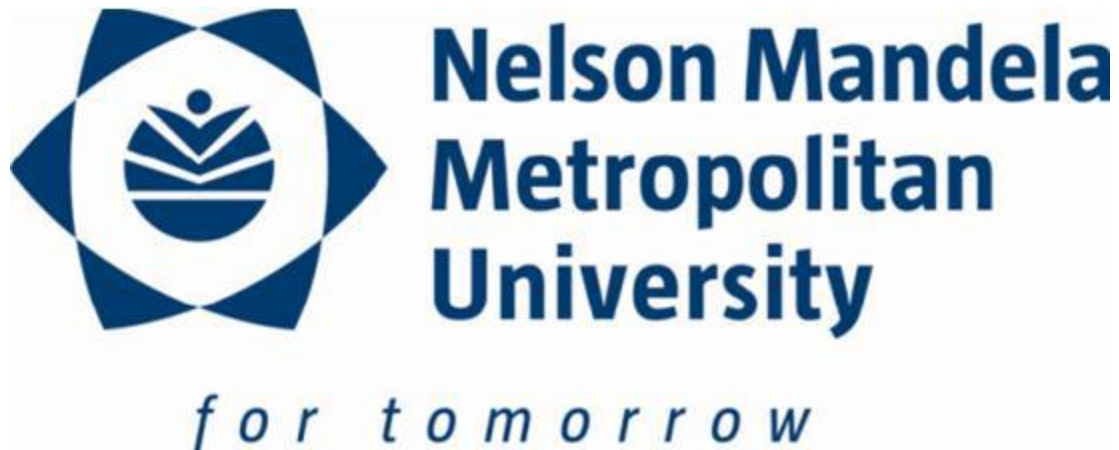


**A Framework for Grain Commodity  
Trading Decision Support in  
South Africa**

**Kayode Anthony Ayankoya**

**2016**



**A Framework for Grain Commodity Trading Decision Support in  
South Africa**

**Kayode Anthony Ayankoya**

Submitted in fulfilment of the requirements for the degree of  
Philosophiae Doctor  
in the Faculty of Science at the Nelson Mandela Metropolitan University

**Promoter:** Professor André Calitz  
**Co-Promoter:** Professor Jean Greyling

March 2016

## **Declaration**

I, Ayankoya Kayode Anthony 212400096, hereby declare that the thesis for the degree of Philosophiae Doctor is my own work and that it has not previously been submitted for assessment or completion of any postgraduate qualification to another University or for another qualification.

Ayankoya Kayode Anthony

## Summary

In several countries around the world, grain commodities are traded as assets on stock exchanges. This indicates that the market and effectively the prices of the grain commodities in such countries, are controlled by several local and international economic, political and social factors that are rapidly changing. As a result, the prices of some grain commodities are volatile and trading in such commodities are prone to price-related risks.

There are different trading strategies for minimising price-related risks and maximising profits. But empirical research suggests that making the right decision for effective grain commodities trading has been a difficult task for stakeholders due to high volatility of grain commodities prices. Studies have shown that this is more challenging among grain commodities farmers because of their lack of skills and the time to sift through and make sense of the datasets on the plethora of factors that influence the grain commodities market.

This thesis focused on providing an answer for the main research problem that grain farmers in South Africa do not take full advantage of all the available strategies for trading their grain commodities because of the complexities associated with monitoring the large datasets that influence the grain commodities market. The main objective set by this study is to design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities.

This study takes advantage of the developments in Big Data and Data Science to achieve the set objective using the Design Science Research (DSR) methodology. The prediction of future prices of grain commodities for the different trading strategies was identified as an important factor for making better decisions when trading grain commodities and the key factors that influence the prices were identified. This was followed by a critical review of the literature to determine how the concepts of Big Data and Data Science can be leveraged for an effective grain commodities trading decision support. This resulted in a proposed framework for grain commodities trading.

The proposed framework suggested an investigation of the factors that influence the prices of grain commodities as the basis for acquiring the relevant datasets. The proposed framework suggested the adoption of the Big Data approach in acquiring, preparing and integrating relevant datasets from several sources. Furthermore, it was suggested that algorithmic models for predicting grain commodities prices can be developed on top of the data layer of the proposed framework to provide real-time decision support. The proposed framework suggests the need for a carefully designed visualisation of the result and the collected data that promotes user experience. Lastly, the proposed framework included a technology consideration component to support the Big Data and Data Science approach of the framework.

To demonstrate that the proposed framework addressed the main problem of this research, datasets from several sources on trading white maize in South Africa and the factors that influence market were streamed, integrated and analysed. Backpropagation Neural Network algorithm was used for modelling the prices of white maize for spot and futures trading strategies were predicted. There are other modelling techniques such as the Box-Jenkins statistical time series analysis methodology. But, Neural Networks was identified as more suitable for time series data with complex patterns and relationships.

A demonstration system was setup to provide effective decision support by using near real-time data to provide a dynamic predictive analytics for the spot and December futures contract prices of white maize in South Africa. Comparative analysis of predictions made using the model from the proposed framework to actual data indicated a significant degree of accuracy. A further evaluation was carried out by asking experienced traders to make predictions for the spot and December futures contract prices of white maize. The result of the exercise indicated that the predictions from the developed model were much closer to the actual prices. This indicated that the proposed framework is technically capable and generally useful. It also shows that the proposed framework can be used to provide decision support about trading grain commodities to stakeholders with lesser skills, experience and resources.

The practical contribution of this thesis is that relevant datasets from several sources can be streamed into an integrated data source in real-time, which can be used as input for a real-time learning algorithmic model for predicting grain commodities prices. This will make it possible for a predictive analytics that responds to market volatility thereby providing an effective decision support for grain commodities trading. Another practical contribution of this thesis is a proposed framework that can be followed for developing a Decision Support System for trading in grain commodities. This thesis made theoretical contributions by building on the information processing theory and the decision making theory. The theoretical contribution of this thesis consists of the identification of Big Data approach, tools and techniques for eradicating uncertainty and equivocality in grain commodities trading decision making process.

## **Acknowledgement**

My profound gratitude goes to the almighty God for the making it possible for me complete this study and providing everything that I required along the way. Indeed, I could not have finished this thesis without God's grace and the wonderful people that surrounds me. I am thankful for the support that I received from all and sundry in completing this study.

I will like to thank my promoters, Professor Andre Calitz and Professor Jean Greyling for their enthusiasm, unwavering support, guidance and extraordinary leadership. I am very grateful to my promoters for believing in me and allowing me to grow, yet providing essential inputs for the success of this thesis. Moreover, I will also like to thank the academic and non-academic staff of the Department of Computing Sciences at NMMU for being there and always willing to help. My gratitude also goes to Professor Margaret Cullen for all her support and encouragement during this study.

Thanks to all my colleagues that shared the PhD laboratory with me over the period of my study, for always willing to assist and for sharing my moments of joy and frustrations. Particularly, I am thankful to Ifeoluwapo Fashoro for always making time to edit and provide constructive criticism for my work. I will also like to express my gratitude to Dr Anthony Simpson of BKB GrainCo for the professional advice and support that he offered several times during my research. My gratitude also goes to SYSPRO for their support.

Lastly, I will like to express a deep appreciation to my wife, Adeyemi Ayankoya, for the support, encouragement, understanding and sacrifices that she made for me to be able to complete my studies. To my children, Dominic, Derek and Divine, thank you for your understanding and sacrifices.

## Table of Contents

Declaration .....	i
Summary .....	ii
Acknowledgement .....	v
List of Figures .....	xi
List of Tables .....	xv
List of Abbreviations .....	xvii
Chapter 1 : Research Context.....	1
1.1 Background .....	1
1.2 Problem Statement.....	5
1.3 Thesis Statement.....	7
1.4 Research Objectives.....	7
1.5 Research Questions .....	8
1.6 Scope and Constraints .....	9
1.7 Proposed Thesis Structure .....	10
1.8 Summary .....	13
Chapter 2 : Research Design and Methodology.....	15
2.1 Introduction .....	15
2.2 Research Design .....	17
2.2.1 Research philosophy.....	18
2.2.2 Adopted philosophy for study .....	20
2.3 Research Approach .....	21
2.4 Research Strategy .....	22
2.5 Research Methodology.....	24
2.5.1 Relevance cycle .....	25
2.5.2 Rigour Cycle.....	26
2.5.3 Design cycle.....	27



2.5.4 Design science research process .....	28
2.6 Ethical Considerations .....	34
2.7 Summary .....	35
Chapter 3 : Decision Support for Grain Commodities Trading .....	37
3.1 Introduction .....	38
3.2 Decision Making in Organisations.....	41
3.2.1 Decision making theory .....	42
3.2.2 Information processing theory .....	44
3.2.3 Improving decision making.....	47
3.3 Decision Support Systems.....	49
3.3.1 Components of DSS .....	50
3.3.2 Tools of DSS .....	55
3.4 Decision Making for Grain Commodities Trading.....	57
3.4.1 Grain commodities trading strategies .....	59
3.4.2 Factors influencing grain commodities trading .....	63
3.5 Grain Commodities Trading DSS Requirement Survey .....	66
3.6 Conclusion .....	69
Chapter 4 : Big Data for Grain Commodities Trading in South Africa.....	72
4.1 Introduction .....	73
4.2 Big Data .....	75
4.2.1 Big Data techniques and technologies .....	79
4.2.2 Big Data challenges and opportunities .....	85
4.3 Data for Grain Commodities Trading Decision Support .....	88
4.3.1 Data Sources .....	88
4.3.2 Integrating disparate data.....	92
4.4 Conclusion .....	93
Chapter 5 : Market Intelligence and Predictive Modelling.....	96

5.1 Introduction .....	97
5.2 Business Intelligence and Analytics .....	99
5.3 Data Science .....	100
5.3.1 Data pre-processing .....	103
5.3.2 Exploratory Data Analysis (EDA) .....	106
5.4 Machine Learning Algorithms .....	107
5.5 Neural Networks .....	110
5.5.1 Backpropagation Neural Networks .....	115
5.5.2 Features selection for the model .....	117
5.5.3 Overfitting and Generalisation .....	120
5.6 Time Series Analysis and Neural Networks .....	121
5.7 Real-time Neural Network Learning for Time Series .....	124
5.8 Conclusion .....	126
Chapter 6 : Proposed Grain Commodities Trading DSS Framework and Implementation .....	129
6.1 Introduction .....	130
6.2 Proposed Grain Trading DSS Framework .....	132
6.2.1 Real-time data acquisition and integration layer .....	133
6.2.2 Modelling layer .....	135
6.2.3 Intelligence layer .....	135
6.2.4 Visualisation layer .....	136
6.2.5 Technological considerations .....	137
6.3 Application of Framework .....	137
6.4 Implementation of Proposed Framework .....	138
6.4.1 Technology consideration .....	139
6.4.2 Data acquisition .....	140
6.4.3 Integration of datasets .....	144

6.4.4 Exploratory analysis .....	146
6.5 Modelling and Predictions.....	160
6.5.1 Neural Networks modelling experiments.....	162
6.6 Visualisation of Market Intelligence.....	181
6.7 Conclusion.....	181
Chapter 7 : Empirical Evaluation .....	185
7.1 Introduction.....	186
7.2 Evaluation of DSS.....	187
7.3 Performance Evaluation Experiments.....	189
7.3.1 Pre-August Iteration .....	191
7.3.2 Post-August Iteration.....	200
7.3.3 Comparison of Iteration 2 results and predictions by panel of experts ...	205
7.4 Conclusion.....	214
Chapter 8 : Recommendations and Conclusions .....	216
8.1 Introduction.....	217
8.2 Achievement of Research Objectives .....	219
8.2.1 Data-related requirements for a grain commodities trading DSS (RO <sub>1</sub> ) .	220
8.2.2 Modelling techniques for predicting the prices of grain commodities (RO <sub>2</sub> )	..... 222
8.2.3 Developing the framework for grain commodities trading DSS (RO <sub>3</sub> ) ....	223
8.2.4 Evaluation of the proposed framework (RO <sub>4</sub> ).....	226
8.3 Research Contributions .....	230
8.3.1 Theoretical contributions .....	231
8.3.2 Practical contributions .....	233
8.4 Limitations and Challenges.....	236
8.5 Recommendations.....	237
8.6 Future Research.....	239

8.7 Summary .....	241
REFERENCES .....	245
Appendix A: Ethics Clearance.....	255
Appendix B: Invitation to participate in survey on Landbou.com .....	256
Appendix C: Questionnaire for traders’ survey on grain commodities trading .	257
Appendix D: Questionnaire for farmers’ survey on grain commodities trading	261
Appendix E: Permission from JSE to use data.....	265
Appendix F: Sample of commodities futures transaction data .....	266
Appendix G: Sample of grain commodities spot transaction data .....	267
Appendix H: Sample of Chicago Board of Trade transactions data .....	268
Appendix I: Sample of grain commodities demand and supply data for South Africa.....	269
Appendix J: Sample of grain commodities demand and supply data for USA.	270
Appendix K: Peer-reviewed conference paper .....	271
Appendix L: Submitted journal article.....	278

## List of Figures

Figure 1.1: Chapter outline.....	12
Figure 2.1: Steps in research design (Collis and Hussey, 2009).....	17
Figure 2.2: The research onion (Saunders, Lewis and Thornhill, 2009).....	18
Figure 2.3: Design science research cycle (Hevner, 2007).....	25
Figure 2.4: DSR process model (Peppers et al., 2008).....	29
Figure 2.5: DSR process model (DSR cycle) (Vaishnavi and Kuechler, 2015).....	30
Figure 2.6: Envisaged DSR process and cycle for this study (Adapter from Peppers et al. (2008) and Johannesson and Perjons (2012)). .....	33
Figure 3.1: Chapter outline and deliverables.....	37
Figure 3.2: Role of information in reducing uncertainty and equivocality (Daft and Lengel, 1986) .....	46
Figure 3.3: Data-centric and Organisational mechanism based information processing (Kowalczyk and Buxmann, 2014).....	47
Figure 3.4: Framework showing intrinsic relationship between Data, Information, Knowledge and Decision making. (Adapted from Sabherwal and Becerra-Fernandez (2011)).....	54
Figure 4.1: Chapter outline and deliverables.....	72
Figure 4.2: Big Data techniques (Chen and Zhang, 2014).....	80
Figure 4.3: Initial framework for grain commodities trading DSS.....	94
Figure 5.1: Chapter outline and deliverables.....	96
Figure 5.2: The Data Science process (O'Neil and Schutt, 2014).....	102
Figure 5.3: Simple Neural Network (Engelbrecht, 2007).....	111
Figure 5.4: Diagram showing Neural Networks weighting (Lantz, 2013).....	112
Figure 6.1: Chapter outline and deliverables.....	129
Figure 6.2: Proposed grain commodities trading DSS.....	134
Figure 6.3: Database schema for spot price modelling .....	145
Figure 6.4: Database schema for futures price modelling .....	145
Figure 6.5: Graph showing the spot price of white maize for all historical data .....	147
Figure 6.6: Graphs showing the spot prices of white maize over different periods. ....	148
Figure 6.7: Graphs showing the spot prices of white maize and wheat.....	149
Figure 6.8: Graphs showing the spot prices of white maize and USD-Rand Exchange rates .....	150

Figure 6.9: Graphs showing the spot prices of white maize and Brent crude oil ....	150
Figure 6.10: Graphs showing the spot prices of white maize and interest rates.....	151
Figure 6.11: Graphs showing the spot prices of white maize in South Africa and price of corn in USA .....	152
Figure 6.12: Graphs showing the spot prices of white maize and volume of corn trade on Chicago Board of Trade (USA) .....	153
Figure 6.13: Graphs showing the spot prices and demand of white maize .....	154
Figure 6.14: Graphs showing the spot prices and supply of white maize .....	154
Figure 6.15: Graphs showing the closing price of December futures contract of white maize.....	156
Figure 6.16: Graph showing the spot price against closing price of December futures contract of white maize.....	157
Figure 6.17: Comparison of actual vs predicted spot prices of white maize (1 month in-sample) .....	172
Figure 6.18: Comparison of actual vs predicted spot prices of white maize (1 month out-sample) .....	172
Figure 6.19: Comparison of actual vs predicted spot prices of white maize (3 months in-sample) .....	173
Figure 6.20: Comparison of actual vs predicted spot prices of white maize (3 months out-sample) .....	173
Figure 6.21: Comparison of actual vs predicted spot prices of white maize (6 months in-sample) .....	174
Figure 6.22: Comparison of actual vs predicted spot prices of white maize (6 months out-sample) .....	175
Figure 6.23: Comparison of actual vs predicted December futures contract of white maize (1-month in-sample).....	177
Figure 6.24: Comparison of actual vs predicted December futures contract of white maize (1-month out-sample) .....	177
Figure 6.25: Comparison of actual vs predicted December futures contract of white maize (3-month in-sample).....	178
Figure 6.26: Comparison of actual vs predicted December futures contract of white maize (3-month out-sample) .....	178

