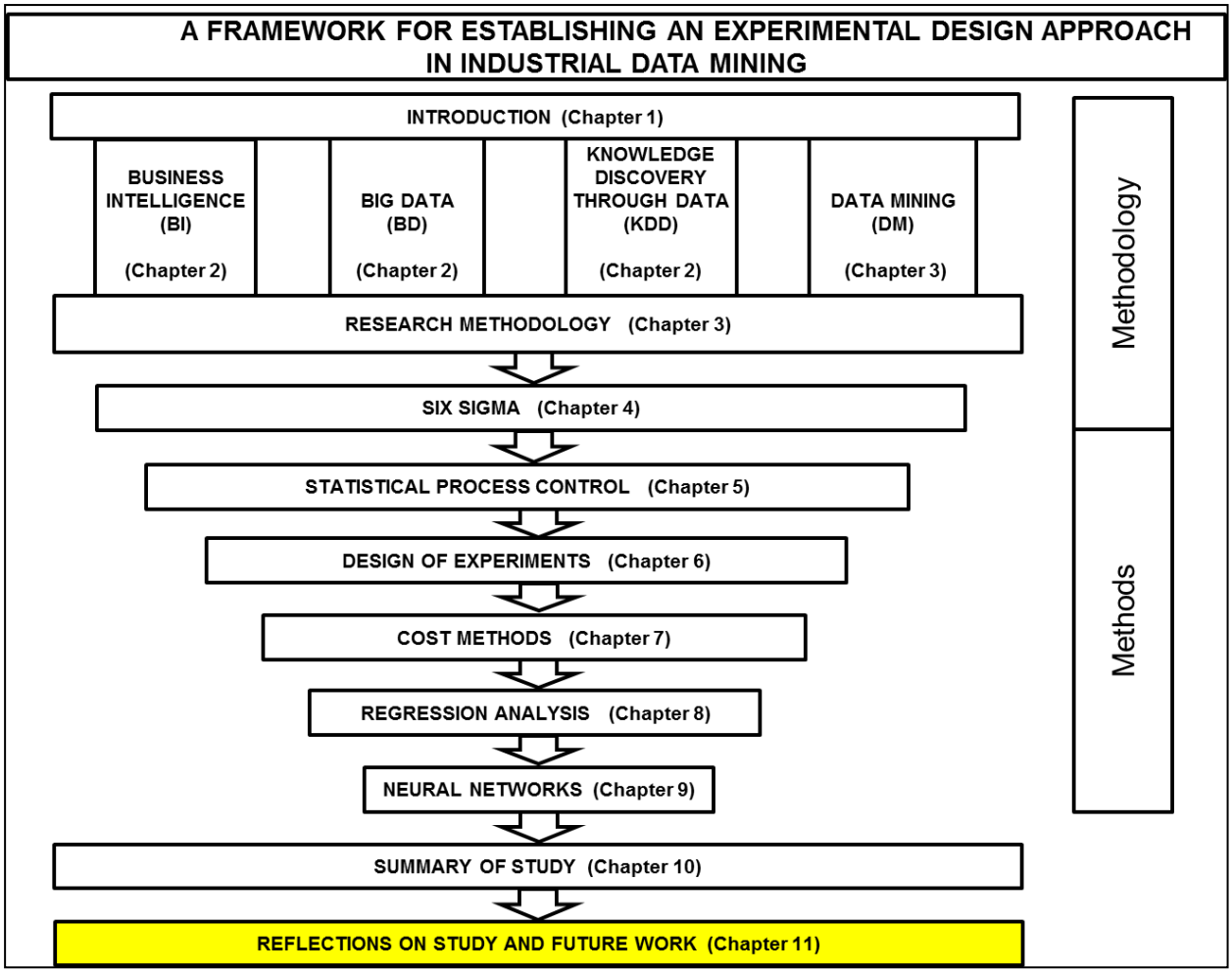


Figure 11.1: Framework of the study

11.1 REFLECTION ON FRAMEWORK

The framework for this study, see Figure 11.1, is a generic framework for analysts and management to follow when an extensive data analysis is considered. This framework with the embedded experimental analysis design serves as an analytical roadmap for process development and improvement and was used specific to the case study. The development process of this framework was focused on achieving the main goals of this research. These main goals were:

Goal 1: To accommodate DOE as a Data Mining Technique in an Industrial Data Mining environment.

Goal 2: To enhance the awareness of expanding DOE as a statistical approach to complement existing methods and methodologies used for Data Mining.

Goal 3: To validate the integrity of captured data through the refining process to determine upper and lower operating conditions required by DOE, any abnormal data points will be exposed.

Goal 4: To focus on Industrial Data Mining, and concentrate on process data, applying DOE rather than generic, traditional Data Mining techniques.

Goal 5: To develop a methodology to accommodate the use of DOE as a Data Mining technique to determine impacts of variables on process outcomes through experimenting with data within current databases.

Proposing DOE as the core for industrial datamining focuses the analytical process on process improvement through scientific experimentation through different operating conditions to determine optimal process conditions with the lowest cost impact for the company.

A difference in the analytical approach presented by this framework to similar frameworks is that it is based on historic data to reduce the risk of costly experimentation with untested data. Working with historic data as the basis from estimation to validation should provide a huge financial benefit to management. The case study presented in this research shows that this approach is viable even though the future is unpredictable. Generally, regression analysis is used for analysing historic data to predict future process levels. For this research, DOE is used for the same purpose through an experimental design approach, with regression analysis as a control technique to evaluate DOE experimental conditions.

Following the framework through using the analytical methods forces the analytical process to follow a sequential process that funnels the empirical analysis from determining the research methodology to NN analysis. From a pragmatic industrial operating perspective, the funnelled sequential empirical analytical process keeps analysts and management focused on a prescriptive analytical roadmap to restrain from including non-value-added issues that could clutter the objective of the study.

Including regression analysis near to the end of the empirical analysis process for this study had the advantage to utilise it as a control analysis to evaluate the DOE operating

conditions done through experimental analysis. This also includes the identifying and validating the red “X” variable that were identified by SPC, DOE, capability analysis and the holistic analytical summary in chapter 10. Regression analysis confirmed the same variable as the variable having the largest impact on process performance. The red “X” seems to be the variable having the largest impact on process changes and therefore may be used for controlling future process performance. Regression analysis for this study was extended for prediction purposes as well.

Lessons learned through this research:

Empirical analysis should follow a bottom up and not a top down approach. This means that the analytical process should start with analysing individual variables, then progress to multi-variate analysis if necessary, not the other way around. It saves time and a better understanding of the process data will be realised.

The challenge for this study was not the empirical analysis, but to present the analytical process, findings and solutions to interested parties in such a way as to maintain interest. Keep it basic, simple and high level, then drill down if need be.

Follow the process, objectively validating the results, refrain from explaining deviations, and jump to conclusions based on the accepted norm and personal experience to fit your understanding. This was extremely difficult in the beginning but became natural towards the end of this research; a valuable personal analytical lesson.

From a Six Sigma (SS) perspective, the study for this research followed the DMAIC methodology. Table 11.1 presents a high level DMAIC representation of this study. The summary shows the comparative chapters to the DMAIC methodology, which strengthens the analytical structure of the study showing that, within the proposed framework (see Figure 11.1) the sequence of applying analytical methods also coincide with the DMAIC methodology.

Although the DMAIC process coincides with the proposed framework on high level, it is of equal importance to recognise that the DMAIC process does occur within each referenced chapter during the analytical process; see Table 11.1.

Six Sigma (DMAIC)	Framework Application	Study chapter reference for DMAIC	Description
Define	Identifying process and data basis	Chapter 4	Project definition and database selection for study
Measure	SPC	Chapter 5	Screening and capability analysis
Analyse	DOE	Chapter 6	Determine and evaluate experimental conditions (runs).
Improve	Cost methods	Chapter 7	Compare experimental conditions (runs) and optimal region.
Control	Regression and NN	Chapter 8	Control analysis of DOE and comparative analysis with regression

Table 11.1: DMAIC methodology showing study framework and chapter reference

The difference for following the proposed approach is that the experimenter will know the probable impact of changing inputs on the process changes before any dynamic in-line process changes are made. In addition, changes can be introduced beyond the test model's parameters with a higher degree of confidence than when developing a DOE process development model based on guesswork, gut feel, experience and the high probability of high cost implications associated with trial and error experimental runs.

11.2 FUTURE WORK

Although significant main, two-way and three-way effects are present for the 2** (7-0) resolution FULL 128 run model, the impact of missing values, sign changes for 2^{nd} and 3^{rd} order interactions on the results are unknown and therefore will not be considered for this study. A more comprehensive analysis is necessary to evaluate the impact sensitivity to missing values and sign changes on significant effects.

Screen out known sources of variation by blocking. When grouping experimental runs into homogeneous blocks, variation such as raw materials, machine differences or shift changes, screen out noise by known variation. Blocking was not considered for this study because raw process data represent one operation and homogeneous blocking of any variable that may contribute to noise variation was extremely difficult to accommodate. Consider the influence of blocking with a DOE design with current data for future work.

The assumption that Independent variables with fixed processing mechanical settings or set points, determined by the company's global R&D team removed during phase 1 of the data cleaning process, do not have a significant impact on the results. Test this assumption through a separate statistical analysis.

A possible reason for small insignificant interactions is that the band for minimum and maximum values selected from the historical data set was too narrow and therefore could not calculate real significant interactions. The selection of minimum and maximum values seem to have a major impact on significant effects. The other possibility is that there are no significant interactions. Investigate the impact of boundary changes on interaction sensitivity.

Additional models to evaluate screened data for this study are:

A fit for inverse of x compared to y

A third order polynomial fit

A log transformation model

Fit proposed and existing models on screened data for the validation period that only includes data applicable for the changed raw material.

By exploring alternative models will assist in choosing possible alternatives to the proposed models.

11.3 OVERALL CONCLUSIONS

The analytical approach initially started with a top down approach using MR analysis to evaluate the significance of independent variable contribution within a MR model. During the analytical process, the need for a bottom up approach became evident for individually evaluating the significance of each independent variable to the process output. By doing both approaches, a holistic analytical approach was achieved that led to multidimensional model options for process improvement.

The process of transforming management thinking from utilising data in existing formats for strategic decision making into a process of effectively transforming and analysing data that will have a sustainable and measurable effect on the strategic direction of a business, was not only a huge opportunity but will be a strategic challenge for management in general.

To separate BI, BD and KDD into independent silos in terms of data management is not pragmatic, because these concepts overlap and focus on large databases. The specifics may vary but the aim is to handle the modern explosion and availability of data for analysts to manage effectively for strategic decision making on an operational and executive level. The challenge for modern management is not only to handle the data but also to be smart in designing databases, extracting useful information, handling large amounts of data that are increasing on a daily basis and ensuring that managing data becomes one of their key performance indicators.

None of the independent or dependent variables in this study are in statistical control; many out of control points as well as pattern-like trends, cycles and shifts were present. Even though these patterns exist, the process specification limits for each of these variables are wide enough to accommodate these patterns. For this reason, the variable selection was based on the ability to split each independent variable with the median into near 50/50 data points.

From a theoretical statistical process control perspective, a process must be in statistical control before changing it. In reality, however, the market mostly dictates the time line for new product development. Most product changes are forced using current reality and discarding the notion of first having process stability before changing a product dictated by the market. This sometimes creates a strategic clash of interest amongst management. This study assisted in evaluating the risks between these two options, and shows that normality is not a prerequisite for analysing data, but it can be approximated by using subgroups of size 40.

To base the selection of an independent variable for a DOE model on the normality assumption should not be the only criterion. The importance of combining a distribution with a trend line is critical to graphically show the time-line of a variable. This will prevent using normality tests only when determining if normal variation is present. Even if the normality test shows non-normal distribution, a graphical representation greatly assists in evaluating an independent variable.

A fundamental criterion for the proposed Experimental design approach to be effective is the risk of having missing data that may cause some DOE runs with no data. This is because for each independent variable that forms the unique combination of high and low values for an experimental run, the database halves. For this reason the database

size is critical, the larger the better. The gaps in the experience region are also identified by the missing values for certain combinations, which give opportunities to experiment and create new data.

For this study, all of the independent variables, irrespective if the DOE regression coefficients are significant, will be used for predicting process output. This is because all seven selected variables are critical to the holistic process, which is a pragmatic management decision for this specific process.

Prediction accuracy is lower for the second period when compared to the first period. The lower prediction accuracy is because of a technical change to a raw material during the second period that caused the process to operate on a different level than the first period.

Prediction error for DOE and MR is low across all categories for BIAS, MSE, MAD and MAPE, but DOE regression results consistently lower and therefore seems the best option for predicting future process behaviour compared to multiple regression. Even with a simple process adjustment to the second period database to align the operating level between the two periods, the best predictor is still the DOE regression.

Calculated cost based on the Nominal the best loss function is similar for all three categories in the estimated period, but is on the lower side of the target value, which in terms of product performance is better than operating towards the high side of the target value. Costs are similar for all three categories in the second period as compared to the first period, but much lower because the experimental outcomes are closer to the target value. The lower costs for the second period in terms of product performance have a small risk of producing out of specification.

The ideal experimental runs for process improvement are those with high S/N ratios because they show minimal external variation influencing operating outcomes. The initial sixteen experimental runs 1 – 9, 11 – 16 and 28 selected for low non-conforming risk based on risk and cost profiles across the three periods did not consider S/N ratios. Each of these runs has a specific signal to noise ratio that may influence the final experimental run selection. Ranking the associated S/N ratios of each selected run from high to low provides a run sequence selection to minimize risk and cost.

Applying the proposed framework for process optimisation studies in any company where needed should enhance process improvement, because this research is about following a new experimental analysis design approach that is generic for any process development and improvement, irrespective of the product rendered. The framework and techniques used in this research are applicable within any processing plant where multiple variables affect product quality.

BIBLIOGRAPHY

Adriaans, P. & Zantinge, D. 1996. Data-mining. London: Addison – Wesley.

Al-Azmi, A.A.R. 2013. Data, text and web mining for Business Intelligence: a survey. *International journal of data mining and knowledge management process*, 3(2):1-21.

Alazmi, A.R. & Alazmi, A.R. 2012. Data mining and visualization of large databases. *International journal of computer science and security*, 6(5):295-314.

Antony, J. 2014. Design of experiments for engineers and scientists. 2nd ed. London: Elsevier.

AT & T Technologies. 1985. Statistical quality control handbook. New York: AT & T Technologies Company.

Atkinson, P. 2014. DMAIC: a methodology for Lean Six Sigma business transformation. *Management services*, 58(1):12-17.