Figure 6.27: Comparison of actual vs predicted December futures contract of white maize (6-month in-sample).....	180
Figure 6.28: Comparison of actual vs predicted December futures contract of white maize (6-month out-sample) .....	180
Figure 7.1: Chapter outline and deliverables.....	185
Figure 7.2: Prediction of spot prices of white maize by experts and DSS (Iteration 1) .....	194
Figure 7.3: Error measurements of experts and DSS predictions for spot prices (Iteration 1).....	194
Figure 7.4: Correlation between predictions of spot prices and actual values (Iteration 1).....	195
Figure 7.5: Prediction of December futures contract prices of white maize by experts and DSS (Iteration 1).....	196
Figure 7.6: Error measurements of experts and DSS predictions for spot prices (Iteration 1).....	198
Figure 7.7: Correlation between predictions of spot prices and actual values (Iteration 1).....	199
Figure 7.8: Graph showing actual and predicted spot prices of white maize with different models.....	203
Figure 7.9: Graph showing actual and predicted December futures prices of white maize with different models.....	205
Figure 7.10: Prediction of spot prices of white maize by experts and DSS .....	208
Figure 7.11: Error measurements of experts and DSS predictions for spot prices .	209
Figure 7.12: Correlation between predictions of spot prices and actual values .....	209
Figure 7.13: Prediction of December futures prices of white maize by experts and DSS.....	210
Figure 7.14: Error measurements of experts and DSS predictions for December futures prices.....	212
Figure 7.15: Correlation between predicted December futures contract prices and actual values .....	212
Figure 8.1: Chapter outline and deliverables.....	216
Figure 8.2: Proposed framework for grain commodities trading .....	225
Figure 8.3: Prediction of spot prices of white maize by experts and DSS .....	229

Figure 8.4: Prediction of December futures prices of white maize by experts and DSS..... 229

Figure 8.5: Decision making model (Adapted from Simon, 1960; Hammond, Keeney and Raiffa, 1999; Bazerman, 2006)..... 232



## List of Tables

Table 1.1: Research questions, objectives and chapter deliverables .....	9
Table 3.1: Phases and steps in decision making .....	43
Table 6.1: Descriptive statistics for spot prices of white maize over different periods .....	147
Table 6.2: Correlation between spot price of white maize and other variables.....	155
Table 6.3: Correlation between price of December futures contract of white maize and other variables.....	158
Table 6.4: Input variables for Neural Network model for WMAZ spot price .....	161
Table 6.5: Input variables for Neural Network model for WMAZ December futures contract price.....	161
Table 6.6: Mandatory parameters for setting BPNN topology in SAP HANA (SAP, 2015).....	162
Table 6.7: Optional parameters for setting BPNN topology in SAP HANA (SAP, 2015).....	163
Table 6.8: Comparison of BPNN models for spot prices of white maize .....	167
Table 6.9: Comparison of BPNN models for December futures contract prices of white maize .....	169
Table 6.10: Summary of verification of BPNN model for spot prices .....	171
Table 6.11: Summary of verification of BPNN model for December futures prices	176
Table 7.1: Evaluation criteria for DSS .....	188
Table 7.2: Comparison between predictions from experts and implemented DSS for spot prices of white maize (Iteration 1).....	193
Table 7.3: Comparison between predictions from experts and implemented DSS for December futures contract prices .....	197
Table 7.4: Tables showing the input datasets used in category C modelling .....	201
Table 7.5: Actual and predicted spot prices of white maize.....	202
Table 7.6: Actual and predicted December futures prices of white maize .....	204
Table 7.7: Comparison between predictions from experts and implemented DSS for spot prices of white maize .....	207
Table 7.8: Comparison between predictions from experts and implemented DSS for December futures contract prices .....	211
Table 8.1: Research objectives and questions addressed in study .....	220

Table 8.2: Summary of identified datasets and their sources.....	221
Table 8.3: Input variables for Backpropagation Neural Network model for predicting spot prices of white maize .....	226
Table 8.4: Input variables for Backpropagation Neural Network model for predicting December future contract prices of white maize .....	227

## List of Abbreviations

Abbreviations	Terms in full
API	Application Program Interface
AR	Auto-Regression
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with Exogenous variables
BI	Business Intelligence
BI&A	Business Intelligence and Analytics
BPNN	Backpropagation Neural Networks
CBOT	Chicago Board of Trade
CSV	Comma Separated Values
DAFF	Department of Agriculture, Forestry and Fisheries
DSR	Design Science Research
DSS	Decision Support System
EDA	Exploratory Data Analysis
ERS	Economic Research Service
HTML	Hyper Text Markup Language
ICT	Information and Communication Technology
IS	Information Systems
JSE	Johannesburg Stock Exchange
JSON	JavaScript Object Notation
MA	Moving Averages
ML	Machine Learning
MSE	Mean Square Error
NMMU	Nelson Mandela Metropolitan University
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
PDF	Portable Document Format
RBF	Radial Basis Function
RMSE	Root Mean Square Error
SAFEX	South African Futures Exchange

SAGIS	South African Grain Information Services
SARB	South African Reserve Bank
USA	United States of America
USDA	United States Department of Agriculture
VAR	Vector Autoregressive
WMAZ	White Maize

# Chapter 1 : Research Context

## 1.1 Background

The production of grain commodities is an important agricultural industry in South Africa. Output from the industry includes major staples for human consumption and important components of animal feeds (DAFF, 2014). Besides what is consumed locally, the grain commodities produced in South Africa are also exported to neighbouring countries and to other parts of the world, making the trade a source of foreign exchange income. Grain commodities marketing is currently facilitated by the Johannesburg Stock Exchange (JSE), operating according to the provisions of the Agricultural Marketing Act (Act No. 47, 1996) (Doyer, D'Haese, Kirsten and Van Rooyen, 2007). The core function of the JSE is to facilitate the trade of grain commodities and to provide an enabling environment for risk management and price discovery (Venter, Strydom and Grové, 2013).

The trading of grain commodities on the stock exchange in South Africa is *Laissez Faire* in nature. In essence, this means that the market, and effectively the prices of the grain commodities, are controlled by several local and international economic, political and social factors that are rapidly changing. Therefore, stakeholders in the industry, especially the grain farmers, are constantly exposed to price-related risks due to the volatility of prices of grain commodities (Venter, Strydom and Grové, 2013).

The volatility in the prices of grain commodities and other agricultural products has been a source of concern for academic researchers, governmental and non-governmental organisations for many decades (Trostle, 2008; Wright, 2011). This is because the volatility of the prices of agricultural commodities has dire and multifaceted implications for stakeholders in the industry and on the economy of the nation at large. There are indications that changes in the prices of agricultural commodities have social implications on issues such as the fight against poverty and also have economic implications for the Gross Domestic Product (GDP) of a country and the sustainability of the agricultural sector which is very important in many countries (Headey and Fan, 2008; Trostle, 2008). Hence, governments of different

countries develop policies that are believed to be in the best interest of agricultural commodities trading.

The volatility of grain commodities prices and the associated price-related risks suggest that stakeholders will be confronted with important decisions when marketing their products. Different trading alternatives such as spot, futures contracts, forward contracts and options are available on the stock exchange for trading grain commodities. The choice of the right trading alternative can be used to manage price-related risks in trading grain commodities. It has been noted that trading grain commodities using the spot alternative is more risky (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013). Previous studies, however, have shown that a number of South African farmers and some other stakeholders in the industry still trade their grain commodities mainly by using the spot alternative (Jordaan and Grové, 2010). As a result, these stakeholders do not take full advantage of all the alternatives that are available to them for managing price-related risks and increasing profitability. This has been attributed to the complexities associated with obtaining grain market intelligence and determining a future outlook (Jordaan, Grové, Jooste and Alemu, 2007; Venter, Strydom and Grové, 2013).

In order to optimise income and reduce price risks by choosing the correct trading alternative, knowledge about the future outlook and performance of the grain market for the different trading alternatives is required. To achieve this, it is required at present, that stakeholders in the industry sift through volumes of economic, political and social data (Wright, 2011; Trostle, 2008) that has to be sourced from various places. Moreover, they are required to make sense out of the changes in this data as it relates to the price of grain commodities on a regular basis (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013). Therefore, a system for supporting decision making for trading in grain commodities will be beneficial to the industry. Such a decision support system should be able to provide price predictions for the different trading options, thereby improving decision making about grain commodities trading.

In a rapidly changing environment, the ability to make the right decisions at the right time has been found to be directly responsible for increased efficiency and productivity

(Brynjolfsson, Hitt and Kim, 2011). On the other hand, organisations that consistently make poor decisions, especially regarding the management of risks and capturing value that will increase profitability might have their sustainability threatened (Bazerman and Chugh, 2006; Davenport, 2009). Two major problems that decision makers face are uncertainty and equivocality. The former is a result of the lack of sufficient information needed for decision making, while the latter can be described as the lack of ability to comprehend the available information. This could be due to the volume of the information or the lack of capacity to comprehend (Kowalczyk and Buxmann, 2014).

Contextually, this could be described as the dilemma of the average grain commodities farmers that enjoy farming activities but are unable to get the best price for their produce. Within the value chain of the grain commodities production and trade in South Africa, the grain commodities farmers seem to be price-takers, because they are compelled to take the prevailing market prices without being able to influence the prices and are left with limited options. In the long run, this can be seen as a threat to the sustainability of the operations of such farmers due to the price-related risks that they face yearly. This could be a contributing factor to the dwindling national production of grain commodities as indicated in reports by the South Africa Department of Agriculture, Forestry and Fisheries (DAFF, 2014). Therefore, a Decision Support System that helps such groups to make better decisions in managing their price-risks and thus increase profitability will be beneficial.

Computer-based Decision Support Systems (DSS) bring together a set of tools, techniques and practices that provide interventions that can be used to improve decision making (Sauter, 2010; Demirkan and Delen, 2013). This is achieved by enabling the gathering, sorting and manipulation of data for the purpose of extracting valuable information and knowledge. Studies show that the use of data and different levels of analytic tools and techniques are the bedrock of DSS (Sauter, 2010; Chen, Chiang and Storey, 2012). Recently, more data has become available and is easily accessible on several subjects, creating new opportunities and challenges for both researchers and practitioners to extract information and knowledge and create value by using data-centric systems.

The volume of data available globally has grown significantly and the rate of growth is increasing by the minute (Manyika et al., 2011). Organisations are now able to capture large volumes of data through transactions with their customers, and suppliers, and through business operations, social networks, online shopping applications, mobile communication, etc. (Manyika et al., 2011). A report from IBM in 2011 indicated that 90% of the world's data had been created in the previous two years and that all the data created in time past formed only 10% of the global data repository (IBM, 2011).

Big Data has been defined as large datasets that require much more than the capabilities available with conventional tools for collecting, storing, managing and analysing data (Manyika et al., 2011; Minelli, Chambers and Dhiraj, 2013). The peculiar characteristics of Big Data have also been used in defining the phenomenon. Big Data can also be defined as very large data sets (Volume), that are created at unusually high rate (Velocity), generally heterogeneous (Variety) and susceptible to inaccuracies (Veracity) (Chen, Chiang and Storey, 2012; Mayer-Schonberger and Cukier, 2013; Goes, 2014).

This deluge of data generated is now described as Big Data and has changed the way people live, how organisations operate and how business is conducted. The impact is also being felt in the field of science and academic research (Chae and Olson, 2013). Some of the forward-thinking organisations now consider Big Data as part of their organisational assets. These organisations are now exploring the possibilities of deriving competitive advantage from data. This is achieved by looking into their internal and external sources of data to gain a competitive edge and some of these organisations are now referred to as data-driven organisations (McAfee and Brynjolfsson, 2012).

Developments regarding Big Data have made it possible to obtain, analyse and derive value from more data. The evolution of Big Data tools, techniques and approaches makes it possible to acquire large volumes and varieties of structured or unstructured data in real-time (Chen and Zhang, 2014; Goes, 2014). This evolving concept provides a platform for integrating data of any kind, thereby making it possible to extract more value that will benefit decision making for organisations or industries. The possibilities



of Big Data are unleashed through the use of business analytic tools that are based on mathematical and statistical models and presented in simplified formats to extract business intelligence to support decision making at all levels (Dhar, 2013). Big Data is complemented by data mining concepts, business analytics and business intelligence to extract embedded and actionable knowledge from data sources (Năstase and Stoica, 2010).

The availability and integration of relevant data from different sources about the grain commodities market in South Africa could offer more insight about grain commodities trading in South Africa. Moreover, the availability and analysis of such data in real-time for market intelligence, could offer new perspectives and support for decisions regarding trading in grain commodities in South Africa. This study will attempt to investigate how traders in grain commodities can receive support for making decisions about trading their commodities to reduce price-related risks and maximise profits.

## **1.2 Problem Statement**

Price-related risks are a major concern in the trading of grain commodities as a financial derivative on stock exchanges all over the world. Several trading strategies are available; besides selling grain commodities on a cash basis, known as the spot, there are also forward contracts, futures contracts and options trading alternatives (Hull, 2012). The forward contract is an agreement to sell a commodity to a buyer for an agreed price and at a set future date. The future contract is similar to the forward contract, but the transaction takes place through the stock exchange and not between the buyer and the seller. On the other hand, the option strategy allows the buyer or the seller, to signify intention to buy or the sell on the stock exchange, but without an obligation to honour such transaction. All of these alternatives provide an opportunity for effective managing of price-related risks and the discovery of the best prices in the trading of grain commodities (Venter, Strydom and Grové, 2013). However, it has been found that many people, especially farmers, do not take advantage of all the alternatives that are available to them in managing price-related risks and maximising profits with the exception of the cash (spot) sales (Jordaan and Grové, 2010).

Making use of the different trading alternatives will require a more involved participation in the financial markets and a better understanding of the economic climate. This requires keeping abreast of the rapidly changing local and global markets to interpret implications for the present and also to predict the future (Venter, Strydom and Grové, 2013). By so doing, farmers and traders will have to deal with an enormous amount of financial, economic, political and social data from several sources (Trostle, 2008; Wright, 2011). The farmers are very important stakeholders in the grain commodities industry as producers, but most of them do not possess the skills, expertise and willingness to engage in activities such as scouting for data and market intelligence (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013). Hence, most of the farmers would rather sell their produce on a cash (spot) basis, which makes them vulnerable to more price-related risks.

Hence the main problem that will be addressed in this study is:

***Grain farmers in South Africa do not take full advantage of all the available strategies for trading their grain commodities because of the complexities associated with monitoring the datasets that influence the grain commodities market.***

This study will attempt to address the challenge by exploring the use of Computer-based Decision Support Systems that can help stakeholders such as farmers, with limited skills and expertise for collecting and interpreting datasets in order to increase the benefit from the market. Besides the farmers, other stakeholders in the industry may also benefit from the outcome of this study that is expected to provide a framework for developing a DSS for grain commodities trading in South Africa. This will be achieved by exploring the possible role that the evolving Big Data and associated concepts such as Data Science can play in improving decision making in the industry.

### 1.3 Thesis Statement

The proposed thesis statement for this study is as follows:

***A framework for making effective decisions about trading grain commodities can be developed, which utilises the Big Data approach in collecting and analysing datasets that influence grain commodities prices in South Africa for assisting farmers in taking full advantage of the available strategies for trading their grain commodities.***

### 1.4 Research Objectives

The purpose of this study is to propose a framework that can be followed for the development of a Decision Support System that will enable relevant stakeholders to make decisions regarding trading grain commodities in South Africa. Stakeholders, such as farmers, with limited skills and expertise will be the primary focus of this research. The components of the DSS will focus on predicting the price of grain commodities for the different trading alternatives available. Real-time extraction of market intelligence and price prediction will also be examined. In order to achieve this purpose, the following objectives have been set.

The main research objective (**RO<sub>m</sub>**) of this study is:

***RO<sub>m</sub>: To design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities.***

To achieve the main research objective, the secondary objectives below have also been set:

**RO<sub>1</sub>:** To identify data-related requirements for a system to support decisions on trading grain commodities in South Africa.

**RO<sub>2</sub>:** To identify modelling techniques for predicting the future prices of grain commodities in South Africa.

**RO<sub>3</sub>:** To develop a framework to support decisions on grain commodities trading.

**RO<sub>4</sub>:** To evaluate the capabilities of a Decision Support System that is developed by following the proposed framework in predicting grain commodities prices.

## 1.5 Research Questions

Based on the purpose of this research, the problem that has been identified and the objectives that have been set, the main research question (**RQ<sub>m</sub>**) of this study is:

**RQ<sub>m</sub>:** *How can decision making of grain commodities farmers about trading in grain commodities be improved using the Big Data approach?*

In order to answer the main research question, it has been broken down further into the following sub-questions:

**RQ<sub>1</sub>:** What are the local and international factors that influence the grain commodities market in South Africa?

**RQ<sub>2</sub>:** What strategies in trading grain commodities are available for minimising price-related risks and increasing profitability?