Ayobami, A. & Rabi'u, S. 2012. Knowledge discovery in database: a knowledge management strategic approach. Paper presented at the 6th Knowledge Management International Conference, Johor Bahru, Malaysia, July. http://ssrn.com/abstract=2088965 Date of access: 4 Jun. 2015.

Azhar, S., Ahmad, I. & Sein, M. 2010. Action research as a proactive research method for construction engineering and management. *Journal of construction engineering and management*, 136(1):87-98.

Baran, A. 2011. Managing quality. http://www.slideshare.net/abrn/managing-quality-8256932 Date of access: 23 Jul. 2015. [PowerPoint presentation].

Berry, M.J.A. & Linoff, G. 1997. Data mining techniques for marketing, sales, and customer support. New York: John Wiley.

Berson, A. & Smith, S.J. 1997. Data warehousing, data mining and OLAP. New York: McGraw Hill.

Bigus, J.P. 1996. Data mining with neural networks. New York: McGraw Hill.

Boslaugh, S. 2013. Six Sigma: quality control standard. Salem Press Encyclopedia. http://eds.a.ebscohost.com.nwulib.nwu.ac.za/eds/detail/detail? sid=0d11d5d0-af59-4b16-8a5d-55378b394047%40sessionmgr4004&vid=0&hid =4110&bdata= JnNpdGU9ZWRzL WxpdmU%3d#db=ers&AN=89677627 Date of access: 16 Sep. 2014.

Bowman, D. 2009. Enterprise Business Intelligence. http://www.information-management-architect.com/enterprise-business-intelligence.html Date of access: 30 Apr. 2011.

Box, G.E.P., Hunter, W.G. & Hunter, J.S. 1978. Statistics for experimenters: an introduction to design, data analysis, and model building. New York: John Wiley.

Box, G.E.P. & Draper, N.R. 1969. Evolutionary operation: a statistical method for process improvement. New York: John Wiley.

Brešić, B. 2012. Knowledge acquisition in databases. *Management information systems*, 7(1):32-41.

Brightman, H.J. 1999. Data analysis in plain English with Microsoft Excel. Brooks/Cole: Duxbury Press.

Brown, M.L. & Kros, J.F. 2003. Data mining and the impact of missing data. *Industrial management and data systems*, 103(8):611-621.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. 1997. Discovering data mining from concept to implementation. New Jersey: Prentice-Hall.

Cano, E.L., Moguerza, J.M. & Redchuk, A. 2012. Six Sigma with R: statistical engineering for process improvement. New York: Springer.

Cao, L. 2007. Domain-driven, actionable knowledge discovery. *IEEE intelligent systems*, 22(4):78-88.

Creswell, J. 2008. Educational research: planning, conducting and evaluating quantitative and qualitative research. 3rd ed. New Jersey: Prentice Hall.

De Carvalho, G., Da Silva, C.E. & Costa, K. 2014. Application of Six Sigma methodology in improving of the industrial production processes. *Applied mechanics and materials*, 627(Sep):327-331.

Desale, S.V. & Deodhar, S.V. 2013. Lean Six Sigma principal in construction: a literature review related to abstract. *Journal of information, knowledge and research in civil engineering*, 2(2):133-139.

Dunham, M.H. 2003. Data mining introductory and advanced topics. New Jersey: Prentice Hall.

Dutta, D. & Bose, I. 2015. Managing a Big Data project: the case of Ramco Cements Limited. *International journal of production economics*, 165(Jul): 293-306.

EI-Haik, B.S. & Shaout, A. 2011. Software design for Six Sigma: a roadmap for excellence. Hoboken, NJ: John Wiley.

Enoch, O.F., Shuaib, A. & Hasbullah, A.H.B. 2015. Applying P-diagram in product development process: an approach towards design for Six Sigma. *Applied mechanics and materials*, 789-790(Sep):1187-1191.

Erohin, O., Kuhlang, P., Schallow, J. & Deuse, J. 2012. Intelligent utilisation of digital databases for assembly time determination in early phases of product emergence. *Procedia CIRP*, 3(Dec):424-429.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37-54.

Feldkamp, N., Bergmann, S. & Strassburger, S. 2015. Knowledge discovery in manufacturing simulations. Proceedings of the 3rd ACM SIGSIM Conference on principles of advanced discrete simulation. http://dl.acm.org/citation. cfm?id=2769468 Date of access: 12 Apr. 2016.

Forsman, S. 1997. What is OLAP? http://www.olapcouncil.org/research/ whtpaply.htm Date of access: 10 Mar. 2015.

Freund, J.E. 1988. Modern elementary statistics. 7th ed. New Jersey: Prentice-Hall.

Global Intel. 1998. What is data-mining? http://www.globalintel.com/data.htm Date of access: 7 Apr. 2009.

Goes, P.B. 2013. Editor's comments: commonalities across IS silos and intradisciplinary information systems research. *Management information systems quarterly*, 37(2):iii-vii.

Gray, D.E. 2012. Doing research in the real world. 3rd ed. Los Angeles: Sage.

Gygi, C., DeCarlo, N. & Williams, B. 2005. Six Sigma for dummies. Hoboken, NJ: John Wiley.

Gygi, C., DeCarlo, N. & Williams, B. 2010. Six Sigma for dummies. 2nd ed. Hoboken, NJ: John Wiley.

Hamdy, A.T. 2007. Operations Research: An introduction. 8th ed. New York: Pearson Prentice Hall.

Han, J. & Kamber, M. 2006. Data mining concepts and techniques. 2nd ed. San Francisco: Diane Carra.

Hand, D.J. 1998. Data-mining: statistics and more? *The American statistician*, 52(2):112-118.

Hedges, L.V. 2008. Basic experimental design. www.ipr.northwestern.edu/.../NCER2008/8hedges130pm/basicdesign.ppt Date of access: 10 Sep. 2014. [PowerPoint presentation].

Hermida, J.M., Meliá, S., Montoyo, A. & Gómez, J. 2013. Applying model-driven engineering to the development of rich internet applications for business intelligence. *Information systems frontiers*, 15(3):411-431.

Hines, W.H. & Montgomery, D.C. 1990. Probability and statistics in engineering and management science. New York: John Wiley.

Horníková, A., Durakbasa, N.M., Güclü, E. & Bas, G. 2011. Data mining: novel statistical method. Paper presented at the 8th International Conference, Smolenice, Slovakia. www.measurement.sk/M2011/doc/proceedings/ 012_Durakbasa-1.pdf Date of access: 4 Apr. 2015.

Hosking, J.R. & Wallies, J.R. 2005. Regional frequency analysis. Cambridge: Cambridge University Press. http://www.olapreport.com/FASMI.HTM Date of access: 10 Jan. 2015.

Jha, R. 2014. Association rules mining for Business Intelligence. *International journal of scientific and research publications*, 4(5):1-5.

Kai, Y. & Basem, S.E. 2009. Design for Six Sigma: a roadmap for product development. 2nd ed. New York: McGraw-Hill.

Karout, R. 2015. A DMAIC framework for improving software quality in organizations: case study at RK company. Quebec : Concordia University. (Dissertation - Master of Applied Science).

Khuri, A.I. & Cornell, J.A. 1987. Response surfaces: designs and analyses. New York: ASQC Quality Press.

Kleppmann, W. 2014. Design of Experiments (DoE): optimizing products and processes efficiently. *Chemical engineering*, 121(11):50-57.

Koch, R. 2015. From Business Intelligence to predictive analytics. *Strategic finance*, 97(1):56-57.

Kopčeková, A., Kopček, M. & Tanuška, P. 2013. Business Intelligence in process control. *Research papers Faculty of Materials Science and Technology Slovak University of Technology*, 21(33):43-53.

Kučerová, M. & Fidlerová, H. 2014. Improving the quality of manufacturing process through Six Sigma application in the automotive industry. *Applied mechanics and materials*, 693(Dec):147-152.

Kumar, P. 2012. Impact of Business Intelligence systems in Indian telecom industry. *Business Intelligence journal*, 5(2):358-366.

Lamont, J. 2012. Big data has big implications for knowledge management. *Knowledge management world*, 21(4):8-11.

Lasi, H. 2013. Industrial intelligence: a business intelligence-based approach to enhance manufacturing engineering in industrial companies. *Procedia CIRP*, 12(Sep):384-389.

Launsby, R.G. & Schmidt, R.S. 1991. Understanding industrial designed experiments. Colorado: Air Academy Press.

Liu, Q.Y., Shi, J.L. & Xing, S.Y. 2014. The application research of statistical process control in quality management. *Advanced materials research*, 926-930(May):4000-4003.

Lochner, R.H. & Matar, J.E. 1990. Designing for quality: an introduction to the best of Taguchi and western methods of statistical experimental design. New York: ASQC Quality Press.

Mackenzie, N. & Knipe, S. 2006. Research dilemmas: paradigms, methods and methodology. *Issues in educational research*, 16(2):193-205.

Mackinnon, M.J. & Glick, N. 1999. Data mining and knowledge discovery in databases: an overview. *Australian and New Zealand journal of statistics,* 41(3):255-275.

Madhuri, V.J. 2013. Significance of data warehousing and data mining in business applications. *International journal of soft computing and engineering*, 3(1):329-333.

Maimon, O. & Rokach, L. 2010. Data mining and knowledge discovery handbook. 2nd ed. New York: Springer.

Mason, R.L., Gunst, R.F. & Hess, J.L. 1989. Statistical design and analysis of experiments with applications to engineering and science. New York: John Wiley.

Mikroyannidis, A. & Theodoulidis, B. 2010. Ontology management and evolution for Business Intelligence. *International journal of information management*, 30(6):559-566.

Moeinaddini, M., Asadi-Shekari, Z. & Shah, M.Z. 2014. Analysing the relationship between park-and-ride facilities and private motorised trips indicators. *Arabian journal for science and engineering*, 39(5):3481-3488.

Nedelcu, B. 2013. About Big Data and its challenges and benefits in manufacturing. *Database systems journal*, 4(3):10-19.

Oakland, J.S. 2012. Statistical process control. Florence, Kentucky: Routledge.

Opresnik, D. & Taisch, M. 2015. The value of big data in servitization. *International journal of production economics*, 165(Jul):174-184.

Otero, M., Pastor, A., Portela, J., Viguera, J. & Huerta, M. 2012. Standard methodology for establishing the "state of the art" based on Six Sigma. *American Institute of Physics Conference Proceedings*, 1431(1):933-938.

Pande, P. & Holpp, L. 2002. What is six sigma? New York: McGraw-Hill.

Parminter, T., Botha, N. & Small, B. 2003. Appreciating the influence of our own and others' world views upon extension strategies. http://www.regional.org.au/au/apen/2003/refereed/095parmintertg.htm Date of access: 3 Jun. 2015.

Perrons, R.K. & Jensen, J.W. 2015. Data as an asset: what the oil and gas sector can learn from other industries about Big Data. *Energy policy*, 81(Jun):117-121.

Peterka, P. 2008a. The Six Sigma method and design of experiments. https://www.statease.com/pubs/sixsigma&DOE.pdf Date of access: 13 Sep. 2014.

Peterka, P. 2008b. The Six Sigma method and design of experiments. http://www.6sigma.us/six-sigma-articles/the-six-sigma-method-and-design-of-experiments/ Date of access: 16 Sept. 2014.

Petre, R. 2013. Data mining solutions for the business environment. *Database systems journal*, 4(4):21-29.

Phadke, M.S. 1989. Quality engineering using robust design. New Jersey: Prentice Hall.

Purohit, N., Purohit, S. & Purohit, R. 2012. Data mining, applications and knowledge discovery. *International journal of advanced computer research*, 2(6):458-462.

Rajkumar, S. & Ramesh, S. 2014. Impact of implementing Six-Sigma on the workplace efficiency in service industries in Karnataka. *Asia Pacific journal of research*, 1(13):65-79.

Render, B., Stair, M.R. & Hanna, M.E. 2012. Quantitative analysis for management. 11th ed. Harlow: Pearson Prentice Hall.

Render, B., Stair, M.R., Hale, T.S. & Hanna, M.E. 2015. Quantitative analysis for management. 12th ed. Harlow: Pearson Prentice Hall.

Ryan, T.P. 1989. Statistical methods for quality improvement. New York: John Wiley.

Santana, L. 2015. Statistical analyses using the computer packages SAS and R. 4th ed. Potchefstroom: ANDCORK Publishers.

Saravanana, S., Mahadevanb, M., Suratkara, P. & Gijoc, E.V. 2012. Efficiency improvement on the multicrystalline silicon wafer through Six Sigma methodology. *International journal of sustainable energy*, 31(3):143-153.

Sharman, N.K., Cudney, E.A., Ragsdell, K.M. & Paryani, K. 2007. Quality loss function: a common methodology for three cases. *Journal of industrial and systems engineering*, 1(3):218-234.

Sowmya, Y., Ratna, M. & Bindu, C. 2015. A review on Big Data mining, distributed programming frameworks and privacy preserving data mining techniques. *International journal of advanced research in computer science*, 6(1):121-126.

Stat Trek. 2014. Teach yourself statistics. http://stattrek.com/statistics/charts/scatterplot.aspx Date of access: 24 Jul. 2014.

Steiner, S. & MacKay, J. 2000. An overview of the Shainin system for quality improvement. https://uwaterloo.ca/business-and-industrial-statistics-research-group/sites/ca.business-and-industrial-statistics-research-group/files/uploads/files/rr-06-03.pdf Date of access: 13 Sep. 2016.

Sundararajan, K. 2010. Design of experiments: a primer. http://www.isixsigma.com/tools-templates/design-of-experiments-doe/design-experiments-%E2%90% 93-primer/ Date of access: 20 Sep. 2014.

Surange, V.G. 2015. Implementation of Six Sigma to reduce cost of quality: a case study of automobile sector. *Journal of failure analysis and prevention*, 15(2):282-294.

Swan, M. 2015. Philosophy of Big Data: expanding the human-data relation with Big Data science services. Paper presented at the IEEE First International Conference on Big Data computing service and applications. www.melanieswan.com/documents/Philosophy_of_Big_Data_SWAN.pdf Date of access: 20 Sep. 2014.

Taylor, B.W. 2013. Introduction to management science. 11th ed. Harlow: Pearson Prentice Hall.

Tembhurkar, M.P., Tugnayat, R.M. & Nagdive, A.S. 2014. Overview on data mining schemes to design Business Intelligence framework for mobile technology. *International journal of advanced research in computer science*, 5(8):128-133.

Thota, L.S. & Rao, A.A. 2013. Overview of empirical data mining research. *International journal of advanced research in computer science*, 4(10):49-55.

Turban, E. & Aronson, J.E. 2001. Decision support systems and intelligent systems. Harlow: Pearson Prentice Hall.

Tyagi, A.K., Priya, R. & Rajeswari, A. 2015. Mining Big Data to predicting future. *Journal of engineering research and applications*, 5(3):14-21.