**RQ<sub>3</sub>:** What datasets influence the prices of grain commodities in South Africa?

**RQ<sub>4</sub>:** What are the modelling techniques utilised for discovering patterns and making predictions from datasets?

**RQ<sub>5</sub>:** How can a framework for a system to support decisions about trading grain commodities be developed and implemented?

**RQ<sub>6</sub>:** How well does a DSS perform, which was developed by utilising the framework?

Table 1.1 presents a structural layout of how each of the research questions will lead to achieving the objectives that have been outlined above. Moreover, the table indicates the chapter layout of this thesis.

Table 1.1: Research questions, objectives and chapter deliverables

Research Questions	Research Objectives	Thesis chapter
RQ <sub>m</sub>	RO <sub>m</sub>	Chapter 1 – Research context
RQ <sub>1</sub>	RO <sub>1</sub>	Chapter 3 – Decision support for grain commodities trading
RQ <sub>2</sub>		
RQ <sub>3</sub>		Chapter 4 – Big Data for grain commodities trading in South Africa
RQ <sub>4</sub>	RO <sub>2</sub>	Chapter 5 – Market intelligence and predictive modelling
RQ <sub>5</sub>	RO <sub>3</sub>	Chapter 6 – Proposed grain commodities trading DSS framework and implementation
RQ <sub>6</sub>	RO <sub>4</sub>	Chapter 7 – Empirical evaluation
RQ <sub>m</sub>	RO <sub>m</sub>	Chapter 8 – Recommendations and conclusions

## 1.6 Scope and Constraints

The primary deliverable of this research is a framework for improving and supporting decision making regarding trading grain commodities. It is expected that the resulting framework from this study will form an abstraction for the development of a decision support system that takes advantage of the opportunities of Big Data and associated concepts. It is also expected that this research will outline the components of such a DSS and important considerations in bringing the components together. This study will outline the approach that should be taken for the acquisition and integration of data from disparate sources to form an integrated source of data that can support decisions about trading grain commodities in South Africa. Furthermore, a demonstration of how the integrated data can be used to predict prices of grain commodities in South Africa will also be carried out in this study.

The resulting framework of this study will be developed with the focus on white maize. However, it is assumed that the framework would be sufficient for developing a DSS that encompasses other grain commodities. It is expected that the framework could be used to extract market intelligence, such as predictions, recommendations and new discoveries from the integrated data source that will be created as a component of the DSS. However, the demonstration and evaluation of the framework will only focus on

using the dataset for predicting the spot and December futures contract prices of white maize in South Africa.

This study will outline the opportunities, processes and major considerations in the acquisition and integration of datasets that are relevant to grain commodities trading in South Africa. The necessary process of identifying the subset of such datasets that are required for extracting market intelligence, which can support decisions about trading grain commodities in South Africa, will also be outlined. However, the datasets that will be acquired might not represent all the factors that influence the grain commodities market in South Africa.

## **1.7 Proposed Thesis Structure**

Based on the purpose, the research objective and the research questions that have been identified in this chapter, the proposed structure of this thesis is presented below:

**Chapter 1 – Research context:** This chapter provides a general overview of the research. It explicates the identified problem, opportunity and the relevance of this research for the stakeholders trading in grain commodities and how this study will add value to the knowledge base of the subject matter. Chapter 1 also outlines the research objectives and questions that will serve as the foundation and guidelines for this study.

**Chapter 2 – Research design and methodology:** This chapter will discuss and motivate the research philosophy and chosen methodology. It will explain the relevance of design science to this study and how each chapter and component of this research fits into the design science research process.

**Chapter 3 – Decision support for grain commodities trading:** This chapter will outline the requirements and components of a computer-based Decision Support System (DSS). The requirements for decisions about trading grain commodities in South Africa will be identified. Chapter 3 will, furthermore, discuss the type of information and knowledge that can support decision making regarding grain trading in South Africa. In order to achieve this, the factors that influence the

grain commodities market in South Africa and the available trading alternatives on the stock exchange will be discussed.

**Chapter 4 – Big Data for grain commodities trading in South Africa:** This chapter will explore the opportunities that the Big Data approach, tools and techniques can offer in developing a rich source of data about grain commodities trading in South Africa. The types and sources of datasets that can influence grain commodities trading in South Africa will be addressed. Moreover, this chapter will explore how to acquire and integrate the disparate datasets in real-time and the technological considerations of doing so.

**Chapter 5 – Market intelligence and predictive modelling:** Data Science is one of the concepts evolving with Big Data. Chapter 5 will explore the Data Science process and concept for extracting actionable insights from large datasets. Specifically, the application of Neural Networks for modelling time series data, such as the grain commodities trading data, will be explored. A model for the use of Neural Networks for predicting the prices of grain commodities will also be developed.

**Chapter 6 – Proposed grain commodities trading DSS framework and implementation:** A framework that can be followed in the development of a grain commodities trading DSS in South Africa will be proposed. This will be based on the review of literature and the scientific grounding from Chapters 3, 4 and 5. Furthermore, it will report on an implementation that demonstrates the ability of the proposed framework to address the problem that has been identified in this study.

**Chapter 7 – Empirical evaluation:** Chapter 7 will evaluate the proposed framework to ascertain how well the framework can solve the problem that has been identified by this study. An evaluation of the framework's usefulness and the technical ability of the developed DSS, based on the proposed framework, will be carried out in this chapter. This will be achieved by comparing the predictions from the developed DSS, based on the proposed framework, and the predictions made by a panel of experts who make predictions based on their skills and many years of experience.

**Chapter 8 – Recommendation and conclusions:** This chapter will be used to summarise the knowledge developed during this study as concluding remarks.

Chapter 8 will outline the practical and the theoretical contribution of this study, and thereafter make recommendations for the industry. Furthermore, the limitations and challenges encountered during the study will be discussed and suggestions for future research will also be outlined.

Figure 1.1 shows a graphical representation of the structure of this thesis as described above. It also depicts the research objectives and the research questions that will be focused on in each of the chapters of this thesis as indicated in Table 1.1. It is expected that by addressing the relevant research questions and by achieving the set objectives, each of the chapters will lead to specific outcomes. It is envisaged that these outcomes will eventually lead to providing a solution to the problem that has been identified.

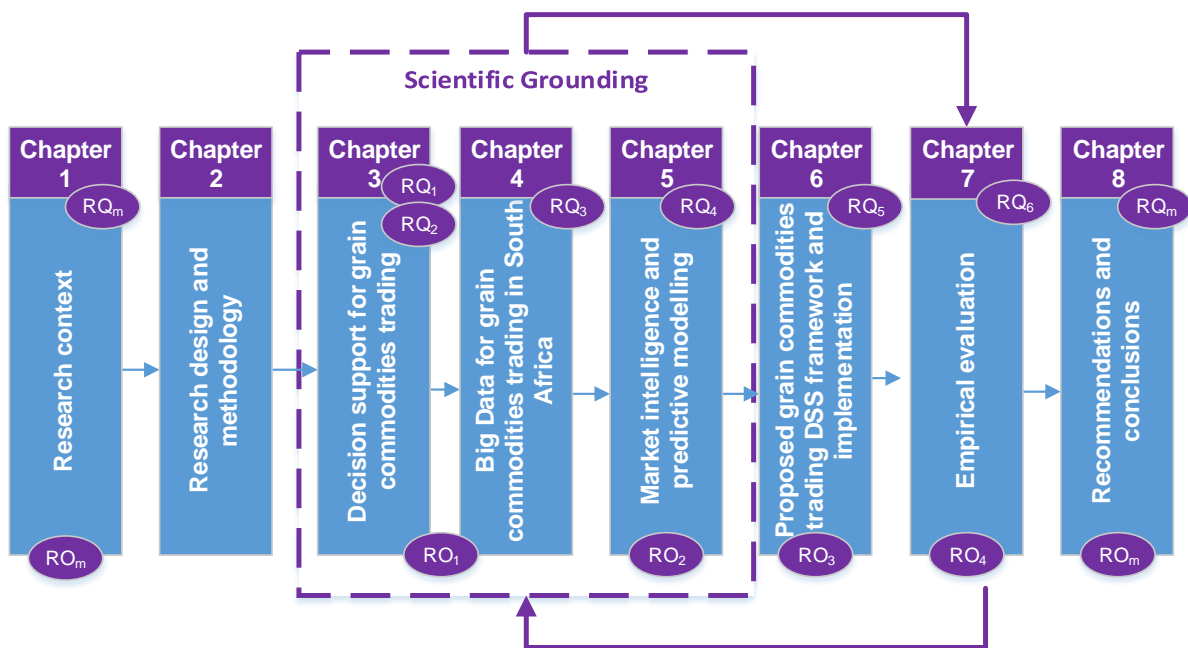


Figure 1.1: Chapter outline

As shown in Figure 1.1, it is envisaged that Chapters 3, 4 and 5 will provide the theoretical background, ideas, methods and concepts that will be synthesised as the scientific grounding for this study. It is expected that this scientific grounding will lead to the development of a framework that can be followed in the development of a system that can support farmers and other stakeholders in making decisions about trading grain commodities. There is a possibility of an iterative process between the



scientific background and the evaluation of the proposed framework to ensure that the outcome of the study provides an adequate solution for the identified problem. A suitable research methodology for an iterative research process in the development of an artefact will be explored in Chapter 2.

## **1.8 Summary**

There is a need to provide support for grain commodities farmers regarding making decisions about trading their produce (Section 1.1). They require decision support that will empower them to make use of the correct trading alternatives that will minimise price-related risks when selling their grain commodities. In order to do this, there is a need for them to predict the future prices of grain commodities for different trading alternatives, so that they can make the best choice. However, it was established in this chapter that taking advantage of all of the available trading alternatives will require the farmers to sift through and make sense of disparate data from different sources. Hence, some of the farmers are systematically excluded from grain commodities trading strategies that can help them to manage price-related risks and increase the profitability of their operations. This in effect could affect the sustainability of such farms and in the long run, the overall production of grain commodities in South Africa.

This study identifies an opportunity for real-time collection and integration of disparate datasets that influence the grain commodities market in South Africa into a single data source. This data can then be analysed to extract insights, such as the price prediction of grain commodities, to assist in making better decisions regarding grain commodities marketing in South Africa. It was proposed that the Big Data approach and associated technologies could offer leverage for the collection, integration and analysis of relevant datasets to achieve the purpose of this study. This study was broken down into research objectives (Section 1.4) and research questions (Section 1.5) in order to provide a solution to the identified problem. The scope and constraints of this study were identified in Section 1.6 as being limited to the use of daily data, although, it is possible to collect more granular data for live implementation.

Section 1.7 outlines the proposed structure of this thesis and the expectation from each chapter. The next chapter will provide an overview of the research design and methodologies. It will discuss and motivate the research philosophy, the research methodology and the research strategies that will be adopted during this study.

## **Chapter 2 : Research Design and Methodology**

### **2.1 Introduction**

The previous chapter presented the motivation for this study by describing the research problem and the community that is affected by the problem. The identified problem formed the basis of asking research questions and setting the research objectives itemised in Table 1.1. This study aims to provide a practical solution to the identified problem by attempting to answer the research questions that have been raised and achieve the research objectives that have been set forth. This chapter provides the choice, motivation and description of the research design and methodology that will be followed in the rest of this study to ensure that a logical and unambiguous solution is provided for the identified problem and is backed by scientific rigour.

Research can be defined as an undertaking to make discoveries, to increase knowledge or produce new knowledge (Saunders, Lewis and Thornhill, 2009). Generally, research activities are expected to be carried out as a structured, systematic and methodical process of investigation and collection of evidence that leads to new knowledge (Collis and Hussey, 2009). This requires that a choice of the procedure of investigation and methods that will be used in gathering data and evidence be carefully selected at the beginning of the research. Selecting the right procedure and methods of investigation for the research is encapsulated in the choice of a suitable research design, approach, strategy and methodology that will be discussed further in this chapter. Making an appropriate choice should be a function of the nature of the problem that is being addressed and possibly of the field of interest (Creswell, 2014).

The main objective of this study (RO<sub>m</sub>) is to design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities. Achieving this objective may require that this research integrate knowledge from different disciplines, such as Agricultural Economics, Finance, Statistics and Computer Science. It is expected that ICT will be used as an enabler in providing a solution to the identified problem, which can be

described as organisational or even institutional when the industry is considered. This qualifies this study as Information System (IS) research.

IS brings software, hardware, data resources, people and processes together to achieve organisational success (Vahidov, 2012). It makes use of technology and technological abilities by creating structures that enable organisations to meet their needs or improve their functionality. Thus, IS research could be concerned with the development of artefacts that provide IS support within an organisation. Additionally, research in the field of IS could focus on providing IS practitioners with “*knowledge for developing and improving IS-enabled initiatives as well as knowledge for implementing and integrating solutions in an organisational context*” (Carlsson, Henningsson, Hrastinski and Keller, 2011).

Traditional research paradigms that are used in natural and social science research are considered to be suitable for carrying out IS-related research. Such research paradigms, however, have been found to be limited in dealing with the nature of IS research (Peffer, Tuunanen, Rothenberger and Chatterjee, 2008). However, Design Science Research (DSR) has continued to gain acceptance as being suitable for IS research (Hevner, March, Park and Ram, 2004; Hevner, 2007; Hevner and Chatterjee, 2010; Beck, Weber and Gregory, 2013; Vaishnavi and Kuechler, 2015). Natural and social science research focuses on behavioural science to describe, explain, explore and predict (Johannesson and Perjons, 2012). DSR can be used to create artefacts that solve problems and create new opportunities for an organisation while taking all the necessary systems, structures, processes and resources in the organisation into consideration (Hevner et al., 2004; Peffer et al., 2008; Carlsson et al., 2011).

The DSR methodology will be adopted in this study as the suitable research methodology for addressing the identified problem. Section 2.2 will provide an overview of the research design; the research approach will be discussed in Section 2.3 and the research strategy in Section 2.4. Thereafter, Section 2.5 will provide an overview of the DSR methodology and the motivation for its choice for this study. Section 2.5 will also explain the implementation of the DSR process for this study.

Ethical considerations will be addressed in Section 2.6 and a summary of the chapter will be provided in Section 2.7.

## 2.2 Research Design

Research design is the blueprint of how a research project will be carried out to ensure that the research effort is structured and systematic. It is the planning of the tasks and procedures that will be followed during the research process in order to ensure that the outcome of the process is a valid research output (Collis and Hussey, 2009). The research design guides the plan of action on how to find answers to the questions and achieve the set objectives of a research exercise (Blumberg, Cooper and Schindler, 2011). This makes the research design a systematic outline of how the research question will be carried out as a research project (Saunders, Lewis and Thornhill, 2009). This implies that a valid and well thought-through research question is a prerequisite for selecting the research design. (Collis and Hussey, 2009) present the steps that should be followed in research design as shown in Figure 2.1.



Figure 2.1: Steps in research design (Collis and Hussey, 2009)

Figure 2.1 reiterates that the research design starts with the identification of a valid research problem that is followed by setting the aims and objectives of the research.

This is followed by the identification of questions that the research process will seek to find answers to. Based on the previous steps, the strategy and methods that will be followed in the research process should be determined and a list of deliverables can be set for the different phases of the research. The outcome of each of these steps can then be put together as a blueprint for the study, which is the research design. In order to effectively carry out the first three steps of the research design presented in Figure 2.1, there is a need for the researcher to decide on an appropriate research philosophy/paradigm. In explaining how the research design depends on the selected research philosophy, (Saunders, Lewis and Thornhill, 2009) present a framework described as the “research onion”. It shows how the research philosophy, approach, strategy and methods are inter-related in the form of layers, with each layer depending on the other, from the core to the outer layer.