Tyagi, P., Tiwari, G. & Singh, A. 2014. Six Sigma approach to reduce the TE/FE defects in optical disc (DVD). *International journal of application or innovation in engineering and management*, 3(12):138-144.

Van Blerk, W.H. 2006. Data-mining in a marketing environment. Vanderbijlpark: NWU. (Dissertation - Msc).

Vercellis, C. 2009. Business Intelligence: data mining and optimization for decision making. New York: John Wiley.

Wagner, T. 2016. Applied business statistics: methods and excel-based applications. 4th ed. Cape Town: Juta.

Washio, T. 2007. Applications eligible for data mining. A*dvanced engineering informatics,* 21(3):241-242.

Weber, S. 2013. Mining Big Data: cover story. *NACD directorship*, 39(5):18-22.

Weisburg, G.S. 1985. Applied linear regression. New York: John Wiley.

Wheeler, D.J. & Chambers, D.S. 1986. Understanding statistical process control. Knoxville: Keith Press.

Wong, J.Y. & Chung, P.H. 2007. Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques. *Journal of air transport management,* 13(6):362-370.

Wright, P. 1998. Knowledge discovery in databases: tools and techniques. *Crossroads*, 5(2): 23-26.

Xie, L. & Kruger, U. 2012. Advances in statistical monitoring of complex multivariate processes: with applications in industrial process control. Hoboken: John Wiley.

Yu, C. & Shan, J. 2014. The application of web data mining technology in e-commerce. *Advanced materials research*, 1044-1045(Oct):1503-1506.

Yu, L. & Zhang, Z.F. 2014. The framework of product definition based on data mining techniques. *Applied mechanics and materials*, 687-691(Nov):4874-4877.

Zhong, R.Y., Huang, G.Q., Lan, S., Dai, Q.Y., Xu, C. & Zhang, T. 2015. A Big Data approach for logistics trajectory discovery from RFID-enabled production data. *International journal of production economics*, 165(Jul):260-272.

Zhou, H., Li, R.Q. & Yu, Y. 2014. Investigation of the datamation of manufacturing industrial chain in the Big Data era. *Applied mechanics and materials*, 670-671(Oct):1629-1632.

APPENDIX 1

PORTION OF ORIGINAL DATABASE (30 RECORDS)

Date	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11	Var 12	Var 13	Var 14	Var 15	Var 16	Var 17	Var 18	Var 19	Var 20	Var 21	Var 22	Var 23	Var 24	Var 25	Var 26	Var 27	Var 28	Var 29	Var 30	Var 31	Var 32	Var 33	Var 34	Var 35
2003/04/13 05:02	96	34	1	2	2849	01&03	05:00	05:01	70	05:01	163	1	160	05:01	05:01	15	12.99	0	05:02	110	110	109	107	114	107	0	0	0	0	0	0	0	0	0	0
2003/04/13 05:15	96	35	3	4	2849	05&07	05:03	04:02	70	05:01	163	2	158	04:02	05:01	16	15.99	0	05:01	108	110	109	107	114	107	6.1	2.86	5.19	4.58	4.58	2937	54	16	108	26
2003/04/13 05:28	96	36	5	6	2849	09&11	05:00	05:00	70	05:01	162	3	158	05:00	05:02	17	16	0	05:01	108	110	109	107	114	107	5.92	2.86	4.02	3.94	3.94	2937	54	17	106	26
2003/04/13 05:41	96	37	7	8	2849	02&04	04:02	05:03	70	05:01	163	1	161	05:03	05:01	16	12.99	0	05:02	110	110	109	107	114	107	5.97	2.86	4.36	4.48	4.48	2937	47	14	107	26
2003/04/13 06:21	96	38	9	10	2849	06&08	05:01	05:02	70	04:02	162	2	164	05:02	05:01	16	15.99	0	05:01	110	110	109	107	114	107	5.89	2.86	5.1	4.46	4.46	2937	47	15	104	26
2003/04/13 06:34	96	39	11	12	2849	10&12	05:01	70:00	3900	05:03	164	3	161	05:00	05:01	15	16	0	05:03	110	110	109	107	114	107	5.92	2.86	4.57	4.27	4.27	2941	50	14	102	26
2003/04/13 06:48	96	40	13	14	2849	01&03	05:02	05:01	70	05:01	163	1	162	05:01	05:01	16	12.99	0	05:00	110	110	109	107	114	107	5.8	2.86	4.68	4.95	4.95	2941	47	14	112	26
2003/04/13 06:59	96	41	15	16	2849	05&07	05:01	05:00	70	05:01	164	2	161	05:00	05:02	15	15.99	0	05:01	112	110	109	107	114	107	5.98	2.86	5.29	3.7	3.7	2941	47	15	111	27
2003/04/13 07:12	96	42	17	18	2849	09&11	04:02	05:03	70	05:01	162	3	160	05:03	05:01	15	16	0	05:02	111	110	109	107	114	107	5.87	2.86	3.79	3.88	3.88	2941	50	15	113	27
2003/04/13 07:26	96	43	20	19	2849	02&04	05:01	05:00	70	05:01	163	1	161	05:00	05:00	16	12.99	0	05:01	110	110	109	107	114	107	5.92	2.86	4.39	4.01	4.01	2941	47	14	106	26
2003/04/13 07:53	96	44	21	22	2849	06&08	05:00	05:02	70	04:02	163	2	163	05:02	05:01	17	15.99	0	05:01	107	110	109	107	114	107	5.89	2.86	4.53	3.99	3.99	2941	47	14	110	26
2003/04/13 08:22	96	45	23	24	2849	10&12	05:03	70:00	3900	05:03	165	3	159	05:03	05:02	17	16	0	05:01	107	110	109	107	114	107	5.86	2.86	4.16	3.81	3.81	2941	47	15	94	26
2003/04/13 08:37	96	46	25	26	2849	01&03	05:00	05:01	70	05:01	163	1	161	05:01	05:01	18	15.99	0	05:00	107	110	109	107	114	107	5.87	2.86	4.04	3.83	3.83	2941	50	15	107	26
2003/04/13 08:51	96	47	27	28	2849	05&07	05:01	05:03	70	05:01	164	2	159	05:03	04:02	19	15.99	0	05:03	96	110	109	107	114	107	5.98	2.86	4.13	3.26	3.26	2941	50	14	106	26
2003/04/13 09:03	96	48	29	30	2849	09&11	05:02	05:01	70	05:01	162	3	159	05:01	05:00	17	16	0	05:01	107	110	109	107	114	107	6.02	2.86	3.36	3.57	3.57	2941	50	15	107	26
2003/04/13 09:17	96	49	31	32	2849	02&04	05:02	05:00	70	05:01	163	1	160	05:00	05:03	17	15.99	0	05:01	108	110	109	107	114	107	5.99	2.86	3.78	3.29	3.29	2941	50	15	107	26
2003/04/13 09:30	96	50	33	34	2849	06&08	05:01	05:03	70	04:02	162	2	161	05:03	05:00	16	15.99	0	05:02	111	110	109	107	114	107	6.17	2.86	3.45	3.04	3.04	2941	50	15	107	26
2003/04/13 09:45	96	51	35	36	2849	10&12	05:01	70:00	3900	05:03	163	3	158	05:01	05:01	17	16	0	05:02	108	110	109	107	114	107	6.06	2.86	3.16	3.24	3.24	2941	47	15	111	26
2003/04/13 09:59	96	52	37	38	2849	01&03	05:03	04:02	70	05:01	163	1	160	04:02	05:03	18	15.99	0	04:02	108	110	109	107	114	107	5.95	2.86	3.5	3.41	3.41	2942	47	15	105	26
2003/04/13 10:11	96	53	39	40	2849	05&07	05:00	05:01	70	05:01	163	2	160	05:01	05:01	17	15.99	0	05:00	108	110	109	107	114	107	6.03	2.86	3.6	3	3	2942	47	14	108	27
2003/04/13 10:27	96	54	41	42	2849	09&11	05:00	05:02	70	05:01	162	3	158	05:02	04:02	17	16	0	05:03	108	110	109	107	114	107	6.06	2.86	3.22	3.47	3.47	2942	50	15	109	27
2003/04/13 10:43	96	55	43	44	2849	02&04	05:03	04:02	70	05:01	163	1	160	04:02	05:02	17	15.99	0	05:01	107	110	109	107	114	107	5.94	2.86	3.77	3.57	3.57	2942	47	15	105	27
2003/04/13 10:58	96	56	45	46	2849	06&08	05:00	05:01	70	04:02	162	2	157	05:01	05:02	19	15.99	0	05:00	89	110	109	107	114	107	0	0	0	0	0	0	0	0	0	0
2003/04/13 13:13	96	57	47	48	2849	10&12	05:02	70:00	3900	05:03	164	3	160	05:02	05:02	21	16	0	04:02	105	110	109	107	114	107	0	0	0	0	0	0	0	0	0	0
2003/04/13 13:22	96	58	48	49	2849	01&03	04:02	05:02	70	05:01	162	1	160	05:02	05:01	17	15.99	0	05:01	108	110	109	107	114	107	5.93	2.86	4.06	3.9	3.9	2942	47	15	104	27
2003/04/13 13:36	96	59	50	51	2849	05&07	05:01	05:01	70	05:01	163	2	156	05:01	05:01	18	15.99	0	05:02	111	110	109	107	114	107	6.18	2.86	3.76	2.95	2.95	2942	50	15	107	27
2003/04/13 13:47	96	60	52	53	2849	09&11	05:02	05:01	70	05:01	162	3	158	05:01	05:02	17	16	0	05:01	107	110	109	107	114	107	6.12	2.86	2.99	3.28	3.28	2942	50	15	110	27
2003/04/13 14:32	96	61	56	55	2849	02&04	05:02	04:02	70	05:01	163	1	161	04:02	05:02	18	15.99	0	05:01	108	110	109	107	114	107	5.97	2.86	3.57	3.49	3.49	2942	47	16	106	27
2003/04/13 14:55	96	62	55	59	2849	06&08	05:01	05:03	70	04:02	162	2	162	05:03	05:01	17	15.99	0	05:02	107	110	109	107	114	107	6	2.86	3.7	3.31	3.31	2942	50	16	106	27

APPENDIX 2

REDUCED DATABASE TO 22 VARIABLES (130 RECORDS)