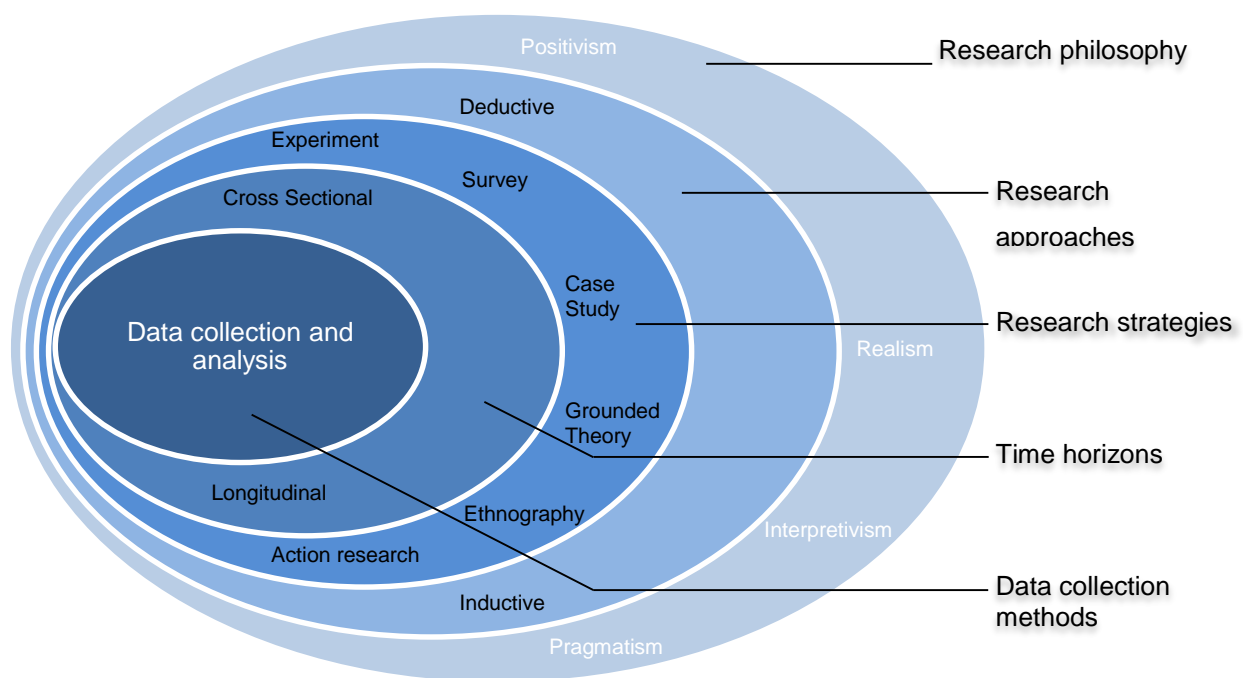


Figure 2.2: The research onion (Saunders, Lewis and Thornhill, 2009)

### 2.2.1 Research philosophy

Research philosophies are the assumptions of worldview and the philosophical framework that will guide the research process (Saunders, Lewis and Thornhill, 2009).

The adopted research philosophy will determine the reasoning pattern in the course of the research, the perception of reality, the nature of knowledge and what is considered as knowledge (Collis and Hussey, 2009; Creswell, 2014). The research philosophy is at the outer layer of the research onion as shown in Figure 2.2 since every other decision that will be made in the research design depends on the research philosophy. Over the years, there has been an evolution of research philosophies; however the following perspectives have remained relevant in literature and encapsulate other philosophies (Collis and Hussey, 2009; Saunders, Lewis and Thornhill, 2009; Creswell, 2014):

- **Positivism:** is an objective philosophy with a general ideal that reality and the process of creating knowledge should be independent of the researcher's opinion. Positivism is mostly used in natural sciences, but has also been adopted in other fields such as social sciences and business research. It guides the use of existing theories to describe, explain or predict occurrences of different phenomena with the goal of validating or refuting hypotheses set during the research.
- **Realism:** has its roots in the positivism philosophy. The perception of realism is that realities are independent of the mind, suggesting that the mind does not need to further interpret the realities captured by the senses. Hence, its main application is in scientific investigation for the collection and interpretation of data.
- **Interpretivism:** proposes that research activities, especially of complex phenomena, should allow for and capture the subjective influence of the perception of social-actors such as the researcher during the research process. It suggests that research involving people needs to accommodate individual differences. Hence, it advocates that the research design should be less structural and contextual in order to capture all the social realities in the process of knowledge creation.
- **Pragmatism:** as a school of thought, prefers to see the choice between the positivism and interpretivism philosophies as a continuum rather than the taking of a position. This philosophical approach suggests that the researcher should not be forced into taking a philosophical position, but that the research problem

and question to be answered should determine the philosophy or philosophies that will guide the research. Pragmatism suggests that the focus should be on carrying out the research in whichever way the researcher deems necessary, thus adding value to the body of knowledge.

The nature of reality and the perception of what is true knowledge are the major defining factors of the different research philosophies as seen in the description of the philosophies above. The former is known as the ontology and the latter is described as the epistemology of the different philosophies.

### **2.2.2 Adopted philosophy for study**

Niehaves (2007) claimed that some earlier studies regard DSR as a new research philosophy, but the author argued succinctly that DSR should not be regarded as a new research philosophy altogether. Rather than a new research philosophy, Niehaves (2007) concluded that DSR provides opportunities for a diversity of philosophical grounding when conducting a research study. On this basis, DSR could be based on positivism, interpretivism or a combination of both (positivism and interpretivism) and should not be regarded as another research philosophy. This notion is supported by Vaishnavi and Kuechler (2015). To further understand the philosophical position of DSR, especially for IS-related research, one can look at the kind of problems where they are applicable.

DSR can be used in addressing problems with complex interaction between systems, processes, resources and people which could be grounded in multiple disciplines (Hevner and Chatterjee, 2010; Vaishnavi and Kuechler, 2015). The nature of these kinds of problems could mean having to contend with conflicting theoretical backgrounds. DSR is able to handle such problems by navigating through the research in an iterative manner in that epistemological and ontological considerations of the study might change during the DSR cycles or within iterations of a particular phase of the research (Vaishnavi and Kuechler, 2015). The researcher is allowed to introduce creativity during the research process from a subjective viewpoint which is based on an interpretivism philosophy.



An example could be the interpretivist, subjective participation in the research in the explication of the problem or identification of opportunities (Johannesson and Perjons, 2012). This could be particularly necessary when a study is on a course of innovating or solving “un-theorised” problems. Thereafter, the researcher can step back as an objective positivist to observe the evolution of the system/artefact being developed and measure it against a relevant theoretical background (Vaishnavi and Kuechler, 2015).

The nature of the problem, objectives that have been set and the research questions of this study as described in Chapter 1 fall into the category of problems for DSR. The problem is a real-world issue and a solution to the problem will bring many components that need to be scientifically grounded together. Moreover, there is a need to ensure that the solution is evaluated scientifically. The pragmatic philosophy will be adopted in this study to allow for a change between positivist and interpretivist philosophies as and when required. The adopted philosophy will influence the research approach, strategy and choice of methodology for this study as explained in the rest of this chapter.

## **2.3 Research Approach**

The research approach explains how a research project will relate to theory and the expected outcome of a study (Blumberg, Cooper and Schindler, 2011). Moreover, the adopted approach of a research project will determine the mode of enquiry, how the research will be designed, what sort of evidence will be collected and how knowledge will be developed from the collected evidence (Saunders, Lewis and Thornhill, 2009). Deductive and inductive approaches are the two major types of research approaches (Saunders, Lewis and Thornhill, 2009; Blumberg, Cooper and Schindler, 2011).

The deductive approach starts by stating a theory or hypothesis and then collects and analyses empirical data to confirm or refute such a theory or hypothesis. The inductive research approach, however, devises a theory at the end of the research based on observation and inference from an empirical analysis of data. However, (Saunders, Lewis and Thornhill, 2009) highlighted that there might be cases where a combination

of both approaches will be more appropriate than using just one of the approaches. The DSR, discussed later in this chapter, suggests that the research process starts with the definition of a problem followed by an iterative research process to find a solution to the problem (Johannesson and Perjons, 2012; Hevner, 2007). This study will be based on existing theories and therefore the approach for this study will be deductive.

## 2.4 Research Strategy

Research strategies are techniques that can be employed during a research process to achieve a set research objective and find answers to research questions (Saunders, Lewis and Thornhill, 2009). Therefore, the research question, objective and specific area of interest, would determine the appropriate research strategy or combination of research strategies that will be employed during a study. Moreover, the choice of research strategy could also be dependent on the choice of philosophy and research approach that were described at the inception of the study. Some possible strategies that can be employed in a research include (Collis and Hussey, 2009; Saunders, Lewis and Thornhill, 2009; Hofstee, 2006):

- **Experiments:** these are most suitable when the research involves investigation into relationships that exist between variables or the testing of theories or hypotheses under certain circumstances. It is required that the environment or circumstances under which the results of the experiments are recorded should be noted alongside the results. This is so that the results can be reproduced by anyone under the same conditions.
- **Surveys:** these are commonly used in the gathering of data for the exploration of descriptive studies that answer questions such as who, when, where, what, how much or how many. The data used can be primary data that is generated specifically for the study or secondary data that has been created previously for other purposes with the purpose of using the data to represent a larger population. Inferential or descriptive statistics can then be used to analyse the data in order to support or refute a theory or hypothesis.
- **Case study:** this is used for obtaining contextual in-depth knowledge by exploring one or a few cases in a structured manner. The selected case could

be an individual, a business process, an organisation or an event. Collis and Hussey (2009) added that a case study could be descriptive, illustrative, experimental or explanatory in nature.

- **Extended literature research:** this can be applied in a research to understand the scholarship about the subject matter. An extended literature review can be used to demonstrate and understand the various sub-components of a field. This can also be useful in synthesising sub-components from different fields as a fundamental explanation for inter-disciplinary concepts.
- **Action research:** this is used in studies that relate to bringing about change in an existing state of affairs. This could involve a collaborative effort between the researcher and other actors that are linked to the study. In which case, a collective effort is invested in the planning, implementation, learning and evaluation of the research project in an iterative process until the desired change is achieved.
- **Grounded theory:** this is focused on the building and development of theories through a systematic execution of procedures. When used, data can be collected without prior theoretical background, but rather theories can be developed by observing the collected data which is tested empirically thereafter.
- **Ethnography:** this is used to acquire knowledge for the purpose of explaining or describing concepts about a social group in the same way as members of such a group would do. Ethnography requires that the study be flexible and responsive to new patterns about the concepts that are being studied in order to be able to capture and interpret social observations from the group that is being studied.

This study will have an inter-disciplinary foundation and it is expected that knowledge will be drawn and synthesised from different fields to address the research questions and achieve the main objective. The research strategies that will be used in this study include literature research, surveys, experiments and case study. Each of the research strategies will be applied during this study based on the needs of the different DSR cycles. An understanding of the DSR and how it will be applied in this study will outline

how the research strategies are used in the different phases of this study as described in the next section.

## 2.5 Research Methodology

Design Science Research methodology has become a prominent choice of methodology for IS research (Carlsson et al., 2011; Vaishnavi and Kuechler, 2015). With its background in engineering (Venable, 2006; Beck, Weber and Gregory, 2013), DSR has found acceptance in research endeavours that involve the design/development of an artefact. DSR, however, differs from design research with respect to the type of problem to which it is applied, the solution created, and the extent of the level of its contribution to knowledge (Johannesson and Perjons, 2012). Design research is a broader study of the different fields of design, but DSR is focused on design as a research method for building an artefact in an iterative manner, where knowledge is produced and consumed during the process (Vaishnavi and Kuechler, 2015). In the IS context, DSR addresses difficult, practical and real-world problems that is of general interest to a wide group of people such as a community of practice (Hevner and Chatterjee, 2010; Johannesson and Perjons, 2012). Furthermore, DSR excels in providing innovative solutions that identify new opportunities.

Peppers et al. (2008) argue that DSR presents IS researchers with the principles, practices and procedures required for the creation and communication of groundbreaking IS research. The research agenda in the IS fraternity can arguably be categorised into the creation of “*knowledge for developing and improving IS-enabled initiatives; and knowledge for implementing and integrating solutions*” (Carlsson et al. 2011:111). DSR is able to achieve this because it provides IS researchers with a model and a guideline for creating, improving and evaluating artefacts that solve problems, affecting a practice which could even be un-theorised (Beck, Weber and Gregory, 2013; Vaishnavi and Kuechler, 2015).

Artefacts that result from a DSR project could include constructs, models, methods, instantiations or a combination of more than one artefact. In IS research, this could be a software solution, process, methodologies or intervention (Vaishnavi and Kuechler,

2015). An all-encompassing description of artefacts that can be generated from DSR is presented as “a thing that has, or can be transformed into, a material existence as an artificially made object (e.g. model, instantiation) or process (e.g. method, software)” by Gregor and Hevner (2013:341). This further confirms the suitability of the DSR as an appropriate methodology for the development of a framework to support trading decisions concerning grain commodities in South Africa. An iterative process is followed in the DSR process which enables the building and evaluation of resulting artefacts. Hevner (2007) presents a framework that shows the knowledge-creation activities of DSR as an integration of 3 cycles as shown in Figure 2.3.

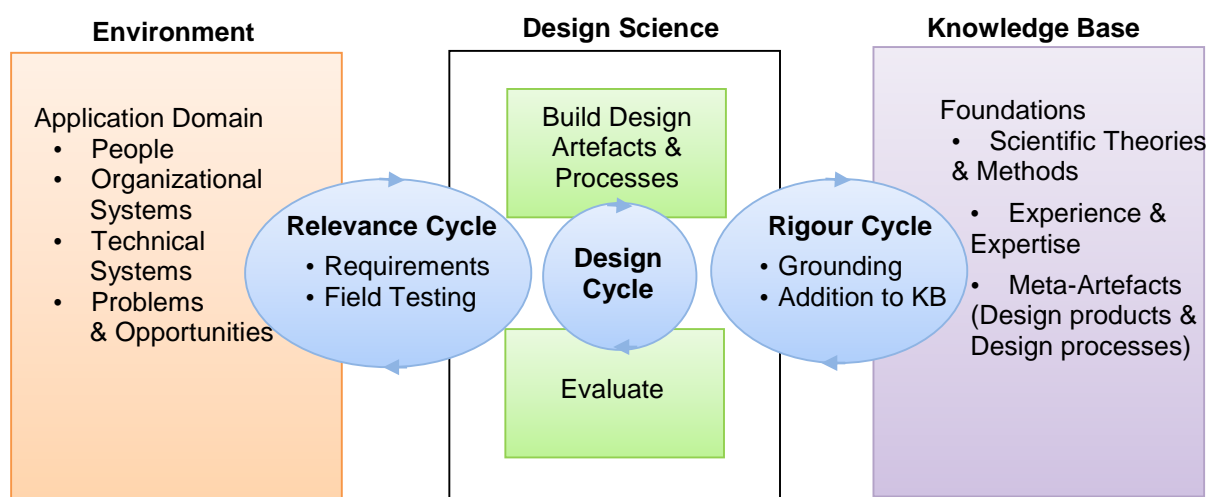


Figure 2.3: Design science research cycle (Hevner, 2007)

### 2.5.1 Relevance cycle

A DSR project is initiated with the relevance cycle by defining the problem or the opportunity for an innovative contribution within the environment where the research is applicable (Hevner, 2007). The relevance cycle also involves the definition of the requirements for developing the desired artefact and the standard of measurement of success of the artefact by taking the environment of the application into consideration. Subsequently, the artefact will be designed and evaluated in the other cycles of the research. Results from testing the artefact will then determine if an iteration of the relevance cycle should be revisited either for further understanding of the environment or the modification of the set question and objective to match the environment of the application (Hevner, 2007).

The focus of this study is to provide a framework that can be followed in developing a decision support system about trading grain commodities, specifically, for grain farmers in South Africa and perhaps for other industry stakeholders. Chapter 1 of this study provided the relevance of this study by defining the identified problem and opportunity that will be tackled in this study with respect to the grain commodities trading industry in South Africa. This study is particularly focused on the benefits of the framework that will be developed as the final artefact of this study into the practices of South African farmers of grain commodities, although the influence could be far-reaching. This will form the basis for designing and evaluating the outcome of this study.

### **2.5.2 Rigour Cycle**

The rigour cycle of the DSR provides the scientific grounding for the development of the artefact. During this cycle, the DSR draws from the knowledge base of relevant theories and methods to form the foundation for the research effort into the development of an artefact, which, in turn, leads to a contribution to the knowledge base (Hevner, 2007). livari (2007) noted that rigorous grounding of the development of an artefact in the implementation of DSR for IS research is what distinguishes such research from the building of an IT artefact in practice. However, the author adds that an IS practitioner that follows the same rigour will be following a DSR methodology.

The rigour cycles for the building of a DSR artefact could draw on and synthesise scientific theories and methods from a knowledgebase in a different field. However, Hevner (2007) contends that limiting the grounding of the building of artefacts with DSR to existing scientific theories and methods could hinder creativity and innovation, which are the hallmarks of using a DSR in a field such as IS research. Hence, the rigour cycle in a DSR can be extended to previous artefacts, analogies and metaphors (such as Neural Networks algorithms) and other sources of creative insights (Hevner, 2007; livari, 2007). This cycle also influences the design and evaluation of the outcome of the DSR such as the relevance cycle, and an iterative process could also ensue during the research process. The purpose of the iteration will be to ensure that the

resulting outcome solves the problem explicated in the relevance cycle and also takes the application environment into consideration.

A literature study will be carried out to explore the scientific theories and methods in relevant fields that provide grounding for the support of decision making regarding the trading of grain commodities in South Africa. The literature study will also cover the environmental requirements by taking the trading practice in grain commodities into consideration. A survey will be carried out among farmers and traders of grain commodities in South Africa to contextualise and confirm that the findings from literature are consistent with the realities of grain commodities trading in South Africa. A further literature study will also be carried out with the aim of synthesising knowledge from different fields of interest to determine the technical requirements of a Decision Support System (DSS) that the resulting framework can be used to build. This study will rely on the use of computer algorithms for decision making. Hence, attempts will be made to find a point of reference from past research and in implementations that have successfully made use of such algorithms in areas that are relevant to this study and how such algorithms have been used to get desired results.