	CODES	var1	var2	var3	var4	var5	var6	var7	var8	var9	var10	var11	var12	var13	var14	var15	var16	var17	var18	var19	var20	var21	var22
		Cat / Block	Input	Cat / Block	Cat / Block	Input	Input	Input	Input	Input	Input	Input	Output	Set points	Set points	Set points	Set points	Set points	Set points	Output	Output	Output	Output
Record Number	Date	Mixer pair	Actual extrusion length	Mixer pair	Cooler number	Mix discharge temp	Cool begin temp	Actual cool time	Actual dump temp	Actual tamp pressure	Actual extrusion rate	Actual extrusion speed	Actual plug temp	Actual mix time	Die top temp	Actual mud cyclinder temp	Actual mud extrution temp	Actual ram temp	Ram set temp	Actual max extrusion pressure	Actual extrusionpressure electrode 1	Actual extrusionpressure electrode 2	Actual average extrusionpressure
1	2003/07/07 19:54	09&11	2003	E	3	162	153	18	108	5.97	54	16	104	70	110	109	107	114	107	2.48	2.49	2.46	2.475
2	2003/07/07 20:07	02&04	2003	B	1	163	154	19	107	5.96	50	16	106	70	110	109	107	114	107	2.58	2.47	2.35	2.41
3	2003/07/07 21:43	09&11	2003	E	3	162	156	18	108	5.99	50	16	103	70	110	109	107	114	107	2.61	2.57	2.35	2.46
4	2003/07/07 21:57	02&04	2003	B	1	164	154	18	108	6.07	47	16	105	70	110	109	107	114	107	2.56	2.47	2.06	2.265
5	2003/07/07 22:37	10&12	2971	F	3	165	155	20	107	6	50	15	104	70	110	109	107	114	107	2.76	2.74	2.63	2.685
6	2003/07/07 22:51	01&03	2971	A	1	163	154	19	107	6.03	50	16	104	70	110	109	107	114	107	2.86	2.79	2.34	2.565
7	2003/07/07 23:07	05&07	2971	C	2	165	156	19	108	5.99	50	16	107	70	110	109	107	114	107	2.63	2.54	2.35	2.445
8	2003/07/07 23:33	02&04	2971	B	1	164	155	20	107	5.99	50	16	106	70	110	109	107	114	107	2.55	2.44	2.34	2.39
9	2003/07/07 23:46	06&08	2971	D	2	164	160	20	108	5.89	50	16	107	70	110	109	107	114	107	2.68	2.61	2.79	2.7
10	2003/07/08 00:00	10&12	2971	F	3	165	156	19	107	5.95	50	16	104	70	110	109	107	114	107	3.03	3.01	2.73	2.87
11	2003/07/08 00:25	01&03	2971	A	1	163	155	18	108	5.98	50	16	106	70	110	109	107	114	107	2.97	2.88	2.32	2.6
12	2003/07/08 00:42	05&07	2971	C	2	165	158	19	106	5.99	50	16	108	70	110	109	107	114	107	2.61	2.59	2.43	2.51
13	2003/07/08 00:57	09&11	2971	E	3	163	156	20	107	5.99	50	16	103	70	110	109	107	114	107	2.73	2.68	2.55	2.615
14	2003/07/08 01:37	06&08	2971	D	2	164	160	19	109	5.96	50	16	109	70	110	109	107	114	107	2.87	2.81	2.97	2.89
15	2003/07/08 01:54	10&12	2971	F	3	165	150	21	104	5.9	50	16	105	70	110	109	107	114	107	3.19	3.18	3.02	3.1
16	2003/07/08 02:06	01&03	2971	A	1	163	155	19	108	6.06	50	16	102	70	110	109	107	114	107	3.16	3.03	2.29	2.66
17	2003/07/08 02:22	05&07	2971	C	2	164	158	19	107	5.94	50	16	107	70	110	109	107	114	107	2.58	2.47	2.5	2.485
18	2003/07/08 02:34	09&11	2971	E	3	162	156	19	108	6.08	47	16	104	70	110	109	107	114	107	2.68	2.64	2.42	2.53
19	2003/07/08 02:47	02&04	2971	B	1	164	155	19	107	6.05	50	16	104	70	110	109	107	114	107	2.55	2.5	2.25	2.375
20	2003/07/08 03:04	06&08	2971	D	2	164	159	20	107	6.02	50	16	110	70	110	109	107	114	107	2.65	2.55	2.66	2.605
21	2003/07/08 03:16	10&12	2971	F	3	164	156	19	107	5.99	50	16	102	70	110	109	107	114	107	2.89	2.86	2.63	2.745
22	2003/07/08 03:31	01&03	2971	A	1	162	155	19	107	6	50	16	105	70	110	109	107	114	107	2.8	2.77	2.44	2.605
23	2003/07/08 03:47	05&07	2971	C	2	164	158	20	107	5.96	50	16	108	70	110	109	107	114	107	2.83	2.75	2.6	2.675
24	2003/07/08 04:06	09&11	2971	E	3	162	157	19	108	6.03	50	16	104	70	110	109	107	114	107	2.8	2.76	2.69	2.725
25	2003/07/08 04:21	02&04	2971	B	1	164	155	19	108	6.05	47	16	105	70	110	109	107	114	107	2.82	2.78	2.39	2.585
26	2003/07/08 04:34	06&08	2971	D	2	164	159	20	105	5.97	50	16	107	70	110	109	107	114	107	2.66	2.57	2.68	2.625
27	2003/07/08 05:01	01&03	2971	A	1	163	156	19	106	6.02	50	16	103	70	110	109	107	114	107	2.73	2.71	2.38	2.545
28	2003/07/08 05:16	05&07	2971	C	2	164	160	20	104	5.96	50	16	108	70	110	109	107	114	107	2.56	2.55	2.6	2.575
29	2003/07/08 05:43	09&11	2971	E	3	162	140	17	108	6.04	50	16	103	70	110	109	107	114	107	2.84	2.77	2.52	2.645
30	2003/07/08 06:00	02&04	2971	B	1	164	154	18	108	6.1	50	16	106	70	110	109	107	114	107	2.64	2.57	2.21	2.39
31	2003/07/08 06:13	06&08	2971	D	2	164	159	19	108	6.02	50	16	108	70	110	109	107	114	107	2.5	2.43	2.49	2.46
32	2003/07/08 06:26	10&12	2971	F	3	165	154	18	108	6.02	47	16	100	70	110	109	107	114	107	2.77	2.7	2.48	2.59
33	2003/07/08 06:41	01&03	2971	A	1	163	154	18	108	6.05	50	16	105	70	110	109	107	114	107	2.71	2.64	2.34	2.49
34	2003/07/08 07:25	05&07	2971	C	2	164	157	18	108	6	50	16	109	70	110	109	107	114	107	2.69	2.67	2.57	2.62
35	2003/07/08 07:47	09&11	2971	E	3	163	154	18	108	6.04	50	16	105	70	110	109	107	114	107	2.58	2.56	2.32	2.44
36	2003/07/08 08:06	02&04	2971	B	1	165	154	18	108	6	50	16	108	70	110	109	107	114	107	2.43	2.39	2.1	2.245
37	2003/07/08 08:23	06&08	2971	D	2	165	160	19	108	6	50	16	108	70	110	109	107	114	107	2.36	2.3	2.54	2.42
38	2003/07/08 10:56	01&03	2971	A	1	163	152	18	108	6.04	50	16	105	70	110	109	107	114	107	3.12	2.83	2.11	2.47
39	2003/07/08 11:10	05&07	2971	C	2	165	156	18	108	5.97	50	16	107	70	110	109	107	114	107	2.25	2.19	2.22	2.205
40	2003/07/08 11:48	02&04	2971	B	1	164	153	19	108	6	47	15	106	70	110	109	107	114	107	5.23	4.97	2.87	3.92
41	2003/07/08 12:38	10&12	2976	F	3	165	154	19	107	5.96	47	15	101	70	110	109	107	114	107	2.81	2.81	2.63	2.72
42	2003/07/08 12:55	01&03	2976	A	1	162	154	19	108	5.97	50	16	105	70	110	109	107	114	107	2.94	2.83	2.57	2.7
43	2003/07/08 13:09	05&07	2975	C	2	165	157	20	106	5.93	50	16	107	70	110	109	107	114	107	2.77	2.78	2.55	2.665
44	2003/07/08 13:25	09&11	2975	E	3	163	155	19	107	5.93	47	16	102	70	110	109	107	114	107	2.66	2.63	2.29	2.46
45	2003/07/08 13:44	02&04	2975	B	1	164	154	19	108	6.01	47	16	103	70	110	109	107	114	107	2.4	2.36	2.12	2.24
46	2003/07/08 13:58	06&08	2975	D	2	165	158	19	108	5.96	47	16	108	70	110	109	107	114	107	2.39	2.39	2.47	2.43
47	2003/07/08 14:17	10&12	2976	F	3	165	158	19	107	5.92	47	16	106	70	110	109	107	114	107	2.75	2.74	2.81	2.775
48	2003/07/08 14:31	01&03	2976	A	1	164	156	20	107	5.95	50	16	103	70	110	109	107	114	107	3	2.98	2.4	2.69
49	2003/07/08 14:48	05&07	2976	C	2	165	159	20	109	5.91	50	16	107	70	110	109	107	114	107	2.73	2.69	2.8	2.745
50	2003/07/08 15:44	02&04	2976	B	1	165	156	21	106	5.96	47	15	105	70	110	109	107	114	107	3.34	3.25	2.89	3.07
51	2003/07/08 15:56	06&08	2976	D	2	164	159	20	108	5.92	47	16	104	70	110	109	107	114	107	3.04	3	2.79	2.895
52	2003/07/08 16:14	10&12	2976	F	3	165	145	20	107	5.91	50	15	105	73	110	109	107	114	107	2.91	2.88	2.54	2.71
53	2003/07/08 16:25	01&03	2976	A	1	164	156	21	107	5.98	50	16	108	70	110	109	107	114	107	2.73	2.68	2.48	2.58
54	2003/07/08 16:39	05&07	2976	C	2	165	159	20	107	5.87	50	16	105	70	110	109	107	114	107	2.85	2.72	2.53	2.625
55	2003/07/08 16:58	09&11	2976	E	3	163	159	22	107	5.93	47	16	104	70	110	109	107	114	107	2.8	2.74	2.72	2.73
56	2003/07/08 17:09	02&04	2976	B	1	165	157	20	108	5.97	47	16	101	70	110	109	107	114	107	2.89	2.86	2.51	2.685
57	2003/07/08 17:23	06&08	2976	D	2	164	159	20	107	5.96	47	16	106	70	110	109	107	114	107	2.8	2.77	2.66	2.715
58	2003/07/08 17:40	10&12	2976	F	3	165	159	22	107	5.89	47	15	107	70	110	109	107	114	107	2.94	2.89	2.94	2.915

APPENDIX 3

MEDIAN SPLIT DATABASE

Mix discharge temp	Mix discharge temp	Cool begin temp	Cool begin temp	Actual cool tim	Actual cool time	Actual dump temp	Actual dump temp	Actual tamp pressur	Actual tamp pressur	Actual extrusion rate	Actual extrusion rate	Actual extrusion speed	Actual extrusion speed
162	-	153	-	18	-	108	+	5.97	+	54	+	16	+
163	-	154	-	19	+	107	-	5.96	+	50	+	16	+
162	-	156	-	18	-	108	+	5.99	+	50	+	16	+
164	+	154	-	18	-	108	+	6.07	+	47	-	16	+
165	+	155	-	20	+	107	-	6	+	50	+	15	-
163	-	154	-	19	+	107	-	6.03	+	50	+	16	+
165	+	156	-	19	+	108	+	5.99	+	50	+	16	+
164	+	155	-	20	+	107	-	5.99	+	50	+	16	+
164	+	160	+	20	+	108	+	5.89	+	50	+	16	+
165	+	156	-	19	+	107	-	5.95	+	50	+	16	+
163	-	155	-	18	-	108	+	5.98	+	50	+	16	+
165	+	158	+	19	+	106	-	5.99	+	50	+	16	+
163	-	156	-	20	+	107	-	5.99	+	50	+	16	+
164	+	160	+	19	+	109	+	5.96	+	50	+	16	+
165	+	150	-	21	+	104	-	5.9	+	50	+	16	+
163	-	155	-	19	+	108	+	6.06	+	50	+	16	+
164	+	158	+	19	+	107	-	5.94	+	50	+	16	+
162	-	156	-	19	+	108	+	6.08	+	47	-	16	+
164	+	155	-	19	+	107	-	6.05	+	50	+	16	+
164	+	159	+	20	+	107	-	6.02	+	50	+	16	+
164	+	156	-	19	+	107	-	5.99	+	50	+	16	+
162	-	155	-	19	+	107	-	6	+	50	+	16	+
164	+	158	+	20	+	107	-	5.96	+	50	+	16	+
162	-	157	+	19	+	108	+	6.03	+	50	+	16	+
164	+	155	-	19	+	108	+	6.05	+	47	-	16	+
164	+	159	+	20	+	105	-	5.97	+	50	+	16	+
163	-	156	-	19	+	106	-	6.02	+	50	+	16	+
164	+	160	+	20	+	104	-	5.96	+	50	+	16	+
162	-	140	-	17	-	108	+	6.04	+	50	+	16	+
164	+	154	-	18	-	108	+	6.1	+	50	+	16	+
164	+	159	+	19	+	108	+	6.02	+	50	+	16	+
165	+	154	-	18	-	108	+	6.02	+	47	-	16	+
163	-	154	-	18	-	108	+	6.05	+	50	+	16	+
164	+	157	+	18	-	108	+	6	+	50	+	16	+
163	-	154	-	18	-	108	+	6.04	+	50	+	16	+
165	+	154	-	18	-	108	+	6	+	50	+	16	+
165	+	160	+	19	+	108	+	6	+	50	+	16	+
163	-	152	-	18	-	108	+	6.04	+	50	+	16	+
165	+	156	-	18	-	108	+	5.97	+	50	+	16	+
164	+	153	-	19	+	108	+	6	+	47	-	15	-
165	+	154	-	19	+	107	-	5.96	+	47	-	15	-
162	-	154	-	19	+	108	+	5.97	+	50	+	16	+
165	+	157	+	20	+	106	-	5.93	+	50	+	16	+
163	-	155	-	19	+	107	-	5.93	+	47	-	16	+
164	+	154	-	19	+	108	+	6.01	+	47	-	16	+
165	+	158	+	19	+	108	+	5.96	+	47	-	16	+
165	+	158	+	19	+	107	-	5.92	+	47	-	16	+
164	+	156	-	20	+	107	-	5.95	+	50	+	16	+
165	+	159	+	20	+	109	+	5.91	+	50	+	16	+
165	+	156	-	21	+	106	-	5.96	+	47	-	15	-
164	+	159	+	20	+	108	+	5.92	+	47	-	16	+
165	+	145	-	20	+	107	-	5.91	+	50	+	15	-
164	+	156	-	21	+	107	-	5.98	+	50	+	16	+
165	+	159	+	20	+	107	-	5.87	-	50	+	16	+
163	-	159	+	22	+	107	-	5.93	+	47	-	16	+
165	+	157	+	20	+	108	+	5.97	+	47	-	16	+
164	+	159	+	20	+	107	-	5.96	+	47	-	16	+
165	+	159	+	22	+	107	-	5.89	+	47	-	15	-
164	+	158	+	21	+	107	-	5.94	+	47	-	15	-
164	+	160	+	21	+	107	-	5.88	-	47	-	16	+
163	-	161	+	21	+	107	-	5.88	-	47	-	16	+
165	+	157	+	20	+	107	-	5.94	+	47	-	16	+
164	+	159	+	20	+	109	+	5.9	+	47	-	16	+
165	+	157	+	20	+	108	+	5.96	+	50	+	16	+
164	+	158	+	20	+	108	+	5.99	+	50	+	16	+
164	+	159	+	21	+	107	-	5.9	+	50	+	16	+
164	+	156	-	20	+	107	-	6	+	47	-	16	+
164	+	142	-	19	+	109	+	5.9	+	47	-	16	+
164	+	137	-	20	+	110	+	5.99	+	50	+	16	+
164	+	157	+	16	-	112	+	6.01	+	50	+	16	+
163	-	156	-	19	+	108	+	5.98	+	47	-	15	-
164	+	155	-	18	-	107	-	5.99	+	50	+	16	+
165	+	157	+	19	+	107	-	5.94	+	47	-	16	+
166	+	160	+	19	+	107	-	5.89	+	47	-	16	+
164	+	157	+	19	+	107	-	5.94	+	50	+	16	+
165	+	159	+	20	+	107	-	5.93	+	47	-	16	+
163	-	157	+	19	+	107	-	5.98	+	50	+	15	-