### **2.5.3 Design cycle**

The design cycle of a DSR methodology is where new knowledge is produced through the building and evaluation of the artefact(s) (Hevner, 2007; Vaishnavi and Kuechler, 2015). Both the building and the evaluation activities of the design cycle are dependent on the rigour and the relevance cycles. The outcome of the relevance cycle will form the initial input at the inception of the research and subsequently as a guideline for iterations during design and evaluation. Furthermore, the rigour cycle provides the scientific basis for any theory, method, tool, technique, or process which are used in the building and evaluation of the desired artefact.

The main artefact that is expected from this study is a framework in the form of an abstraction of knowledge for the implementation and integration of components to solve problems and provide opportunities for grain commodities farmers in South Africa. The use of DSR as the methodology of choice for the design of such an artefact has found scientific grounding in the literature (Carlsson et al., 2011; Gregor and

Hevner, 2013). This study will rely on a literature study, new developments regarding the concepts of Big Data, Data Science and previous use of analytical techniques as the scientific grounding for the design of the framework.

In order to evaluate the artefact, a case study will be carried out by implementing the framework with specific focus on the trading of white maize in South Africa. Experimental implementation of the desired framework will be carried out in order to evaluate the validity and usefulness of the artefact. Furthermore, experts that are traders will be invited to make predictions about prices of white maize. These will be compared to price predictions from the experimental implementation of the framework that will be proposed within the ambits of the relevance and the rigour cycles of this study. Necessary iterations of the design and evaluation stages will be carried out, together with the development of the required rigour, and where necessary the relevance of this study.

#### **2.5.4 Design science research process**

The DSR cycles described in Sections 2.5.1, 2.5.2 and 2.5.3 provide the guideline and the overall methodology for the application of Design Science. However, different models that provide more details about Design Science activities and process have been proposed by different authors, although there are similarities. Peffers et al. (2008) have proposed a DSR process in an iterative model comprising six main phases as presented in Figure 2.4.



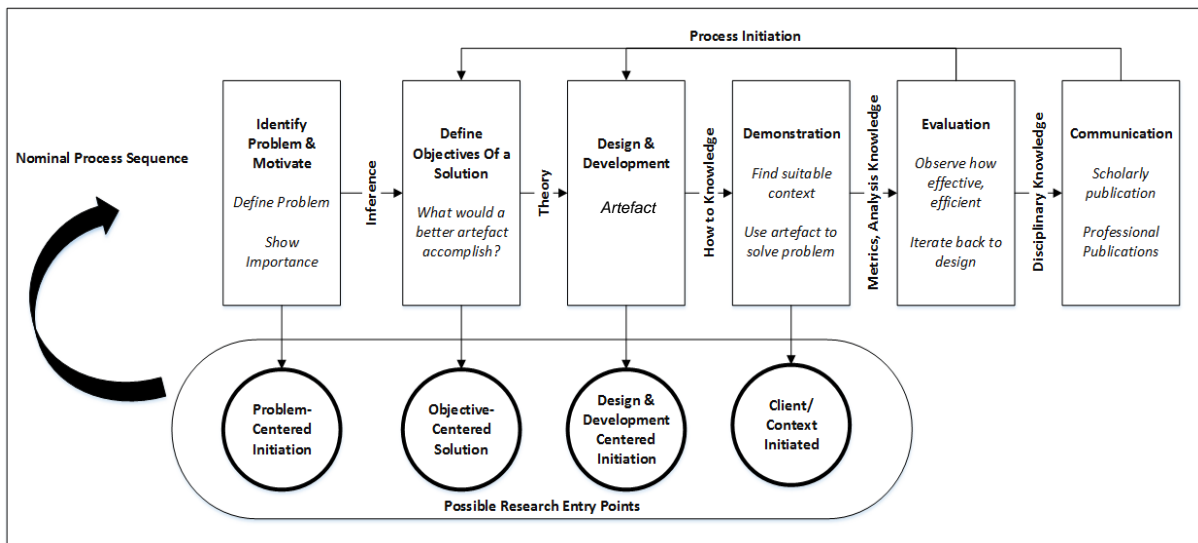


Figure 2.4: DSR process model (Peffer et al., 2008)

The phases in the model include the problem definition and motivation phase that is followed by the phase where objectives and research questions are set. This is followed by the development/design phase where the actual artefact is designed based on theory or other scientific rigour as suggested in the DSR cycle by Hevner (2007). The artefact is demonstrated in the fourth phase of the process by applying it in an environment that represents the practice for which the artefact has been developed. In this demonstration, the success of the artefact is then measured in the fifth phase for applicability, efficiency and effectiveness. This could lead to an iteration in the process; either back to the second phase to review the research questions and objective or to the third phase to redesign the artefact.

The Design Science researcher needs to ensure that the iterations, where necessary, are well grounded in research rigour, and if an iteration is based on creativity, there is still a need for a grounded scientific motivation (Hevner, 2007; Peffer et al., 2008; Johannesson and Perjons, 2012; Vaishnavi and Kuechler, 2015). The last phase of the process is where the knowledge created during the research process is communicated to a relevant community for the purpose of feedback. Such feedback in turn, could also trigger a new iteration for further improvements of the artefact, which could start either from the definition of the research objectives/questions or from the design phase of the process (Peffer et al., 2008). The DSR model by Peffer et al.

(2008) indicates that the DSR research process could be initiated from any of the initial four phases depending on the nature of the problem, the researcher or the circumstance that triggers the research.

Figure 2.5 shows a variant of the DSR process model by Vaishnavi and Kuechler (2015), which reflects the phases in the DSR process as described by Peffers et al. (2008) shown in Figure 2.4, but also incorporates the DSR cycle by Hevner (2007). The emphasis of the model in Figure 2.5 is that DSR should create new and interesting knowledge as much as the process is expected to find proper grounding in the knowledgebase (Vaishnavi and Kuechler, 2015).

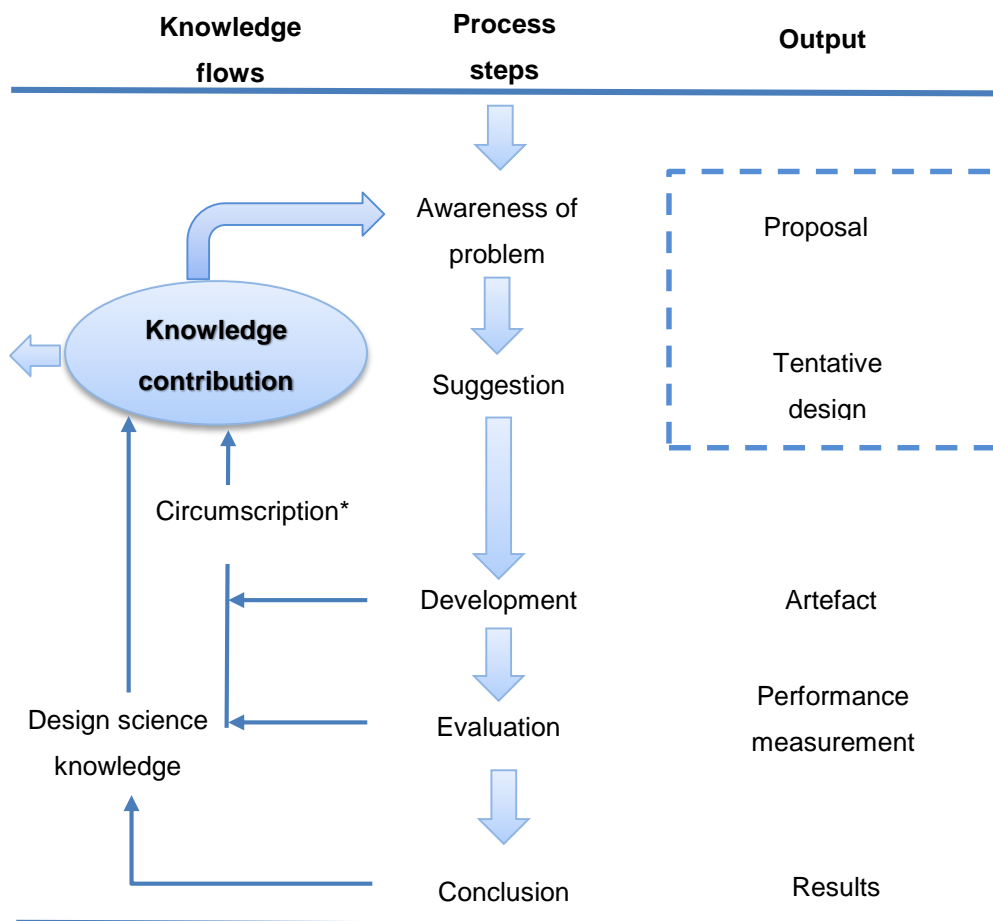


Figure 2.5: DSR process model (DSR cycle) (Vaishnavi and Kuechler, 2015)

The model shows two streams in which the DSR process contributes to the knowledgebase. It shows that there is a learning that contributes to the

knowledgebase as a result of circumscription. This is when iterations in the development or the evaluation phase become necessary when things do not work out as initially expected and perhaps as stated in theory. Secondly, it shows that there should be a contribution to the body of knowledge at the concluding phase of the research which has the potential to initiate new or further research.

In addition to the contribution to the flow of knowledge in the DSR, the model proposed by Vaishnavi and Kuechler (2015) as shown in Figure 2.5 also highlights the output from each of the phases in the research process. The authors propose that a prototype that demonstrates a proof of concept can be developed at the early stage of the process, particularly to inspire creativity during the rest of the DSR process.

The envisaged DSR cycles and process for this study are presented in Figure 2.6. It is presented as an adaptation from the DSR cycle by Hevner (2007) in Figure 2.3 and the DSR process as suggested by Peffers et al. (2008) and Vaishnavi and Kuechler (2015) in Figures 2.4 and 2.5 respectively. The research design in Figure 2.6 takes into consideration, the research philosophy, the adopted research approach and the research strategies that have been adopted for this study.

### **Relevance Cycle**

Figure 2.6 presents the envisaged process, activities and strategies employed during each activity for this study, based on the DSR process and methodology as suggested and adapted from (Peffers et al., 2008; Johannesson and Perjons, 2012). The figure shows that Chapter 1 has been used to explicate and analyse the research problems. The context of the problem in the industry and the opportunities that a solution could offer through the study of existing literatures were discussed in Chapter 1, thereby providing the relevance for this study. Based on this analysis, research questions and objectives that will guide the rest of this study have also been set.

### **Rigour Cycle**

The second activity in the adapted DSR process that is shown in Figure 2.6 outlines the artefact and identifies requirements for the artefact. This activity will span Chapters 3, 4 and 5 of this study. Chapter 3 will explore decision making in business, decision

support requirements for grain commodities trading in South Africa and the basic components of a DSS. Thereafter, Chapter 4 will address the role that the Big Data concept and approach can play in supporting decisions regarding the trading of grain commodities. Chapter 5 will discuss the modelling techniques and related issues in predicting grain commodities prices. Chapters 3, 4, and 5 will form the rigour cycle for this study and it is expected that the rigour cycle will achieve the first research objective (RO<sub>2</sub>) of this study as depicted in Figure 2.6.

A literature study will be carried out for most of the three chapters in this cycle to identify theories, ideas, methods and past experiences that can be used as the foundation for the development of the artefact of this study. Moreover, surveys will also be carried out among farmers and traders to confirm that their requirements are captured in the literature study. It is expected that the rigour cycle in Chapters 3, 4 and 5 will produce the components that can be used to develop a DSS framework for trading in grain commodities.

### **Design Cycle**

The design cycle of this study will be implemented in Chapters 6 and 7. It will involve the actual development of the artefact in fulfilment of the third research objective (RO<sub>3</sub>) of this study. Based on the findings from Chapters 3, 4 and 5, a framework will be developed and proposed in Chapter 6 as an abstraction of a DSS for trading grain commodities in South Africa. Other activities of DSR that will be carried out during this cycle are the demonstration and the evaluation of the artefact based on the fourth research objective (RO<sub>4</sub>).

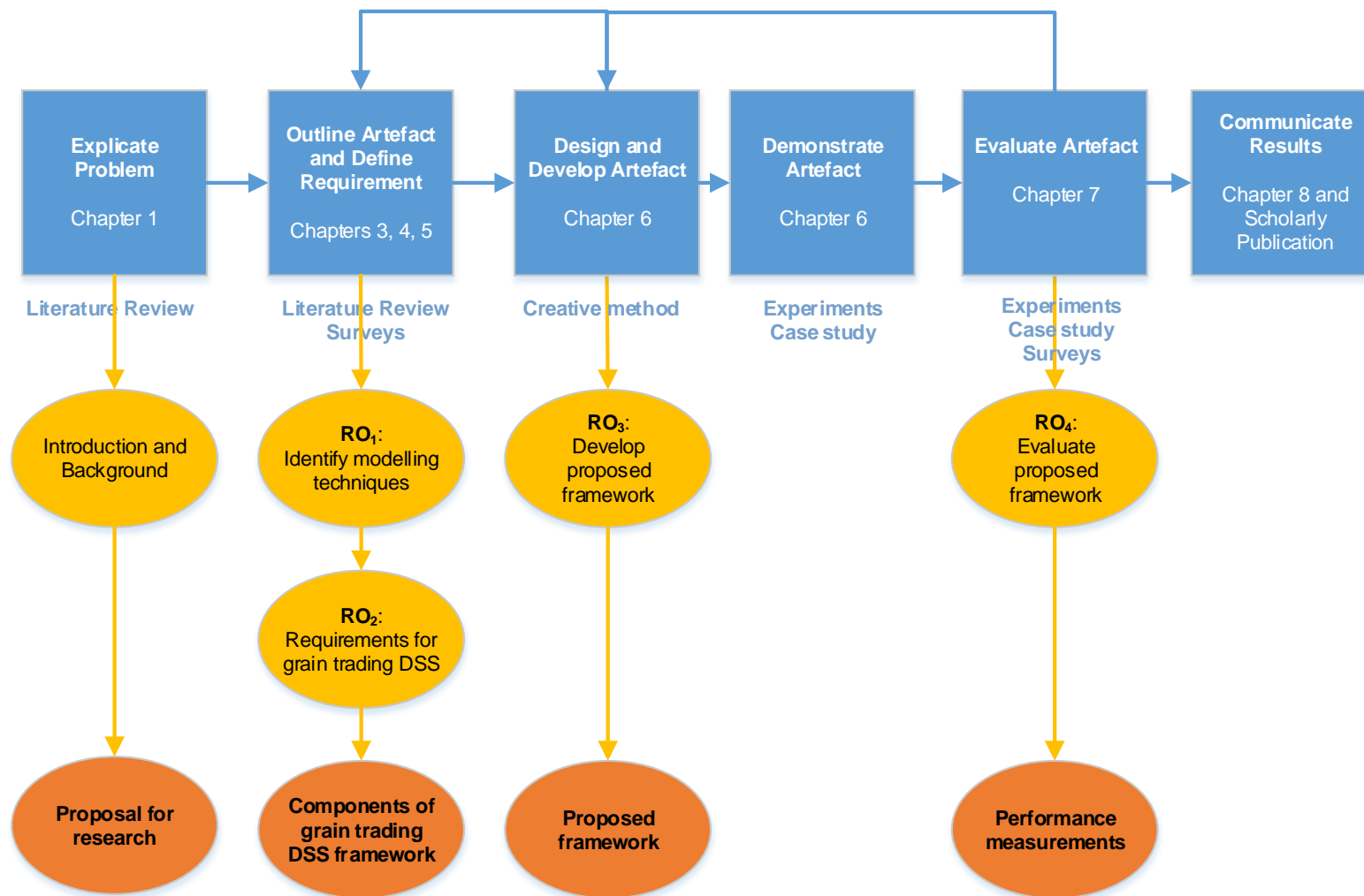


Figure 2.6: Envisaged DSR process and cycle for this study (Adapter from Peffers et al. (2008) and Johannesson and Perjons (2012)).

Chapter 6 will demonstrate the feasibility and applicability of the framework that will be proposed by using the spot and future contract trading of white maize as a case study. Furthermore, the technical abilities and how useful the framework is in supporting grain commodities decision making will be evaluated in Chapter 7. This will be done by comparing predictions made by using an implementation of the framework with price predictions made by a panel of experts that provide decision support about trading in grain commodities, based on their experience and intuition. It is expected that the design cycle at the demonstration and evaluation stage could result in an iterative process leading back to the definition of the requirements for the framework (artefact) or the actual building of the artefact.

The findings, results and contributions of this study will be presented in Chapter 8 of this thesis. These will also be presented to stakeholders by means of academic publications for feedback. It is expected that the feedback received could help in further improving the outcome of this study or could lead to new studies.

## **2.6 Ethical Considerations**

Research ethics have to do with making sure that a research project is conducted in such a way that the design of the research, the collection/usage of data and presentation of results is done with the highest moral standards and a sense of responsibility (Saunders, Lewis and Thornhill, 2009). Ethical considerations during a research project require that the selection and execution of methods during a research process be such that people or organisations are not left vulnerable as a result of the research. Whether the research philosophy is positivism, interpretivism or otherwise, it is still important that researchers carry out their studies in a way that the research output is not misleading.

Some of the major ethical considerations in research include ensuring that participation by third parties is voluntary and the right to withdraw from the study is maintained. Other ethical considerations include seeking of consent from participants with full disclosure but ensure that confidentiality is offered and maintained as much as possible. Moreover, research should be conducted in such a way that the people

who participate in the research are not pained, shamed, harmed or embarrassed (Collis and Hussey, 2009; Blumberg, Cooper and Schindler, 2011). This can also be extended to ensuring that organisations or individuals that partake in a research do not end up with financial or any type of losses in the process.

This study will rely on access to data from several sources and this will be done with permission where necessary. In all cases, attention will be given to ensure that no organisation or individual will be harmed or suffer loss as a result of this study. The surveys conducted in this study will be carried out within the ambits of the ethical clearance granted by the ethics committee of the Faculty of Science of NMMU with reference no: H14-SCI-CSS-12 (Appendix A). All the survey data will be anonymised and the identity of the panel of experts for the evaluation activity will be concealed.

## **2.7 Summary**

An overview of the research design and methodology which serve as the plan for how a research project can be structured and executed was presented in this chapter. After the identification of the research problem, there is a need to choose a suitable research philosophy that will guide the research approach, strategies and methods that will form and guide the execution of the research process. By using the “research onion” as suggested by Saunders, Lewis and Thornhill Adrian (2009), the link between the research philosophies, approaches, strategies and methods was explained.

The problem identified in this study requires the development of an artefact that is scientifically grounded and evaluated, which is able to solve a real-world problem. Therefore, the DSR methodology was adopted in this study. The DSR follows structured and systematic relevance, rigour and design cycles in solving research problems. The general idea is that the DSR process goes through an iteration of six phases in providing a solution to a research problem. The DSR process starts by motivating the problem or opportunity within the context of the application of the solution, which makes DSR even more relevant to IS research. Thereafter, the requirements for developing the artefact are gathered and the artefact is designed in the third phase of the process. This is followed by a demonstration of the artefact’s

ability to perform and then the evaluation phase that is a measurement of how well the artefact solves the identified problem follows. The demonstration and evaluation phases may require going back to the second and third phases of the process in order to ensure that the artefact provides a robust solution. Finally, the outcome of the research is communicated to the relevant stakeholder(s) who will provide feedback on the expected outcome of this study.

This study will adopt a pragmatic viewpoint which allows for a multiple philosophical outlook during the research process depending on the stage and need of the research. The iterative nature of the DSR process is considered suitable and appropriate for this study based on the research questions and the objectives that have been set in Chapter 1 of this study. The next chapter will start with the outlining of the requirements and the components that will eventually make up the final artefact of this study.



# Chapter 3 : Decision Support for Grain Commodities Trading

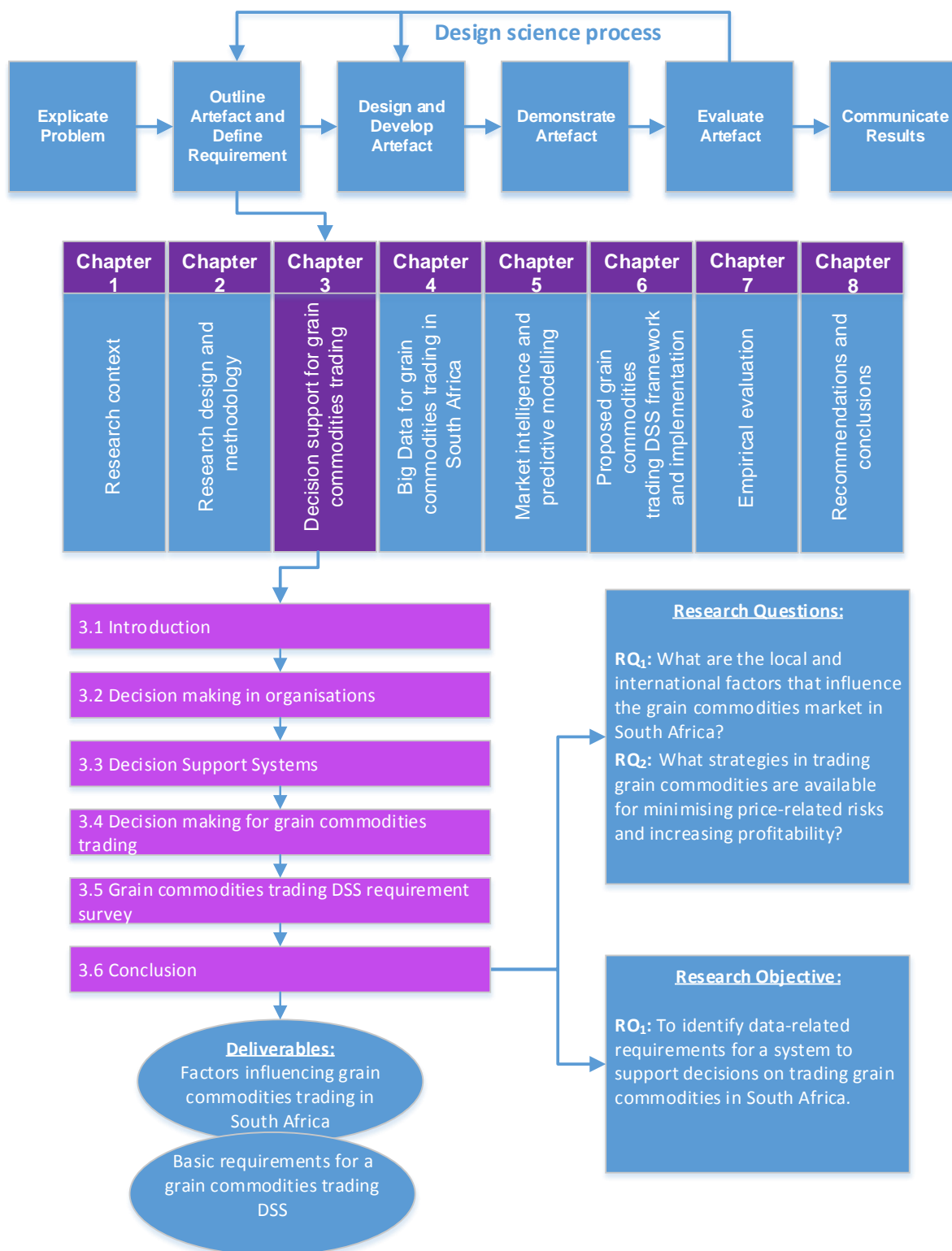


Figure 3.1: Chapter outline and deliverables

### 3.1 Introduction

The previous chapter explained the relevance of design science methodology for this study and how the design science process will be implemented. This chapter will focus on problem identification and research motivation within the design science process as described in Chapter 2. It will define and motivate the opportunities for the use of a system to support decisions made in grain commodities trading.

Information and Communication Technology (ICT) have played an important role in the decision making process within organisations for decades. ICT has played a significant role in enabling organisations to compete properly, respond swiftly to the ever-changing business environment and increase productivity (Davenport and Harris, 2007; Brynjolfsson, Hitt and Kim, 2011; Tambe, Hitt and Brynjolfsson, 2012). The role of ICT for decision making in organisations includes the provision of support systems that enable the collection, organisation, processing and application of data in the decision-making process. However, the evolution of several areas within the CS and IS spheres present new frontiers for the use of these fields in decision making (Roberts, 2008). The concept and evolution of data has played an important role in decision making within organisations over the years.

Data can be defined as a representation of facts that can be collected, recorded and used as a basis for decision making (Collis and Hussey, 2009; Elmasri and Navathe, 1989). Therefore data forms the basis of collecting, organising and describing facts and information. These facts and information form the basis of the strategic, tactical and operational activity of organisations. The place of data has become very important, because the availability of the right type, volume and optimal use of data can provide leverage and increase profitability (Brynjolfsson, Hitt and Kim, 2011; Davenport and Harris, 2007).

However, the amount of data generated globally has increased significantly (Manyika et al., 2011). Hence, the emergence of the Big Data, Data Science and Business Analytics concepts. The new concepts have been described as posing challenges and opportunities that will affect productivity, profitability and efficiency. They also

introduce a complete paradigm shift in the way many things are seen and done (Manyika et al., 2011; Mayer-Schonberger and Cukier, 2013). This chapter will explore the decision making about trading in grain commodities in South Africa with respect to the use of market intelligence that is provided by Decision Support Systems. Thereafter, this chapter will consider the different grain trading strategies and how the availability of the correct data can be used in selecting the most advantageous grain commodities trading strategy. Furthermore, this chapter will consider the factors that influence the grain commodities market in South Africa and the sources of data for each of the factors. Finally, the requirements of a support system for decisions in trading grain commodities will be discussed.

The first research objective (RO<sub>1</sub>) of this study is *“to identify data-related requirements for a system to support decisions on trading grain commodities in South Africa”*. This chapter will fulfil this objective by providing answers to the following research questions:

- **RQ<sub>1</sub>:** What are the local and international factors that influence the grain commodities market in South Africa?
- **RQ<sub>2</sub>:** What strategies in trading grain commodities are available for minimising price-related risks and increasing profitability?

The DSR paradigm requires that the practical relevance of solving the problem concerning a practice should be established at the beginning of the research process (Johannesson and Perjons, 2012; Beck, Weber and Gregory, 2013). Furthermore, there is a need to demonstrate that the research will solve a real-world problem relating to a practice of interest. Thus, Chapter 3 will describe the problem and the related opportunity that has been identified, with grain commodities trading in South Africa, particularly among the grain farmers, as the practice of interest (Hevner, 2007).

Beck, Weber and Gregory (2013) broadly divide the DSR process into the two categories of building and evaluating an artefact and suggested that the building process is a series of tasks with the objective of producing a new solution. Besides establishing the relevance of this study for the grain commodities trading practice in

South Africa, this chapter will also begin the laying of the scientific foundation for the envisaged solution that will be offered by this study. The objectives of the proposed artefacts will be defined in this chapter and attempts will be made to answer the question – “what would a better artefact accomplish” (Peppers et al., 2008). Therefore, this chapter will use a combination of a critical review of literature and the result of interviews with respondents that are involved in grain commodities trading to identify requirements for a grain commodities trading decision support system. This will determine the initial features of the proposed artefact. However, because of the iterative nature of the design science research methodology, the task of defining the requirements and features of the artefact might be revisited if the need should arise. Hence, this chapter will serve as the relevance cycle of the DSR process which will also link to the rigour cycle that will also be initiated in this chapter.

Section 3.2 will provide a theoretical overview and managerial implications, of decision making for businesses. It will introduce the role, opportunities, challenges and the process of decision making in organisations. Section 3.3 will provide a review of literature on the role of technology in decision making in organisations. Section 3.4 will explore the emergence and evolution of data as a decision-support concept and the new paradigm that is promoted by Datafication and Big Data. This section will further review the place of Decision Support Systems in grain commodities trading in South Africa with reference to the use of data as a key component.

Based on the review of literature and the result of a conducted survey, Section 3.5 will identify the requirements for the development of a data-driven Decision Support System for grain commodities trading in South Africa. Thereafter, Section 3.5 will also propose a conceptual framework for making decisions in trading grain commodities that is based on the concepts of Big Data and analytics. Finally, Section 3.6 will provide a summary of Chapter 3 with an outline of objectives and deliverables that have been achieved.

The next section provides the background and theoretical framework for data-driven decision making in organisations.

## 3.2 Decision Making in Organisations

Making decisions is an important part of the responsibilities of business owners, executives and managers. This is because the livelihood of their business or organisation depends on their ability to make the right decisions at the right time (Rogers and Blenko, 2006). The need to make the right decision is even more important because of the need to adapt quickly in the rapidly changing economic, political, technological and social environment. High quality decision making has been linked directly to increased efficiency, productivity and profitability (Brynjolfsson, Hitt and Kim, 2011). But poor, untimely or a total lack of decision making in organisations could lead to the loss of revenue, opportunities or other dire consequences (Davenport, 2009; Bazerman and Chugh, 2006). Hence, high-performing organisations are known to take a holistic approach to decision making. The following factors have a direct impact on the ability to make effective decisions in an organisation (Davenport, 2009);

- Technology,
- Information,
- Organisational structure,
- Methods,
- Personnel.

However, the availability of any type of resources does not necessarily guarantee effective decision making within an organisation. But, it is important to be able to synchronise and create a synergy among these factors to get the best out of them (Courtney, Lovallo and Clarke, 2013). Getting the best out of resources that are supposed to support decision making in itself requires making right decisions. But the circumstances that surround decision making and the environmental factors that influence decision making make it a complex task, especially within organisations. Hodgkinson and Starbuck (2008) noted that decision makers are constantly faced with situations where they have to make complex decisions using incomplete information and amidst uncertainty and ambiguity. Hence, there is a need to continuously study and improve an organisation's decision-making systems. The next section of this chapter reviews previous studies on theories relating to the concept of decision making for the purpose of improvement.

### **3.2.1 Decision making theory**

The theoretical foundation of decision making in businesses and organisations has its roots in novel studies carried out several decades ago, but still relevant in today's body of knowledge. This is because decision-making theory provides key understanding of how organisations function (Kowalczyk and Buxmann, 2014). Studies on the theory of decision making suggest that the key responsibility of business leaders and managers is decision making (Hodgkinson and Starbuck, 2008). Therefore, the understanding of decision making from the theoretical and practical perspective could help decision makers in organisations and businesses to acquire and deploy resources effectively to maximise utilities or increase value of the organisation.

One of the salient findings of the theory of decision making is that decision making is expected to be an ordered process task. Simon (1960) categorised decision making into the three phases of identifying the opportunity to make a decision, identifying possible alternatives and the phase where a selection is made from the available list of alternatives. Other research work has presented different ideas that elaborate on steps taken during decision making which still fit into the categorisation by Simon (1960). Table 3.1 presents a diagrammatic representation of how the decision-making steps as suggested by Bazerman (2006) and Hammond, Keeney and Raiffa (1999) fit into the suggestion of Simon (1960).

A comparison of the decision-making steps suggested by different authors in Table 3.1 emphasise that decision making is a process. However, it also shows the rationality of the decision-making process. This implies that the decision maker follows a logical and explicit process that is based on information at the disposal of the decision maker, by considering the possible eventualities and before selecting the optimal option that is informed by a scale of preference that is consistent (Bazerman, 2006).

Table 3.1: Phases and steps in decision making

	Simon (1960)	Hammond, Keeney and Raiffa (1999)	Bazerman (2006)
<b>Phase 1</b>	Identify opportunity for decision making	1. Define the problem	1. Identify the right problem 2. Understand consequences
<b>Phase 2</b>	Identify possible alternatives	2. Identify the criteria 3. Weight the criteria 4. Generate alternatives	3. Define objectives 4. Identify alternatives
<b>Phase 3</b>	Select a course of action	5. Rate each alternative 6. Determine the optimal decision	5. Identify trade-offs 6. Clarify uncertainties 7. Consider the risks 8. Identify optimal decision and consider linked decisions

However, the complete rationality of the decision-making process is unrealistic in practice (Simon, 1997). Hence, the assumption of the “bounded rationality” introduced into the decision-making theoretical framework (Simon, 1997). The bounded rationality assumption agrees that decision making should follow a logical process that results in an optimal choice being made (Bazerman, 2006). But it also acknowledges the limitation of incomplete information, the capacity of the decision makers and limitations of the context or the environment where the decision is being made (Simon, 1997).

It has been identified that decision makers function relying on factors such as intuition, experience and emotions (Sauter, 2010). This is not just because of incomplete information, ambiguity and complexities that characterise many of the decision-making processes in organisations (Hodgkinson and Starbuck, 2008). But Simon (1997) asserted that the decision-making process is bounded by the limitations of the human mind to store and process information especially in a complex situation. Hence, decision makers make “satisficing” decisions where a choice is based on the limitation of the decision maker or other external factors instead of the optimal decision (Simon, 1979).

Decades after its initial proposition, the discussion on the bounds of rationality has continued among researchers. However, the assumption has not gone uncriticised, especially by studies that considered rationality by looking at it from the perspective of many decision makers that are responding to the same problem under the same circumstance (Hodgkinson and Starbuck, 2008). Many organisations, however, are still confronted with limitations in their decision making as proposed by the bounded rationality assumption (Courtney, Lovallo and Clarke, 2013). Hence the evolution of information, processing theories, systems and innovations are important factors in the theory and practice of decision making.

### **3.2.2 Information processing theory**

Scholarship on the organisational information processing theory provides further insight into the management and usage of information as the backbone of decision making in organisations. The theory suggests that the flow of information could provide solutions for the bounds of rationality in decision making. The theory of organisational information-processing proposes that the availability, processing and management of information can be used to deal with the issues of uncertainty and equivocality in structuring an organisation and its decision-making systems (Galbraith, 1974; Daft and Lengel, 1986).