APPENDIX 4

PREDICTION ERROR EXAMPLES

Dependent variable - Ave extrusion																
Average																
1 st Period DOE Model		1 st Period DOE regression		1 st Period regression		Signal noise ratio	2 nd Period DOE Base		2 nd Period DOE regression		2 nd Period regression		Signal noise ratio	2nd period prediciton DOE regression		2nd period prediciton normal regression	
Model rank		A - B		A - C													
A		B		C			D		E		F			G		H	
Run number	Model	Run number	Model	Run number	Model	S/N	Run number	Test	Run number	Test	Run number	Test	S/N	Run number	Test	Run number	Test
1	2.962	1	2.907	1	2.962	19.506	1	3.327	1	3.256	1	3.169	21.160	1	2.907	1	2.878
2	2.924	2	3.079	2	3.003	20.481	2	3.487	2	3.504	2	3.406	17.857	2	3.079	2	3.000
3	3.160	3	3.030	3	3.081	24.902	3	3.661	3	3.704	3	3.533	18.372	3	3.030	3	**3.011**
4	2.908	4	2.921	4	2.926	23.780	4	3.316	4	3.357	4	3.195	19.632	4	2.921	4	**2.863**
5	3.044	5	2.981	5	3.011	20.796	5	3.332	5	3.339	5	3.271	17.517	5	2.981	5	2.985
6	2.983	6	2.992	6	2.958	20.316	6	3.398	6	3.327	6	3.254	23.434	6	2.992	6	2.919
7	2.965	7	2.943	7	2.963	21.712	7	3.648	7	3.527	7	3.451	18.932	7	2.943	7	2.913
8	3.020	8	2.995	8	3.026	23.627	8	3.518	8	3.440	8	3.451	18.904	8	2.995	8	2.997
9	3.080	9	3.111	9	3.069	20.739	9	3.498	9	3.560	9	3.426	19.246	9	3.111	9	**3.040**
10	3.004	10	3.003	10	2.990	20.933	10	3.172	10	3.213	10	3.144	19.811	10	3.003	10	2.929
11	2.970	11	2.953	11	2.972	22.527	11	3.351	11	3.413	11	3.227	18.823	11	2.953	11	2.901
12	3.184	12	3.125	12	3.056	20.341	12	3.804	12	3.662	12	3.549	17.541	12	3.125	12	**3.062**
13	2.940	13	3.024	13	3.015	18.801	13	3.274	13	3.383	13	3.159	18.773	13	3.024	13	2.981
14	3.088	14	3.077	14	3.084	21.086	14	3.195	14	3.297	14	3.301	17.300	14	3.077	14	3.056
15	2.964	15	3.027	15	3.036	20.662	15	3.510	15	3.497	15	3.483	18.122	15	3.027	15	**3.046**
16	3.015	16	3.038	16	3.012	22.676	16	3.474	16	3.485	16	3.400	21.736	16	3.038	16	2.961
17	2.810	17	2.786	17	2.779	18.426	17	3.034	17	3.098	17	3.152	19.012	17	2.786	17	**2.698**
18	2.843	18	2.838	18	2.794	21.628	18	3.021	18	3.012	18	3.128	20.797	18	2.838	18	**2.760**
19	2.654	19	2.789	19	2.773	22.388	19	3.196	19	3.212	19	3.257	18.027	19	2.789	19	**2.785**
20	2.814	20	2.800	20	2.728	26.946	20	3.165	20	3.200	20	3.183	18.970	20	2.800	20	2.647
21	2.862	21	2.860	21	2.886	19.716	21	3.317	21	3.182	21	3.149	19.692	21	2.860	21	2.629
22	2.720	22	2.751	22	2.763	20.646	22	2.806	22	2.835	22	2.950	19.235	22	2.751	22	2.624
23	2.726	23	2.702	23	2.775	21.426	23	2.937	23	3.035	23	3.119	19.298	23	2.702	23	**2.661**
24	2.857	24	2.874	24	2.849	21.806	24	3.148	24	3.283	24	3.325	20.040	24	2.874	24	2.679
25	2.896	25	2.870	25	2.894	23.970	25	3.133	25	3.068	25	3.113	20.459	25	2.870	25	2.767
26	2.940	26	2.881	26	2.852	23.384	26	3.077	26	3.056	26	3.114	17.516	26	2.881	26	2.677
27	2.769	27	2.832	27	2.814	26.197	27	3.317	27	3.256	27	3.213	18.965	27	2.832	27	**2.677**
28	2.893	28	2.884	28	2.856	22.841	28	3.183	28	3.170	28	3.251	19.483	28	2.884	28	2.793
29	2.786	29	2.783	29	2.783	21.132	29	2.955	29	2.891	29	3.014	18.550	29	2.783	29	**2.742**
30	3.019	30	2.955	30	2.922	18.981	30	3.134	30	3.140	30	3.233	17.515	30	2.955	30	2.854
31	2.917	31	2.906	31	2.865	21.980	31	3.273	31	3.340	31	3.365	19.413	31	2.906	31	2.735
32	2.802	32	2.797	32	2.801	22.699	32	3.074	32	2.992	32	3.155	18.352	32	2.797	32	**2.728**

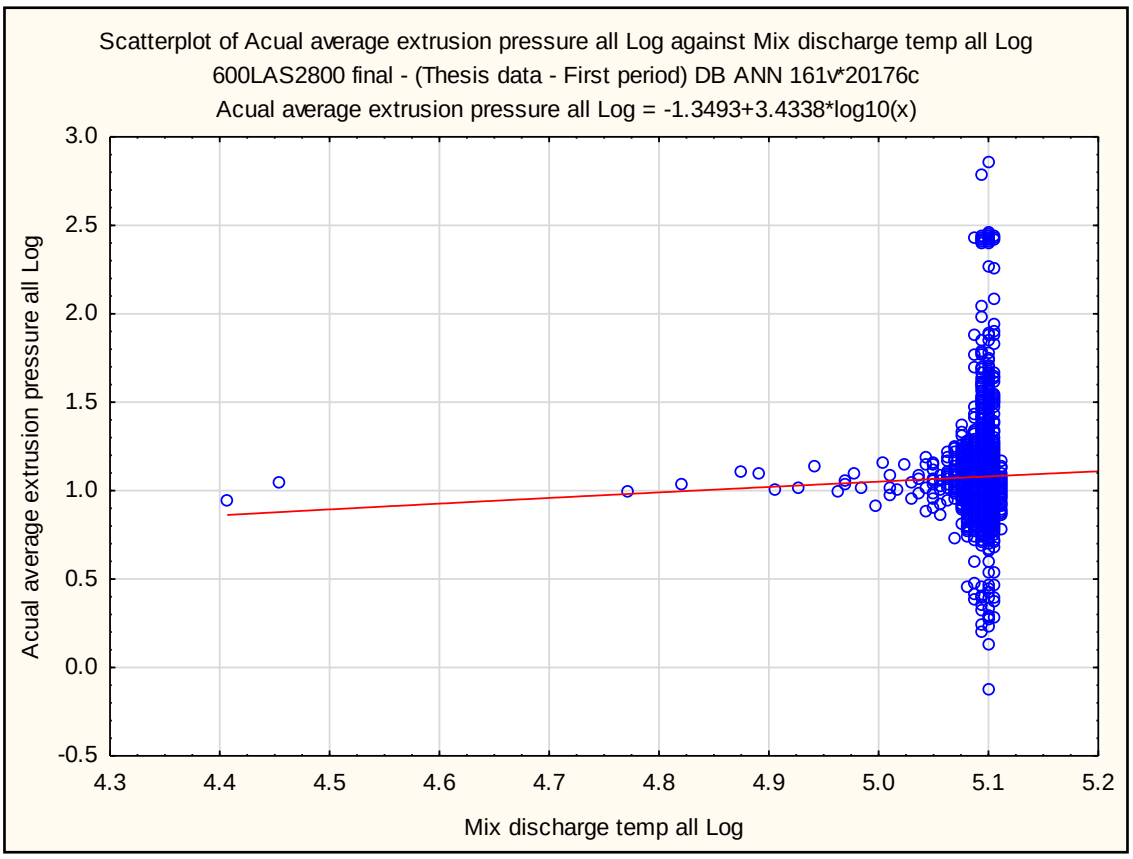
APPENDIX 5

EXAMPLE OF DOE RUN SUMMARY

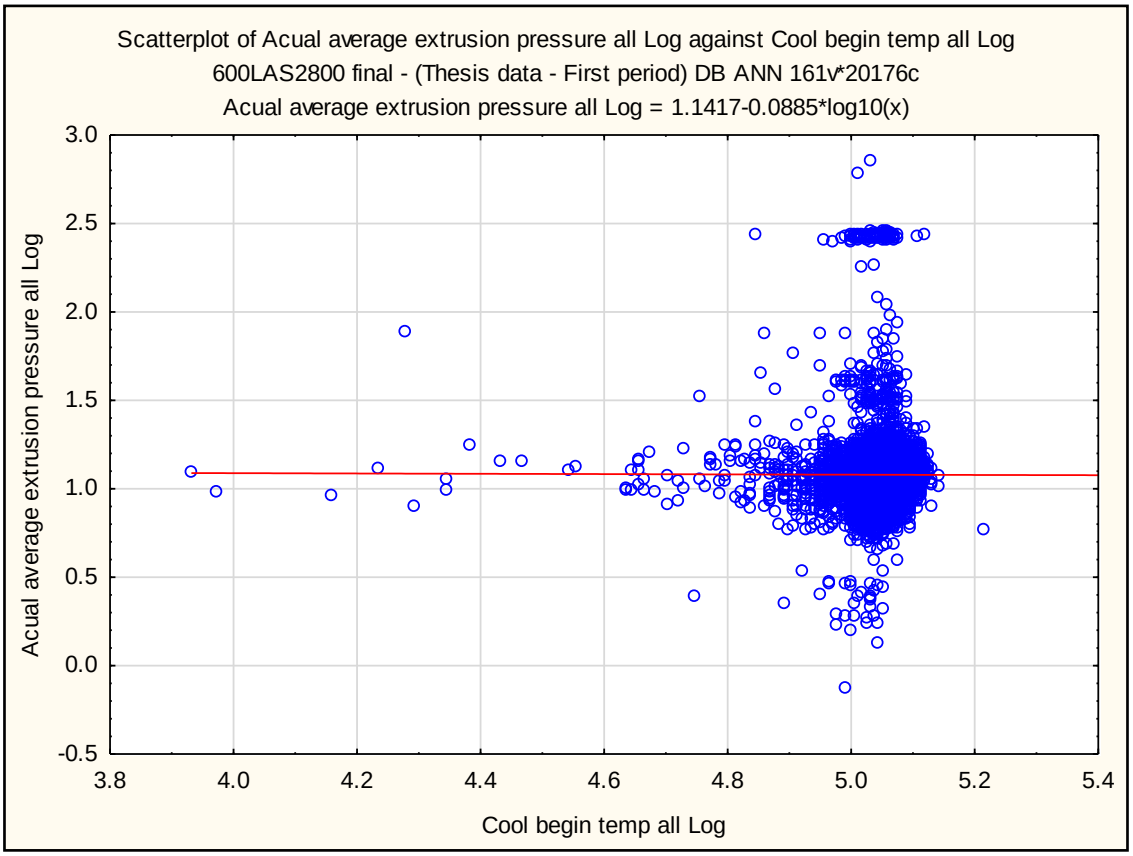
Y (3) - (Actual average extrusionpressure)																																
Ave Average	DOE reg Max ave ave	Error	Abs Error	Error sq	% error	Norm Reg ave Max DOE ave ave	Error	Abs Error	Error sq	% error	Ave Median	DOE reg Max ave Median	Error	Abs Error	Error sq	% error	Norm Reg ave Max ave Median	Error	Abs Error	Error sq	% error	Ave Stdev	DOE reg Max ave stdev	Error	Abs Error	Error sq	% error	Norm Reg ave stdev DOE ave ave	Error	Abs Error	Error sq	% error	n
2.962	2.907	0.054	0.054	0.003	0.018	2.962	-0.001	0.001	0.000	0.000	2.890	2.928	-0.038	0.038	0.001	0.013	2.962	-0.072	0.072	0.005	0.025	0.314	0.266	0.048	0.048	0.002	0.153	0.050	0.264	0.264	0.070	0.842	109.000
2.924	3.079	-0.156	0.156	0.024	0.053	3.003	-0.079	0.079	0.006	0.027	2.945	3.081	-0.136	0.136	0.018	0.046	2.988	-0.043	0.043	0.002	0.015	0.277	0.282	-0.005	0.005	0.000	0.020	0.075	0.201	0.201	0.041	0.728	19.000
3.160	3.030	0.129	0.129	0.017	0.041	3.081	0.079	0.079	0.006	0.025	3.260	3.045	0.215	0.215	0.046	0.066	3.151	0.109	0.109	0.012	0.033	0.180	0.231	-0.051	0.051	0.003	0.283	0.095	0.085	0.085	0.007	0.473	10.000
2.908	2.921	-0.013	0.013	0.000	0.005	2.926	-0.019	0.019	0.000	0.006	2.870	2.929	-0.059	0.059	0.004	0.021	2.928	-0.058	0.058	0.003	0.020	0.188	0.193	-0.004	0.004	0.000	0.023	0.056	0.132	0.132	0.017	0.701	41.000
3.044	2.981	0.063	0.063	0.004	0.021	3.011	0.033	0.033	0.001	0.011	3.045	2.981	0.064	0.064	0.004	0.021	3.005	0.040	0.040	0.002	0.013	0.278	0.307	-0.029	0.029	0.001	0.106	0.065	0.213	0.213	0.045	0.766	49.000
2.983	2.992	-0.010	0.010	0.000	0.003	2.958	0.025	0.025	0.001	0.008	3.015	2.992	0.023	0.023	0.001	0.008	2.957	0.058	0.058	0.003	0.019	0.288	0.301	-0.013	0.013	0.000	0.046	0.071	0.216	0.216	0.047	0.752	155.000
2.965	2.943	0.023	0.023	0.001	0.008	2.963	0.003	0.003	0.000	0.001	2.985	2.956	0.029	0.029	0.001	0.010	2.956	0.029	0.029	0.001	0.010	0.243	0.249	-0.006	0.006	0.000	0.023	0.079	0.164	0.164	0.027	0.675	65.000
3.020	2.995	0.024	0.024	0.001	0.008	3.026	-0.006	0.006	0.000	0.002	3.005	2.982	0.023	0.023	0.001	0.008	3.026	-0.021	0.021	0.000	0.007	0.199	0.234	-0.035	0.035	0.001	0.177	0.079	0.120	0.120	0.014	0.604	83.000
3.080	3.111	-0.031	0.031	0.001	0.010	3.069	0.011	0.011	0.000	0.004	2.993	3.096	-0.103	0.103	0.011	0.035	3.064	-0.071	0.071	0.005	0.024	0.283	0.296	-0.014	0.014	0.000	0.048	0.077	0.206	0.206	0.042	0.728	66.000
3.004	3.003	0.001	0.001	0.000	0.000	2.990	0.014	0.014	0.000	0.005	3.013	2.980	0.032	0.032	0.001	0.011	2.986	0.026	0.026	0.001	0.009	0.270	0.259	0.011	0.011	0.000	0.042	0.074	0.195	0.195	0.038	0.724	184.000
2.970	2.953	0.017	0.017	0.000	0.006	2.972	-0.002	0.002	0.000	0.001	2.975	2.944	0.031	0.031	0.001	0.010	2.957	0.018	0.018	0.000	0.006	0.222	0.207	0.015	0.015	0.000	0.068	0.059	0.163	0.163	0.027	0.734	25.000
3.184	3.125	0.058	0.058	0.003	0.018	3.056	0.128	0.128	0.016	0.040	3.085	3.098	-0.013	0.013	0.000	0.004	3.063	0.022	0.022	0.000	0.007	0.306	0.223	0.083	0.083	0.007	0.270	0.069	0.237	0.237	0.056	0.775	17.000
2.940	3.024	-0.084	0.084	0.007	0.029	3.015	-0.075	0.075	0.006	0.026	2.975	3.007	-0.032	0.032	0.001	0.011	3.031	-0.056	0.056	0.003	0.019	0.338	0.315	0.022	0.022	0.001	0.066	0.079	0.259	0.259	0.067	0.766	29.000
3.088	3.077	0.011	0.011	0.000	0.004	3.084	0.003	0.003	0.000	0.001	3.020	3.034	-0.014	0.014	0.000	0.005	3.083	-0.063	0.063	0.004	0.021	0.272	0.300	-0.027	0.027	0.001	0.101	0.074	0.198	0.198	0.039	0.728	27.000
2.964	3.027	-0.063	0.063	0.004	0.021	3.036	-0.072	0.072	0.005	0.024	2.965	2.997	-0.032	0.032	0.001	0.011	3.024	-0.059	0.059	0.003	0.020	0.275	0.248	0.026	0.026	0.001	0.096	0.062	0.213	0.213	0.045	0.775	34.000
3.015	3.038	-0.024	0.024	0.001	0.008	3.012	0.003	0.003	0.000	0.001	3.018	3.008	0.009	0.009	0.000	0.003	3.001	0.017	0.017	0.000	0.005	0.222	0.242	-0.020	0.020	0.000	0.092	0.084	0.138	0.138	0.019	0.622	104.000
2.810	2.786	0.024	0.024	0.001	0.008	2.779	0.031	0.031	0.001	0.011	2.925	2.828	0.097	0.097	0.009	0.033	2.776	0.149	0.149	0.022	0.051	0.337	0.251	0.086	0.086	0.007	0.256	0.075	0.262	0.262	0.069	0.778	21.000
2.843	2.838	0.004	0.004	0.000	0.002	2.794	0.049	0.049	0.002	0.017	2.930	2.855	0.075	0.075	0.006	0.026	2.780	0.150	0.150	0.023	0.051	0.236	0.236	0.000	0.000	0.000	0.001	0.075	0.161	0.161	0.026	0.682	48.000
2.654	2.789	-0.135	0.135	0.018	0.051	2.773	-0.118	0.118	0.014	0.045	2.670	2.819	-0.149	0.149	0.022	0.056	2.795	-0.125	0.125	0.016	0.047	0.202	0.184	0.018	0.018	0.000	0.088	0.115	0.087	0.087	0.008	0.430	26.000
2.814	2.800	0.014	0.014	0.000	0.005	2.728	0.086	0.086	0.007	0.031	2.800	2.830	-0.030	0.030	0.001	0.011	2.684	0.116	0.116	0.014	0.042	0.126	0.178	-0.051	0.051	0.003	0.404	0.065	0.062	0.062	0.004	0.490	23.000
2.862	2.860	0.002	0.002	0.000	0.001	2.886	-0.024	0.024	0.001	0.009	2.830	2.881	-0.051	0.051	0.003	0.018	2.862	-0.032	0.032	0.001	0.011	0.296	0.292	0.004	0.004	0.000	0.012	0.118	0.178	0.178	0.032	0.602	22.000
2.720	2.751	-0.031	0.031	0.001	0.011	2.763	-0.043	0.043	0.002	0.016	2.710	2.766	-0.056	0.056	0.003	0.021	2.774	-0.064	0.064	0.004	0.024	0.252	0.254	-0.002	0.002	0.000	0.007	0.078	0.175	0.175	0.031	0.692	390.000
2.726	2.702	0.024	0.024	0.001	0.009	2.775	-0.049	0.049	0.002	0.018	2.740	2.730	0.010	0.010	0.000	0.004	2.783	-0.043	0.043	0.002	0.016	0.231	0.203	0.029	0.029	0.001	0.124	0.083	0.148	0.148	0.022	0.640	219.000
2.857	2.874	-0.017	0.017	0.000	0.006	2.849	0.008	0.008	0.000	0.003	2.865	2.883	-0.018	0.018	0.000	0.006	2.870	-0.005	0.005	0.000	0.002	0.232	0.219	0.013	0.013	0.000	0.056	0.108	0.124	0.124	0.015	0.533	33.000
2.896	2.870	0.026	0.026	0.001	0.009	2.894	0.002	0.002	0.000	0.001	2.920	2.870	0.050	0.050	0.003	0.017	2.920	0.000	0.000	0.000	0.000	0.183	0.250	-0.067	0.067	0.004	0.363	0.108	0.076	0.076	0.006	0.413	51.000
2.940	2.881	0.059	0.059	0.003	0.020	2.852	0.088	0.088	0.008	0.030	3.000	2.881	0.119	0.119	0.014	0.040	2.863	0.137	0.137	0.019	0.046	0.199	0.244	-0.044	0.044	0.002	0.223	0.089	0.110	0.110	0.012	0.554	51.000
2.769	2.832	-0.063	0.063	0.004	0.023	2.814	-0.045	0.045	0.002	0.016	2.700	2.845	-0.145	0.145	0.021	0.054	2.830	-0.130	0.130	0.017	0.048	0.136	0.192	-0.056	0.056	0.003	0.415	0.044	0.092	0.092	0.008	0.678	5.000
2.893	2.884	0.009	0.009	0.000	0.003	2.856	0.037	0.037	0.001	0.013	2.925	2.872	0.053	0.053	0.003	0.018	2.885	0.040	0.040	0.002	0.014	0.209	0.177	0.032	0.032	0.001	0.152	0.086	0.122	0.122	0.015	0.586	22.000
2.786	2.783	0.003	0.003	0.000	0.001	2.783	0.004	0.004	0.000	0.001	2.775	2.781	-0.006	0.006	0.000	0.002	2.777	-0.002	0.002	0.000	0.001	0.245	0.269	-0.024	0.024	0.001	0.098	0.110	0.135	0.135	0.018	0.552	129.000
3.019	2.955	0.064	0.064	0.004	0.021	2.922	0.097	0.097	0.009	0.032	2.910	2.934	-0.024	0.024	0.001	0.008	2.917	-0.007	0.007	0.000	0.002	0.340	0.285	0.055	0.055	0.003	0.161	0.106	0.234	0.234	0.055	0.689	14.000
2.917	2.906	0.011	0.011	0.000	0.004	2.865	0.051	0.051	0.003	0.018	2.958	2.898	0.060	0.060	0.004	0.020	2.863	0.094	0.094	0.009	0.032	0.232	0.233	-0.001	0.001	0.000	0.005	0.073	0.159	0.159	0.025	0.687	16.000
2.800	2.797	0.003	0.003	0.000	0.001	2.801	0.000	0.000	0.000	0.000	2.795	2.783	0.012	0.012	0.000	0.004	2.812	-0.017	0.017	0.000	0.006	0.205	0.195	0.010	0.010	0.000	0.048	0.078	0.127	0.127	0.016	0.619	244.000