Uncertainty in decision making is described as a result of the lack of sufficient information. This can be remedied by increasing the amount of information available for decision making and organisational design (Kowalczyk and Buxmann, 2014). However, equivocality in organisations and their decision-making system refers to having ambiguity, lack of comprehension of the problem at hand or there is a lack of information available to offer solutions (Daft and Lengel, 1986). This might explain why decision makers resort to intuition, emotion or their past experience when they have to make a decision when uncertainty, ambiguity and equivocality exist, especially when there is a degree of complexity in the situation also. But it might be safe to assume systematic processing of the information available can address the problems of uncertainty and equivocality.



However, in dealing with uncertainty, the acquisition of more information, as decision making becomes complex, introduces a new problem of information overload for decision makers. The theory of organisational information processing suggests that organisations can either reduce the amount of information that needs to be processed all the time through the design of the organisations' structure or the organisations can increase their capacity to process information (Galbraith, 1974). To solve the problem of information overload, decision making about a functional area within an organisation or business is sub-divided to reduce the amount and the complexity of information that decision makers have to deal with.

Daft and Lengel (1986) elaborate further on the position of Galbraith (1974) on the solution to uncertainty in decision making as finding ways to deliver the right amount of information to decision makers all the right time. However, Daft and Lengel (1986) suggest the solution to equivocality lies in the ability to process rich information. Information richness is defined "*as the ability of information to change understanding within a time interval*" (Daft and Lengel 1986:560). This suggests the availability of quality information that provide clarity at the right time.

In combining the solution for both uncertainty and equivocality together, Daft and Lengel (1986) suggested the use of a structural mechanism that comprises rules and regulations, formal information system, special reports, planning, direct contact, integrators and group meetings in a continuum as shown in Figure 3.2. Effectively, the purpose of this continuum is to facilitate the delivery of the optimum amount of information and provide the right level of clarity desired. This depends on the degree of uncertainty and the level of equivocality in the decision-making process as shown in the continuum presented in Figure 3.2.

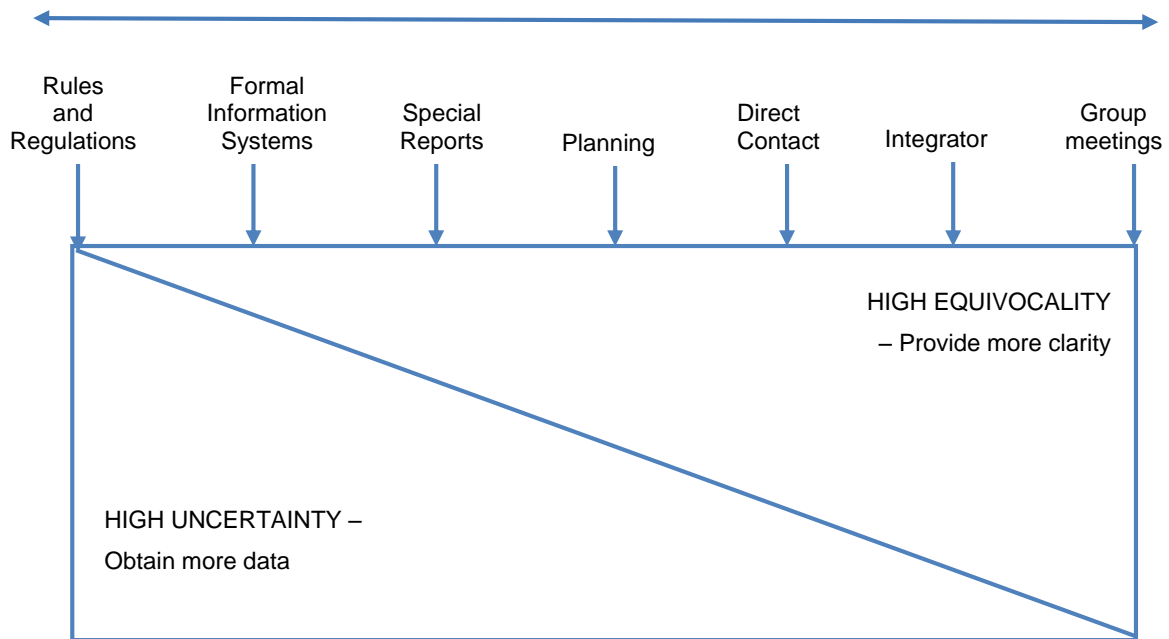


Figure 3.2: Role of information in reducing uncertainty and equivocality (Daft and Lengel, 1986)

The information-processing theory can be linked to the efforts of Information Systems researchers and practitioners. Over the years, there has been an increasing attention given to Information System research and innovation that help organisations to improve decision making (Davenport, 2010). There are increasing opportunities to use information and data to improve decision making. Hence, the emergence of organisations that are structured and designed based on the use and flow of data described as data-driven organisations (Davenport, 2010; Brynjolfsson, Hitt and Kim, 2011; Kowalczyk and Buxmann, 2014). It is therefore important to examine the role of data-centric tools and principles in organisational decision making and how they link to the theory.

Kowalczyk and Buxmann (2014) presented a variant of the role of information in reducing uncertainty and equivocality continuum. The diagrammatic presentation in Figure 3.3 shows a new information-processing mechanism with a new perspective to the organisational information processing theory. The diagram introduces four data-centric mechanisms with the ability to reduce uncertainty and equivocality.

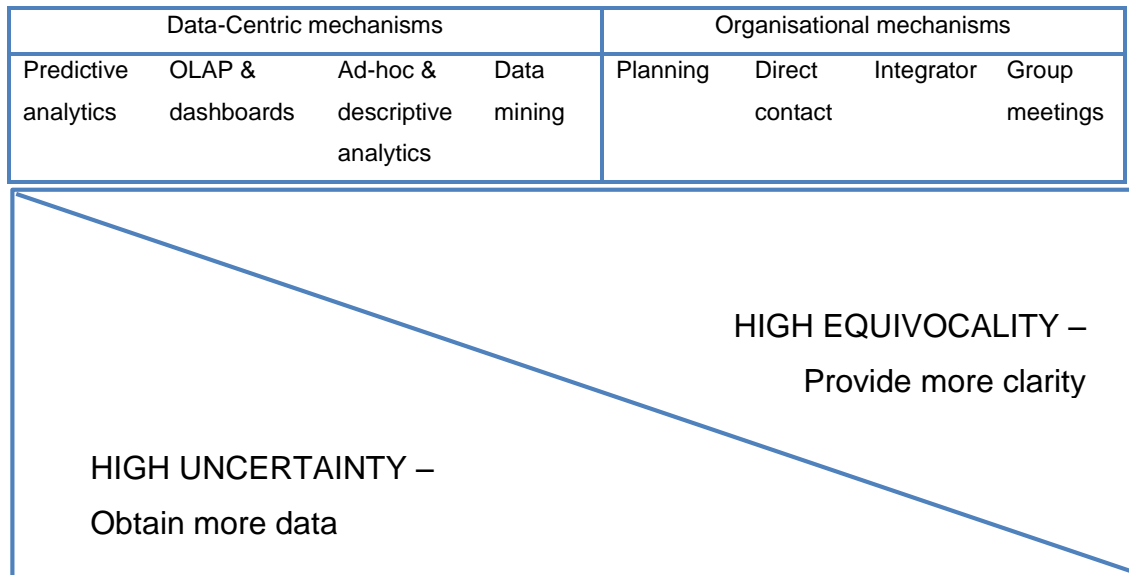


Figure 3.3: Data-centric and Organisational mechanism based information processing (Kowalczyk and Buxmann, 2014)

The theoretical framework that has been discussed in this section will form the foundation of this research. This study will examine the trading decisions on grain commodities trading made by grain farmers in South Africa from the perspective of the theory of decision making and the information-processing theory in organisations. Specifically, this study will consider the components of the framework proposed in Figure 3.3 as the foundation for structured information processing for grain commodities trading by grain farmers in South Africa. Furthermore, this study will examine how each of the evolving areas of Business Intelligence, Business Analytics, Big Data and Data Science combine with the organisational information-processing theory for a framework that provides decision support for grain trading decision making in grain trading South Africa.

### 3.2.3 Improving decision making

The problem of uncertainty, equivocality and ambiguity identified earlier can be categorised as organisational. But, it is also important to note that the personality and disposition of the decision maker can also affect the decision making. The problem of the decision maker may be solved by relying on experienced decision makers

(Bazerman and Moore, 2013). However, an excellent decision maker is unlikely to be successful with a faulty decision-making system, incomplete and confusing information. It is more likely to expect decision makers to be more successful with a systematic decision-making process and necessary support including those elements that reduce uncertainty and equivocality (Davenport, 2009, 2010; Courtney, Lovallo and Clarke, 2013). Hence, Bazerman (2006) argued that the focus of improving decision making should be on making the decision-making process and supporting tools better.

Over the years, research, innovation and development in the fields of Computer Science and Information Systems have been at the centre of information processing and support. Different research into how decision making can be improved in organisations have mentioned the use of technology as a key factor for improving the decision-making system (Davenport, 2009; Courtney, Lovallo and Clarke, 2013). The adoption and implementation of Computer Science and Information-System tools that support the collection, management, processing, presentation and the use of data have been at the forefront of improving decision making for decision makers.

Computer Science and Information System tools enable better communication in organisations through enhanced flow and effective sharing of information at reduced costs. These tools and systems have evolved over the years from merely data-collection systems to tools that decision makers can use to predict the future (Sauter, 2010). With the emergence of Big Data, the amount of data produced and available is increasing considerably. Furthermore, there has also been an evolution of decision-making systems like Business Intelligence, Business Analytics and recently the emergence of Data Science (Chen, Chiang and Storey, 2012; Dhar, 2013). These concepts and tools are able to improve the process and the outcome of decision making (Brynjolfsson, Hitt and Kim, 2011; Kowalczyk and Buxmann, 2014). As a result, empirical research suggest that systems based on these concepts and tools are topmost technology investment priority for organisations (Gartner, 2015). Research by Brown, Sikes and Willmott (2013) of McKinsey and Company indicate that organisations are turning to Big Data, Business Intelligence and Analytics to improve their decision-making processes.

It can therefore be deduced that the use of Computer and Information System based concepts, tools and systems that enable the effective use of information and data is an important factor in improving decision making. The next section of this chapter will review the evolution and the impact of decision support systems that are based on the development of Computer Science and Information Systems.

### **3.3 Decision Support Systems**

Decision Support Systems (DSS) can be described as the foundation for the tools, principles, practice and research of the use of computer-based interventions and Information Systems for decision making (Sauter, 2010; Delen and Demirkan, 2013). Over the past decades, the tools, technologies, driving forces, implementation environments and approaches that support DSS have evolved. But there is a need to use Computer-based and IS support for solving complex problems, planning, management and decision making (Sauter, 2010).

The focus of using Computer-based and Information Systems for decision making is primarily for the gathering, sorting, manipulation and extraction of valuable information from data. Over the years, there has been an emergence of concepts and technologies that support decision making such as Data Warehousing, Data Mining, Business Intelligence, Business Analytics and Data Science. However, DSS remains a common denominator among these concepts, tools and technologies.

Decision support systems are described as *“computer-based systems that bring together information from a variety of sources, assist in the organization and analysis of information, and facilitate the evaluation of assumptions underlying the use of specific models”* (Sauter 2010:5).

In examining the scholarship of DSS, Carlsson and Turban (2002:105) further identified DSS as:

- *Methods and instruments for dealing with unstructured or semi-structured problems, which formed an improvement on management science and operations research methodology;*
- *Interactive computer-based systems, which were built for managers and were more advanced than descriptive systems theory or traditional models;*
- *User-oriented systems, which formed a better platform for decision making than batch-oriented MIS applications; and*
- *The separation of data and models in computer applications, which form the basis for more effective modelling.*

A number of salient themes can be observed from all the perspectives of DSS that have been mentioned, these include the use of computers, decision making, problem solving, the use of information/data, and the development and use of complex models. These identified themes further emphasise the relevance of DSS for decision making in organisations today and suggest a relationship with the new concepts of Business Intelligence, Business Analytics, Big Data and Data Science. This is because the identified themes resonates with the underlying principles and themes that govern the recent developments in the application of Computer Science and Information Systems in organisational decision making.

The availability and ubiquitous access of large volumes and varieties of data that are being created at great speed is described as Big Data (Manyika et al., 2011). Based on the theoretical framework described earlier and the description of DSS above, it can be said that Big Data will facilitate and further enhance the decision-making processes through the use of data-driven DSS. Moreover, this phenomenon presents further opportunities for the use of mathematical and statistical models that are built in DSS for creating insights that enhance decision-making (Dhar, 2013). Hence the need to research how Big Data and analytics fits decision making and the DSS spectrum.

### **3.3.1 Components of DSS**

DSS has evolved significantly over the past decades and this have been evident in the tools used and the components of DSS. There are some of the tools and components

that have evolved fundamentally and a change of nomenclature has taken place for some in relation to DSS. However, the relevance of some of the DSS tools and components have remained the same over the decades, although some of them have evolved as separate concept. The main components of DSS as identified by recent and earlier scholarship include (Sauter, 2010) data, model, intelligence and visualisation.

### Data Component

The data component comprising the data and information can be seen as the most important factor and the foundation of DSS because it provides the raw materials for supporting decision making. Data can be defined as a representation of facts that can be collected, recorded and used as a basis for decision making (Collis and Hussey, 2009). It is described as the codified form of fact that represents what happened (Piccoli, 2012) and the ubiquitous nature of data today attests to this description of data. Therefore data forms the basis of collecting, organising and describing facts. On the other hand Piccoli (2012) makes a distinction between data and information by describing information as contextualised data.

Organisations are able to collect relevant data and information from within and outside their organisations (Manyika et al., 2011) which can be used for decision making, therefore by extension, the foundation of DSS. As a result, some organisations consider data as one of their most important assets (Brynjolfsson, Hitt and Kim, 2011; Davenport and Harris, 2007). Data used for decision making can be in diverse format. Traditional data can be found in the form of measurements like date, weight, height, cost etc. Database Management Systems (DBMS) have allowed organisations to capture, retrieve and manage operational data in organisations to support decision making.

Subsequently, DBMS evolved into Relational Database Management System that allows for the storage of data in rows and columns with the ability to establish links between different sets of data (Sauter, 2010). RDBMS have features that enable organisations to take more advantage of data within their organisations by making data more easily available for decision making at different levels and as the foundation

for implementing DSS. However, the increased amount, new formats and new sources of relevant data have led to a new paradigm leading to the emergence of new concepts, tools and principles concerning data (Chen, Chiang and Storey, 2012). These new developments affect decision making in organisations and the concept of DSS (Kowalczyk and Buxmann, 2014). These will be explored further in subsequent chapters of this study.

### Model Component

The model components in DSS are included to allow a decision maker to make use of structured routines for decision making that have been developed in fields such as Management Science, Statistics, Mathematics etc. (Bazerman, 2006; Sauter, 2010). Models can be used to provide decision makers with necessary support, especially when dealing with complex problems, by using complex theories without having to understand such theories (Shim, Warkentin, Courtney and Power, 2002). It also enables the use of large amounts of data and information in the decision making process and it eliminates human error or bias from the decision-making process. The model component in DSS allows users to contextualise the relationship or associations that exist between the variables identified in data to be used for decision making (Sauter, 2010).

The use of models in DSS makes it possible to simplify, contextualise and eliminate noise so that decision makers can have a better understanding of available decision options and possible outcomes (Sauter, 2010). It is therefore important that when using a model, only the factors that influence the decision at hand should be used. Moreover, the resulting models should only be considered relevant as long as the factors used in building the model remain relevant and the initial assumptions, when the models were developed, are unchanged.

Factors such as the exponential increase in the processing powers of computers have played a significant role in the study and usage of advanced modelling principles of Management Science, Statistical, Mathematical, Financial etc. Data mining and On-line Analytical Processing (OLAP) are some of the first set of tools that are incorporated into DSS, these will be discussed under the tools of DSS. But other tools



and concepts such as Business Analytics are also a basis for the use of models in DSS and will be explored later in this study.

### *Knowledge Management Component (Intelligence)*

The combination of the data and model components in DSS provide decision makers with support to make timely decisions by using theoretical models to generate models that provide a list of alternative solutions (Courtney, 2001). In most cases, the use of models is based on assumptions; this implies that the input of the decision maker is required to ensure that the right assumptions are made for the decision alternatives generated by the models to produce the right result. However, the twenty first century organisations operate in complex environments where the creation of solution alternatives, based on straight-forward mathematical or financial models or the presentation of operational data may not be sufficient for making right decisions. In this type of situation, organisations find it difficult to even define the problem properly and relevant data and information are not just from the internal database but include those fragments from external sources (Courtney, 2001; Nemati, Steiger, Iyer and Herschel, 2002; McAfee and Brynjolfsson, 2012).