APPENDIX 7

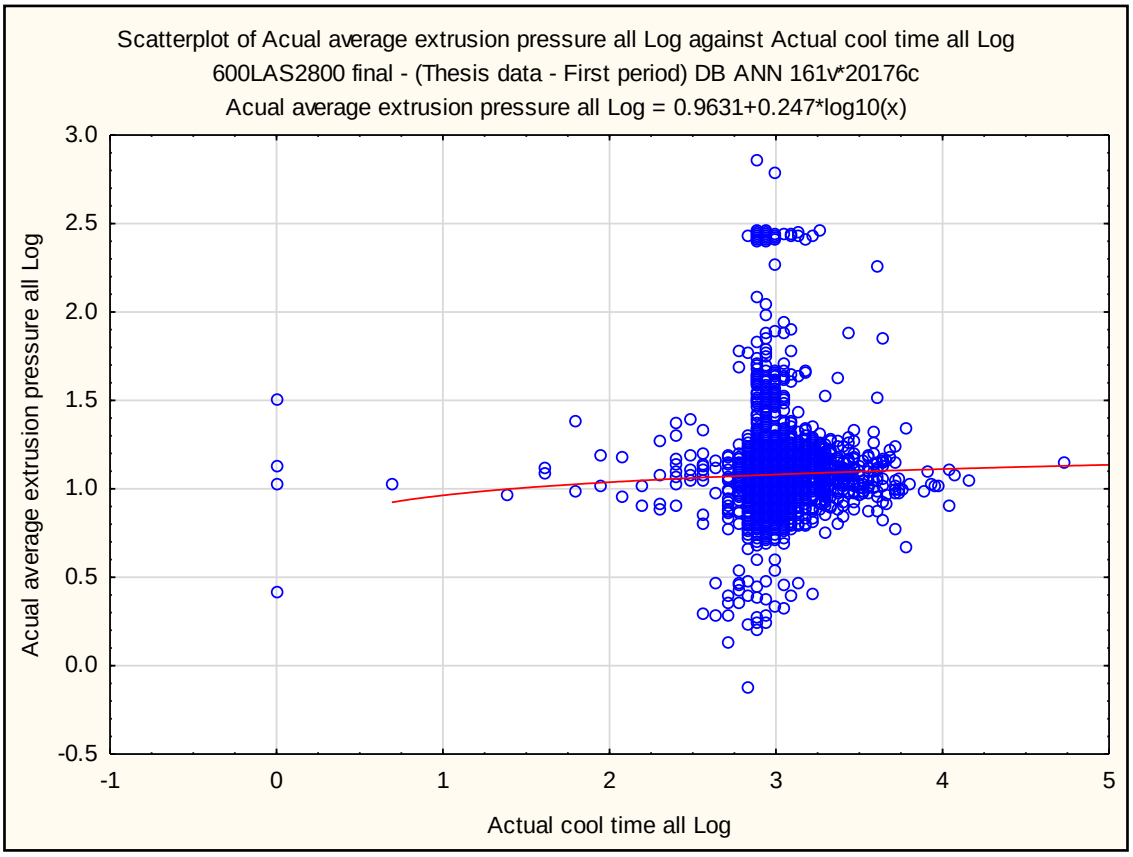
LOG TRANSFORMATIONS



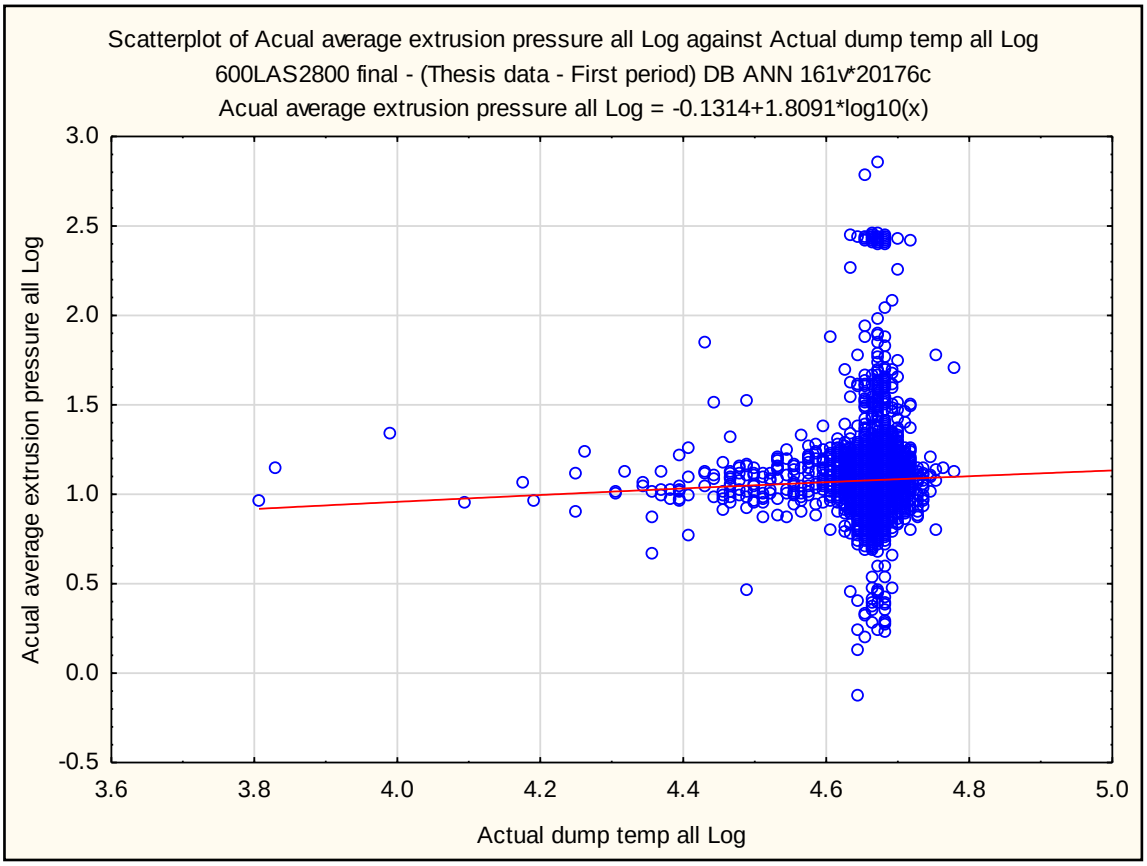
Scatter plot estimated period: (Mix discharge temperature – Raw Log transformed)



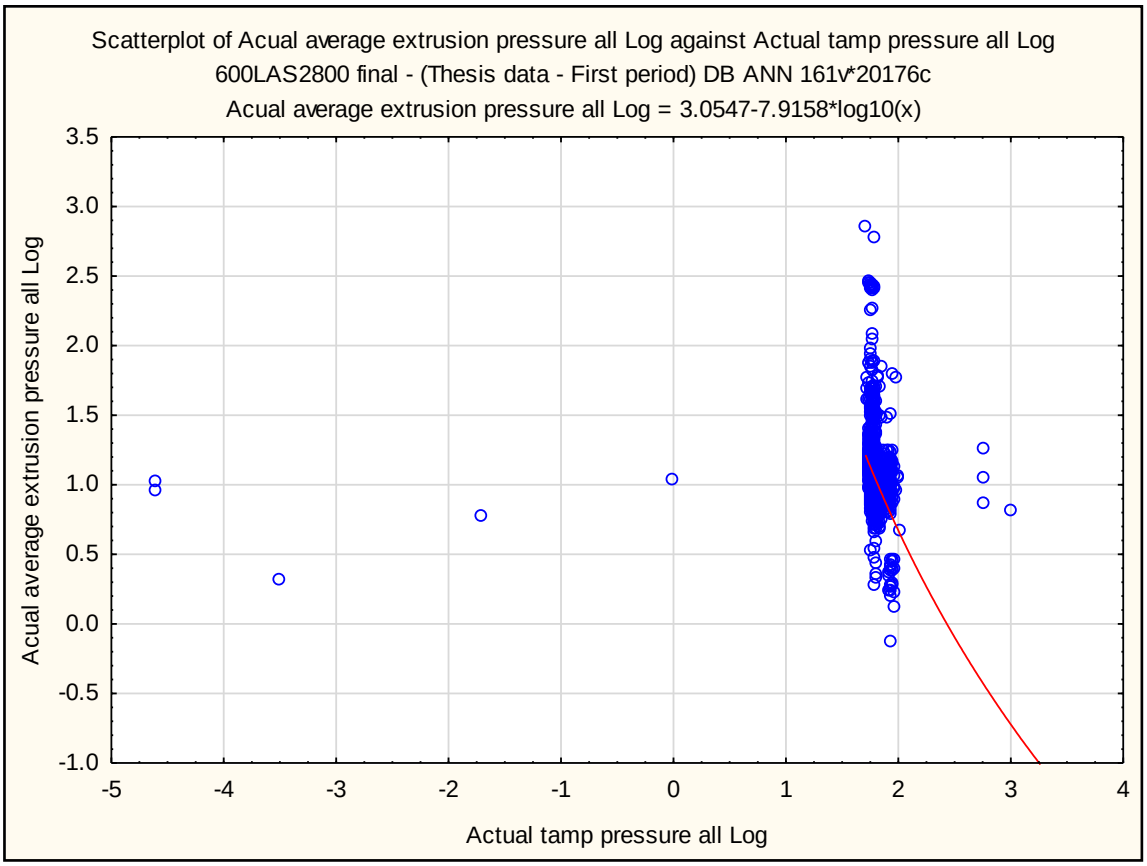
Scatter plot estimated period: (Cool begin temperature – Raw Log transformed)



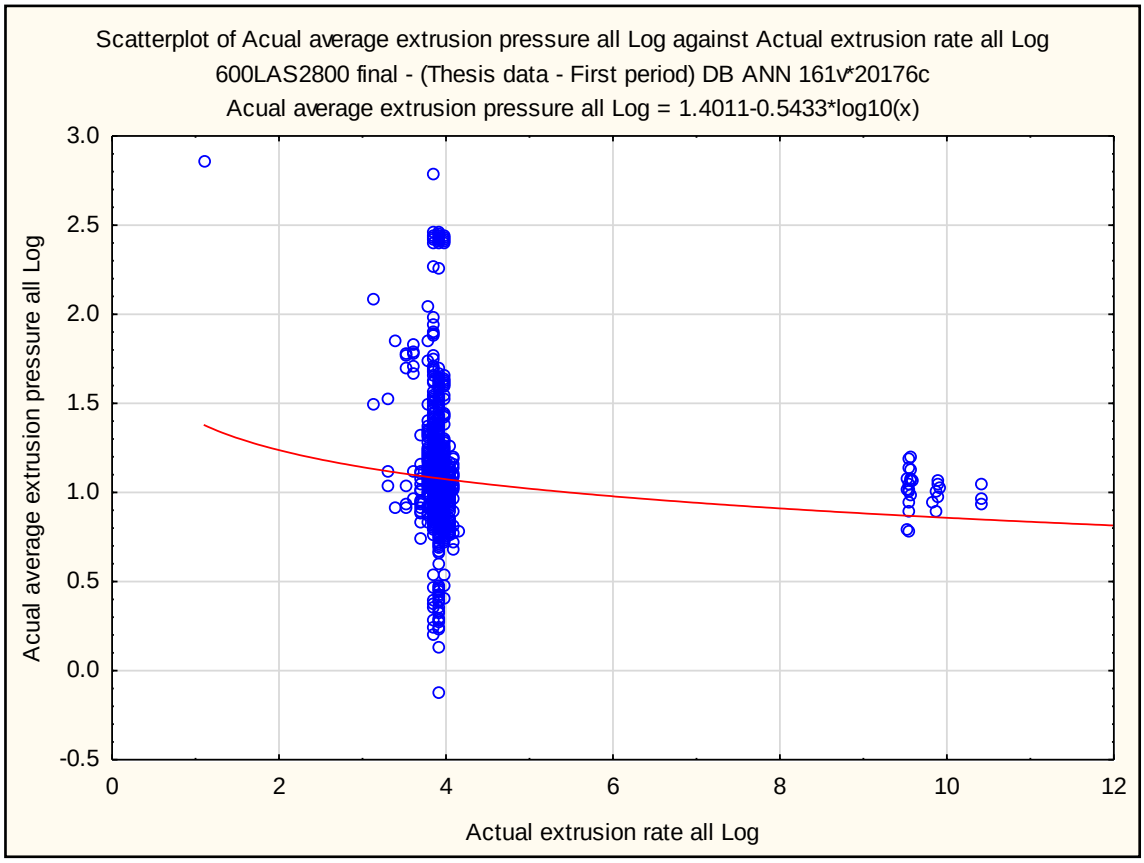
Scatter plot estimated period: (Actual cooling time – Raw Log transformed)



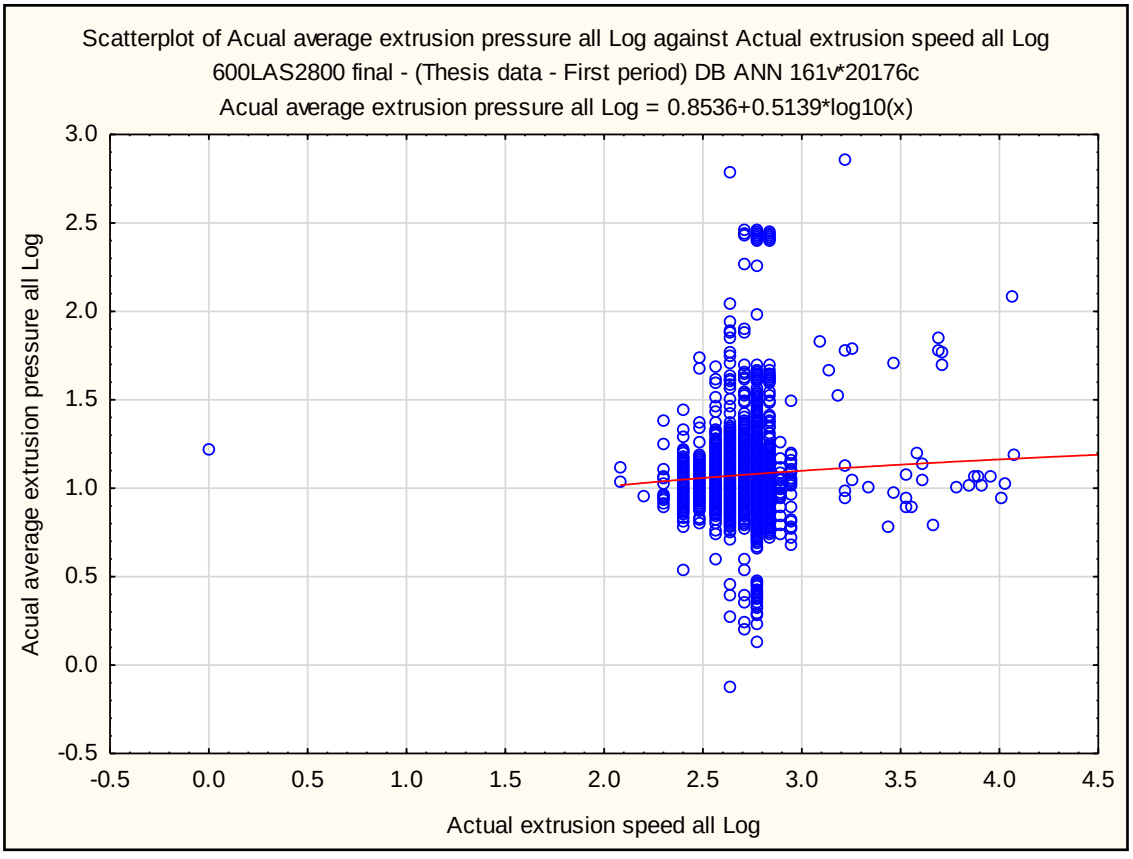
Scatter plot estimated period: (Actual dump temperature – Raw Log transformed)



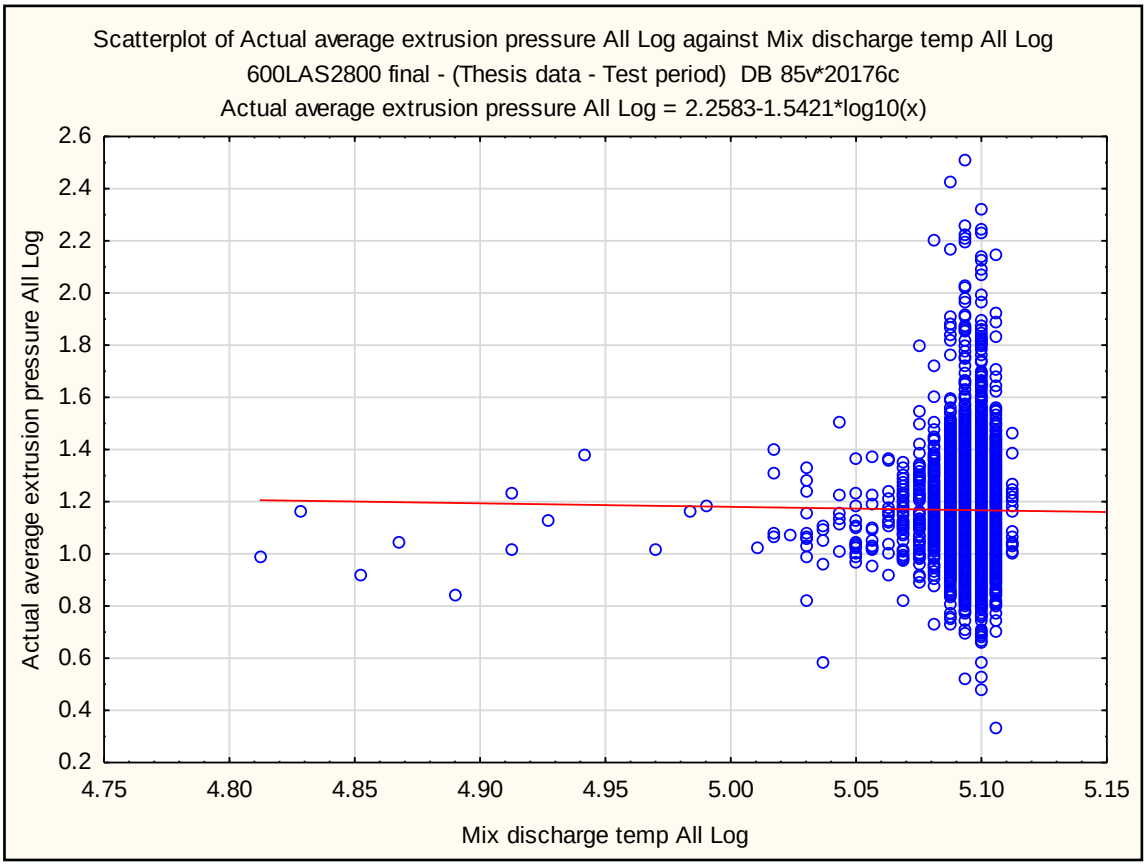
Scatter plot estimated period: (Actual tamp pressure – Raw Log transformed)



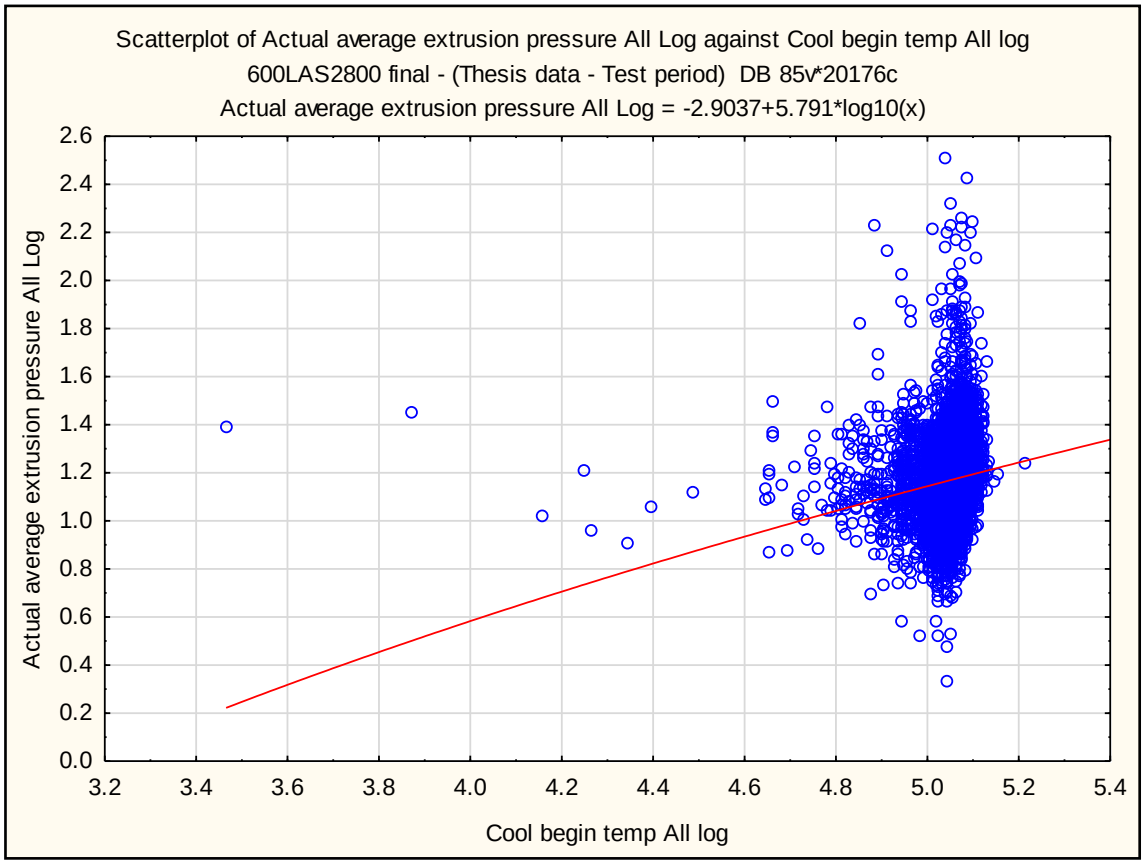
Scatter plot estimated period: (Actual extrusion rate – Raw Log transformed)



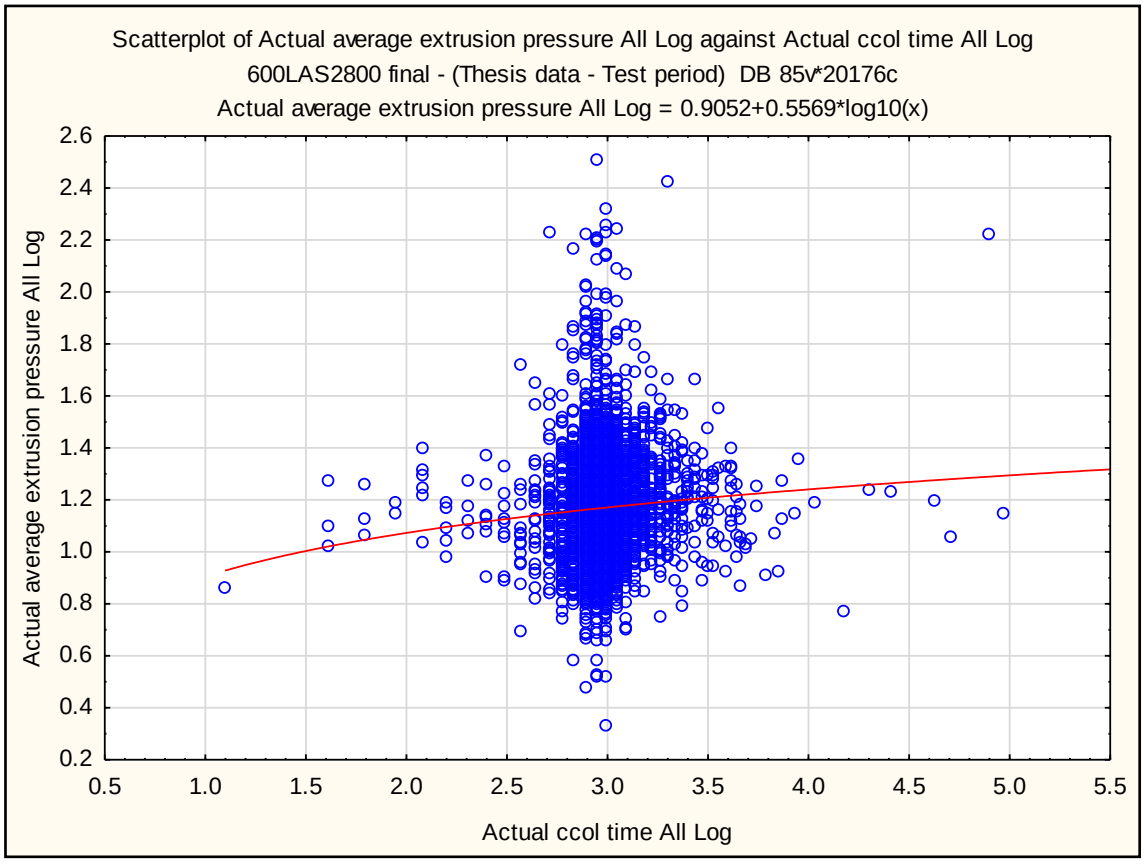
Scatter plot estimated period: (Actual extrusion speed – Raw Log transformed)



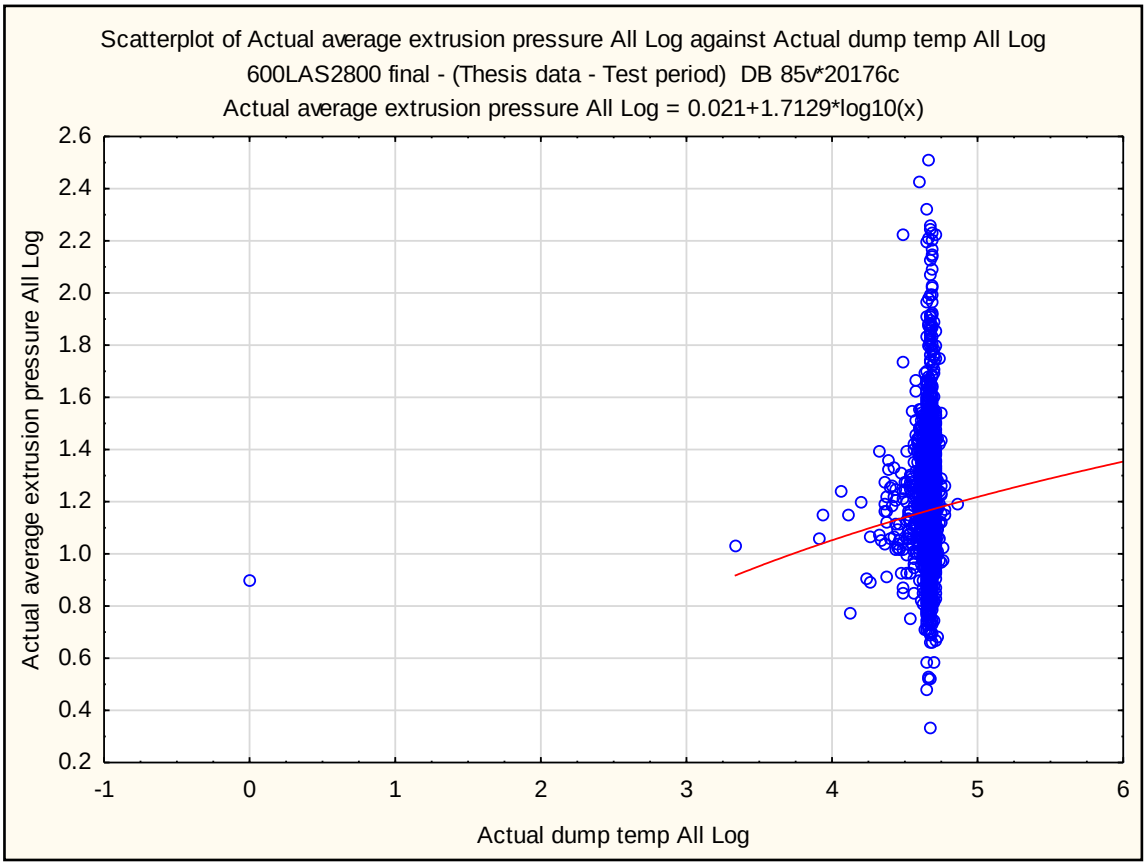
Scatter plot validation period: (Mix discharge temperature – Raw Log transformed)



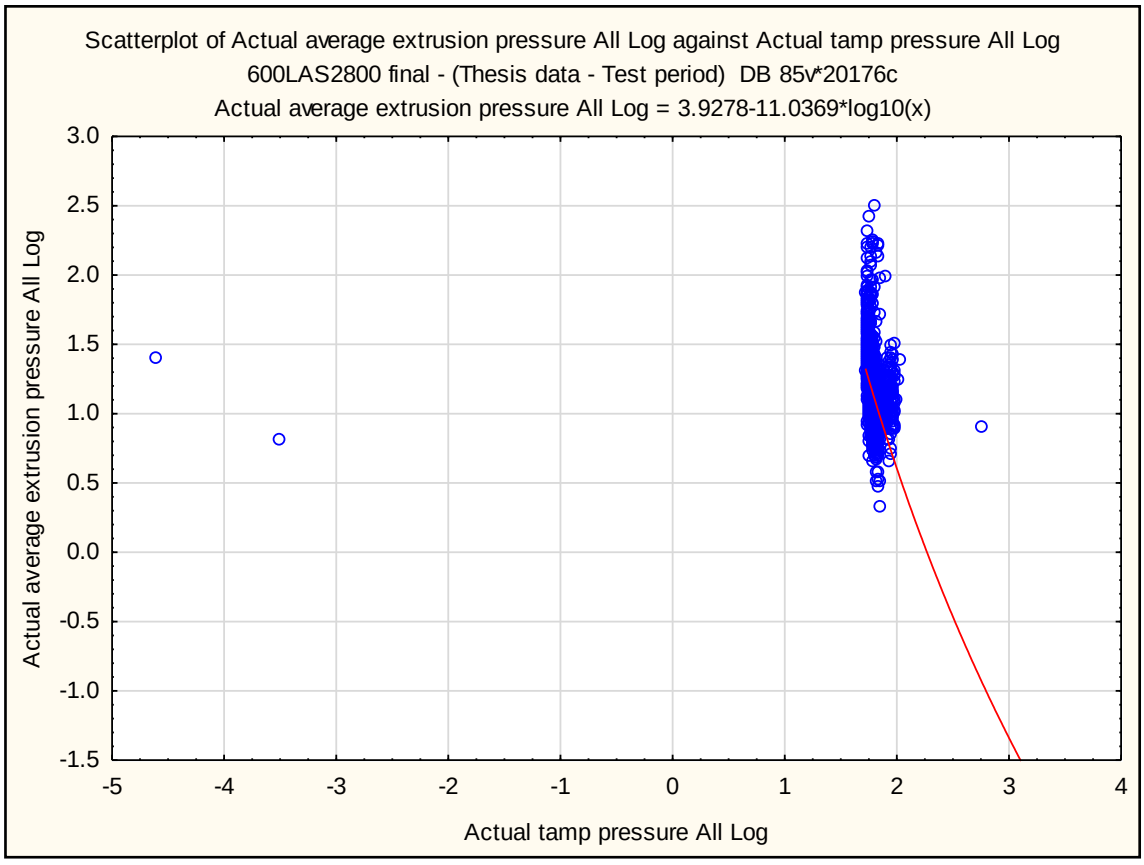
Scatter plot validation period: (Cool begin temperature – Raw Log transformed)



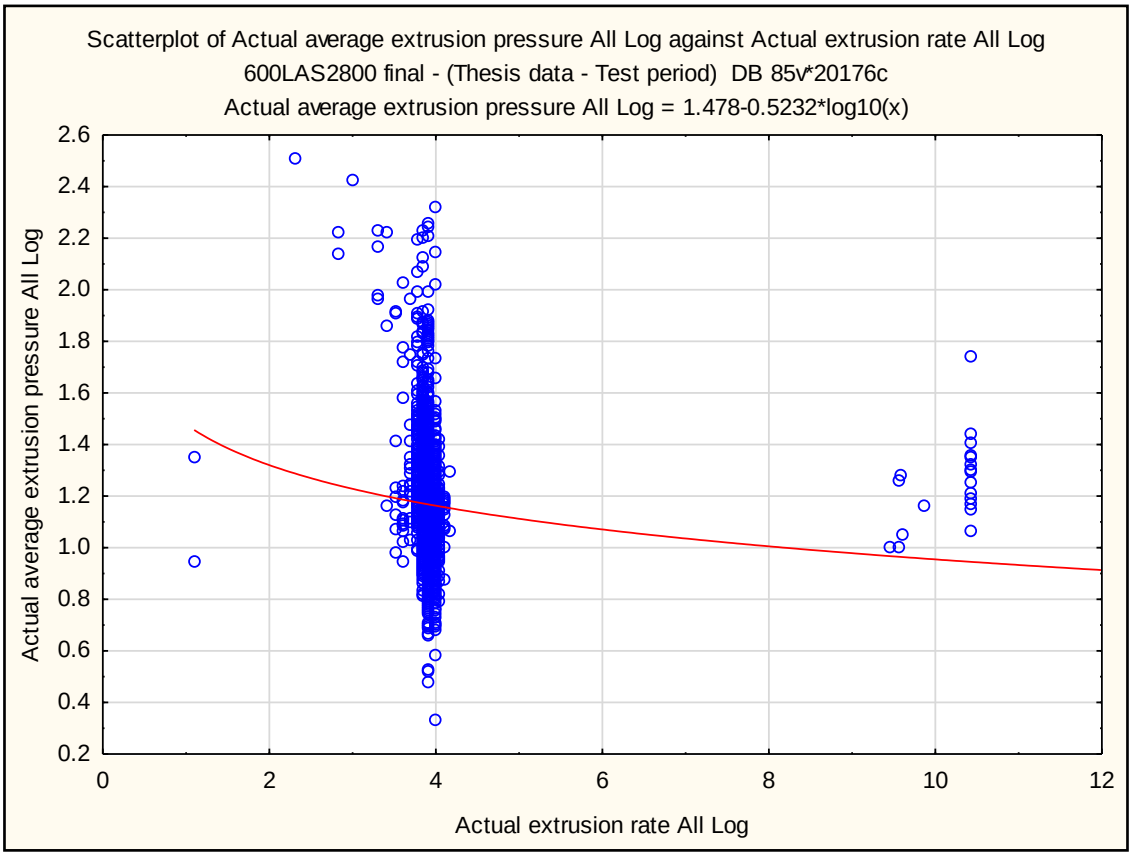
Scatter plot validation period: (Actual cooling time – Raw Log transformed)



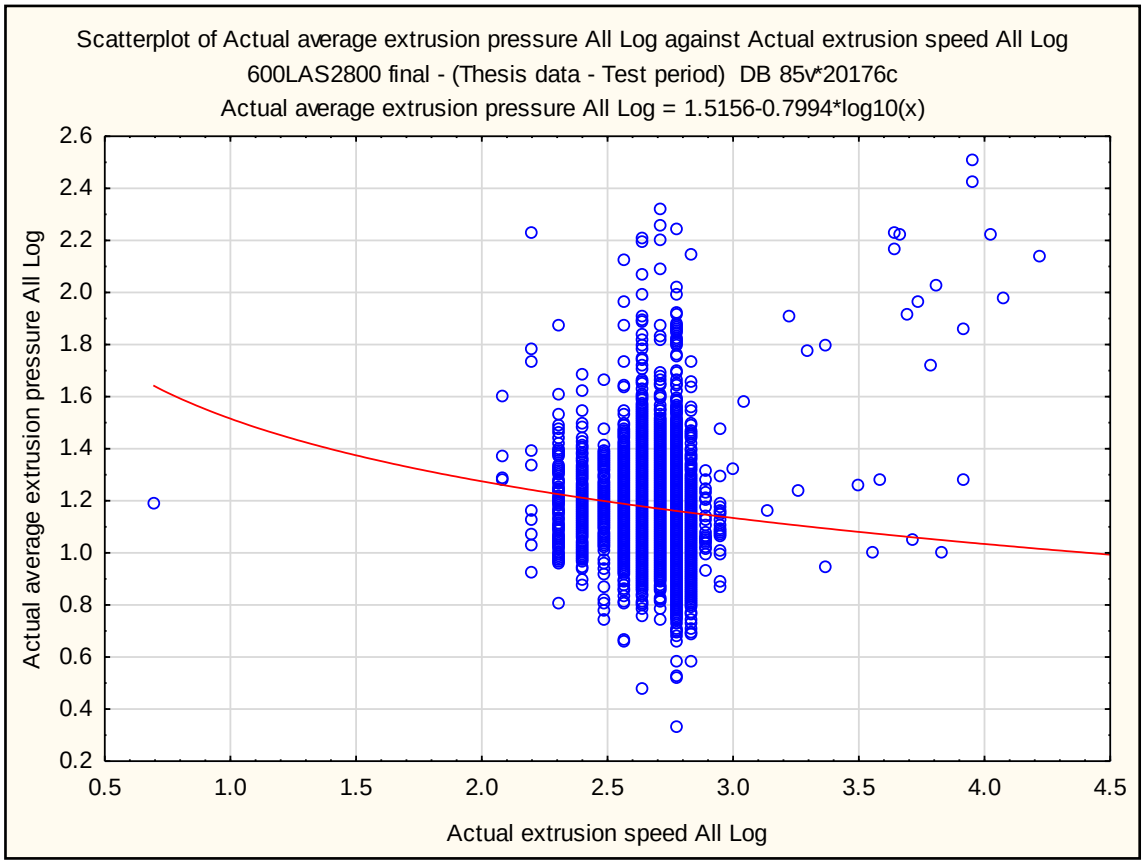
Scatter plot validation period: (Actual dump temperature – Raw Log transformed)



Scatter plot validation period: (Actual tamp pressure – Raw Log transformed)



Scatter plot validation period: (Actual extrusion rate – Raw Log transformed)



Scatter plot validation period: (Actual extrusion speed – Raw Log transformed)