The functionality of DSS that presents data as facts in an easily understandable manner and uses internal data collected to generate solution alternatives has generally become insufficient for the decision makers of the modern day organisations (Nemati et al., 2002; Chen, Chiang and Storey, 2012). Hence, the inclusion and evolution of the knowledge-management component of DSS with the ability to pragmatically add value to data, information and the use of models in the decision-making process (Nemati et al., 2002). Courtney (2001:23) describes knowledge as *“information with guided action and knowing how to act given the information”*. Figure 3.4 presents a diagrammatic framework of the relationship between data, information, knowledge and decision making (Sabherwal and Becerra-Fernandez, 2011).

The framework in Figure 3.4 demonstrates the intrinsic relationship that exists between Data, Information, Knowledge and Decision making. The framework shows a progressive conversion of raw data into information through the processes of analysis that contextualise the data. The information is then converted into knowledge

by interpreting it by guided action which can then be applied in decision making. This explains the functionality of the knowledge management component of DSS as being able to manage large volumes of disparate data and also it's being able to extract valuable knowledge from data (Shim et al., 2002).

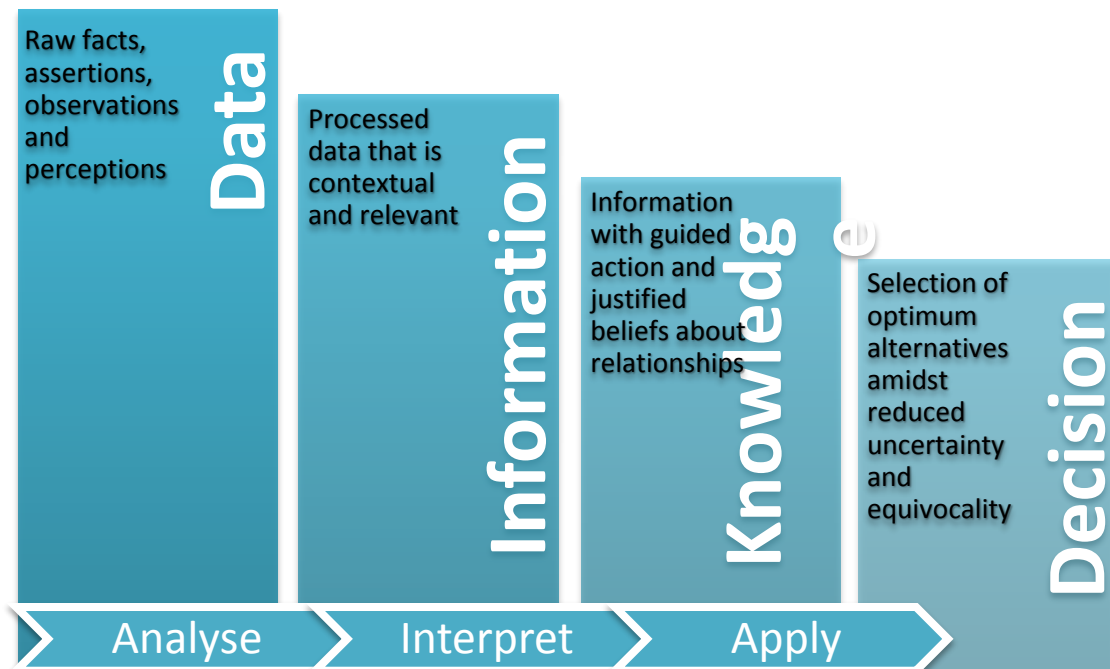


Figure 3.4: Framework showing intrinsic relationship between Data, Information, Knowledge and Decision making. (Adapted from Sabherwal and Becerra-Fernandez (2011))

The contextual description of the knowledge-management component of DSS suggests that the functionality possesses some sort of intelligence. This has been made possible by incorporating concepts from fields such as Artificial Intelligence (AI), Neural Networks (NN) and other mathematical/statistical predictive functions into DSS. These could make computer-based systems to possess inherent ability to take action based on data provided and the process of knowledge creation. The use of intelligent systems in DSS make it possible for decision makers to cover many possible perspectives concerning the problem at hand, identify patterns that are not obvious and determine the best way of presenting the solutions (Sauter, 2010). The data mining tools of DSS discussed later in this chapter facilitates the use of the knowledge-management component of DSS. The concepts of Business Intelligence, Business

Analytics and lately Data Science which are discussed in later chapters of this research present new frontiers for the application of intelligence in decision making (Lim, Chen and Chen, 2013; Dhar, 2013).

#### *User Interface Component (Visualisation)*

The user-interface component of the DSS plays a significant role that connects and enables decision makers to the DSS. The user interface in DSS serves as the means of communication between the computer systems and the decision maker that is described as Human Computer Interaction (HCI) in research and practice (Sauter, 2010). Research in the field of HCI emphasises the importance of designing computer-based technologies for efficiency, effectiveness and most importantly, satisfaction of the users (Dix, 2009). The opportunities to provide user interface to computer-based solutions has evolved significantly with the development of the personal desktop computers, laptop computers, tablets, mobile phones and newer wearable devices.

Guo (2014) describes the research agenda in the field of HCI as understanding how humans interact to system and the development of solutions that improve user experience of system. This encapsulates the importance of the user-interface component of DSS. A well designed user interface component of DSS is expected to enable decision makers to identify and interact with relevant data. It should also simplify the use of models and the knowledge creation/intelligence functionality of the DSS for supporting their decision making.

However, the design and implementation of DSS needs to incorporate technologies that support the realities of the targeted decision maker. With the continuous adoption of mobile technologies for business purposes among decision makers, the evolution of mobile technologies could offer ample opportunities for the future of the user-interface component of DSS.

### **3.3.2 Tools of DSS**

The previous section provides an overview of important components of a Decision Support System. There are several tools that have been used to implement these components in DSS. Earlier literature (Shim et al., 2002) identified four key tools for

DSS - Data Warehousing (DW), On-Line Analytical Processing (OLAP), Data Mining (DM) and interfaces that allow users to interact with computer-based systems, especially web and mobile technologies.

- **Data Warehousing:** This involves the collection, storage and integration of data resulting from transactional processing and operational activities within an organisation (Sauter, 2010). Data Warehousing tools enable the structuring of data using a relational or multi-dimensional approach for the purpose of easy access, reporting and analysis (Sabherwal and Becerra-Fernandez, 2011).
- **OLAP:** On-Line Analytical Processing tools build on the functionality of DW to provide an integrated view or further analysis of transactional data from different sources (Davenport and Harris, 2007; Sauter, 2010). OLAP tools are specially designed for decision making that require access to and analysis of data across multiple dimensions. It provides decision makers with a structured and interactive and multidimensional view of data from different sources with the possibility of drilling down and aggregation of data across several dimensions (Sauter, 2010; Kowalczyk and Buxmann, 2014). OLAP tools have been presented as dashboards that provide an intelligent view of data in recent systems used for decision making in organisations.
- **Data Mining:** Data Mining tools facilitate the modelling component of DSS. DM tools make use of embedded statistical, mathematical, financial models etc. to identify models from data (Sauter, 2010). Lately, some of the DM tools have become more advanced with features that enable the use of Artificial Intelligence and Neural Networks that allow the system to learn by experience from data (Sabherwal and Becerra-Fernandez, 2011; Hardoon and Shmueli, 2013). They enable decision makers to extract valuable knowledge such as describing what happened, understanding what happened, predicting what will happen and providing insight to why something happened (Davenport and Harris, 2007; Sabherwal and Becerra-Fernandez, 2011).

- **User Experience (UX):** These tools allow users to supply input to the DSS, interact with it to perform tasks as required and to receive the outputs that will be used for decision making. The ability to use the web to enable thin-client access and mobile technologies for “on the go” access to the Decision Support Systems have been described as a game changer (Shim et al., 2002). The use of these tools offer decision makers in organisations access to Data, Information and Knowledge required for decision making anytime and anywhere. However, these tools have also been described as the drivers of the overflowing data that is changing the landscape of decision making and the DSS paradigm (Mayer-Schonberger and Cukier, 2013).

Over the years, a number of techniques that are based on the tools identified above have evolved with their origin in DSS. They include Business Intelligence, Business Analytics, Big Data and Data Science (Chen, Chiang and Storey, 2012; Kowalczyk and Buxmann, 2014). These concepts and paradigms support the improvement of decision making by providing contextual information, knowledge and actionable insight (Brynjolfsson, Hitt and Kim, 2011; Chen, Chiang and Storey, 2012; Dhar, 2013). Hence, the tools that were identified as the main DSS tools are now also identified as part of the tools and techniques used for concepts such as Business Intelligence, Business Analytics and so on. These new concepts, tools and techniques will be reviewed further in subsequent chapters of this study. But the remaining of this chapter will explore decision making in grain commodities trading in South Africa and the requirements of a DSS for the same purpose.

### **3.4 Decision Making for Grain Commodities Trading**

The grain commodities industry can be segmented broadly into producers, traders, millers/industrial users and speculators. The grain producers are the farmers that plant and supply the grain commodities while organisations and business that use grain commodities as raw materials to manufacture other products such as animal feed, Bio-fuel products or processed food for human consumption can be described as millers/industrial users. The other segment of the industry includes the speculators that engage in the buying and selling of grain commodities on the market with an

expectation of benefiting from the changes in the prices of grain commodities (Hull, 2012). Within the grain commodities trading industry are also found traders that act as brokers and advisers of financial transactions among stakeholders in the industry.

The different segments in the industry participate differently in the trading of the grain commodities. However, every segment of the industry is always looking for how to maximise profits. The farmers and industrial users are looking for avenues to sell their grain commodities well above their total cost of production. The traders and speculators, however, are focused on the “buy low, sell high” principle to maximise profits (Wright, 2011). However, the segments in the grain commodities industry are all exposed to price-related risks in the trading of grain commodities because of the volatile nature of prices of grain commodities (Geysler and Cutts, 2007; Venter, Strydom and Grové, 2013).

The volatility of the prices of grain commodities and other agricultural products has been a source of concern for academic researchers, governmental and non-governmental organisations for many decades (Wright, 2011; Trostle, 2008). This is because the volatility in the prices of agricultural commodities have dire and multifaceted implications. There are indications that changes in agricultural commodities have social implication on issues like the fight against poverty and economic implication like the GDP and sustainability of the agricultural sector which is very important in many countries (Headey and Fan, 2008; Trostle, 2008). Hence, governments of different countries develop policies that are believed to be in the best interest of agricultural commodities trading.

The prices of grain commodities in South Africa were regulated and controlled by the government under the Marketing Act (Act 59 of 1968, as amended) until 1996 (Mofokeng and Vink, 2013). This meant that the government determined the price that farmers received and at what price agro-processors can buy the commodities. However, from 1996, the grain commodities market was deregulated and became a free market in South Africa (Doyer et al., 2007). The implication of this is that prices were subsequently determined by the markets. As a result, the industry became more competitive and producers became responsible for the trading of their commodities

(Jordaan and Grové, 2010). However, this meant an increase in the volatility of prices of grain commodities with the implication of higher price risks for stakeholders in the industry. This is because the market was now influenced by various economic, social and political factors (Wright, 2011; Venter, Strydom and Grové, 2013).

The volatility of grain commodities prices, the associated price-related risks and the responsibility of trading their own grain commodities suggest that farmers will be confronted with important decisions when trading their products. Previous studies have shown that many South African grain commodities farmers are not participating fully in the market because they do not have the required skills, knowledge and time (Jordaan and Grové, 2010; Venter, Strydom and Grové, 2013).

Based on the requirement and nature of the grain commodities market in South Africa, it can be argued that the farmers do not participate fully because taking full advantage of the possibilities in the market will require that they sift through and understand volumes of economic, political and social data (Wright, 2011; Trostle, 2008) that is scattered in several places. Moreover, they will be required to make a sense out of the changes in these data as it relates to grain commodities trading on a regular basis. This means that the farmers may not be able to focus on their core business of farming. Hence, many of the grain commodities farmers in South Africa focus on the easiest but least optimum strategy for trading their grain commodities (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013).

### **3.4.1 Grain commodities trading strategies**

Grain commodities' trading is facilitated in the South African Futures Exchange (SAFEX), a subsidiary of the Johannesburg Stock Exchange (JSE). This is provided for by the Agricultural Marketing Act (Act No. 47, 1996) (Doyer et al., 2007). The core function of the SAFEX is facilitating the trade in grain commodities and the provision of an enabling environment to provide a platform for risk management and price discovery (Venter, Strydom and Grové, 2013). As a stock exchange, the JSE achieves these objectives by allowing the trading of grain commodities like other financial instruments. Specifically, this allows for the execution of grain commodities trade through the use of derivatives like other commodities such as gold and crude oil.

Derivatives are a means of exchange in the financial market with values that depend on the value of other underlying variables. Examples of such variables include other traded assets like currency exchange rates, stock of traded companies, crude oil, metals such as gold or agricultural commodities such as grains which is the focus of this study (Sundaram and Das, 2011; Hull, 2012). The use of derivatives as the instrument for the trading of any underlying asset or variable serves as a basis for contractual trade. It define the roles and responsibilities of the trading partners and the terms and conditions of the trade. The trading of derivatives is carried out through organisations called an exchange that define the standards of trade (Hull, 2012). Example of such exchanges include the Johannesburg Stock Exchange (JSE) for South Africa, the Chicago Board of Trade established to service farmers and other merchants in USA and Frankfurt Stock Exchange of Germany.

The primary types of derivatives include forward contracts, future contracts and options, although there are several subs for each of them (Sundaram and Das, 2011). These alternatives are applicable in the trading of different commodities just as it is in the case of grain commodities.

### **Forward contracts**

The forward contract is described as the trade between the buyer and the seller, where the buyer agrees to buy an asset for a given price at a set date in the future (Hull, 2012). An example of this, in the trading of grain commodity, is when a farmer agrees to sell his commodity to a willing buyer at an agreed price and date. Hull (2012) emphasised that forward contracts are transactions between a willing buyer and a willing seller outside of an exchange but mentioned the spot contract as a contrast.

### **Spot transactions**

The use of spot transactions is described as the baseline strategy alternative among the alternatives that are available to grain commodities farmers to market their commodities (Venter, Strydom and Grové, 2013). It is regarded as a cash transaction because the seller offers an asset and receives the current market price. For grain



commodities trading, the farmer offers his commodities for sale on the market and receives the applicable market price for the period of transaction for his commodities.

### **Future contracts**

A future contract is functionally similar to the forward contract in that it is an agreement to buy or sell an asset at a future date. However, unlike the forward contract, it is not a contract between individual parties. Rather, the buyer and the seller individually undertake standardised contracts with the exchange (Hull, 2012). In which case, the buyer and the seller are not exposed to the risks associated with a party defaulting because the exchange assumes the responsibility and does what is necessary to ensure that the terms of the contracts it made with the buyer and seller are met (Sundaram and Das, 2011).

An example of a future grain commodity contract could be *“100 tons, of WM1 grade white maize for JULY 2015 basis Randfontein”* of which a farmer agrees to deliver 100 metric tons of WM1 grade of white maize to the Randfontein silos in July 2015 at the futures market of July 2015 as at the time that the contract become effective. This trading strategy is useful for trading grain commodities when there is a concern that the prices might decline by the time of harvest (Venter, Strydom and Grové, 2013). With this strategy, the grain farmer can decide to take future contracts that will expire on different dates to manage his price risks.

### **Options**

The options strategy allows a buyer or seller to take a position in the market but without an obligation to fulfil it. An option instrument can be used as a price risk management for farmers or industrial users of grains (millers) while the speculators can use the strategy to manage their trade in order to take advantage of the price movements in the market (Venter, Strydom and Grové, 2013). The option derivative instrument can be a call option or a put option. The call option instrument allows the holder to buy an asset at a specific date and for a specific price while the put option instrument allows the holder to sell a particular commodity at a specific date and for a specific price (Hull, 2012). However, in both cases, the holders of both instruments reserves the right not

to exercise their instrument without any penalty besides the loss of premiums associated with purchasing the instruments.

The options instruments provide the farmers with some degree of flexibility in the trading of their grain commodities. If there is an indication that the price of a commodity that a farmer has planted will decline, the farmer could purchase a put option that coincides with the time of harvest. This can be exercised if there is a price decline or allowed to expire if it is profitable to enter a new contract or sell the commodity using the spot alternative.

All the grain commodities trading strategies described above are available to South African grain farmers. But the exposure of the farmers to price-related risks is highest with the spot alternative because they are vulnerable and are forced to accept the market price (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013). Furthermore, Jordaan and Grové (2010) concluded that since the deregulation of grain trading in South Africa, there has not been a significant increase in the use of other grain trading alternative beside the spot. This is largely due to the fact that using these alternatives will require specialised skills, knowledge and market intelligence on local and global market, ability to comprehend current trends and decipher future outlooks (Venter, Strydom and Grové, 2013). However, these authors further suggest that the reality is that many South African grain farmers consider all of these requirements to be out of their reach neither do they have the time and knowledge for such practises.

Moreover, the grain commodities market in South Africa is Laissez Faire in nature. In essence, this means that the market and effectively the prices of the grain commodities are controlled by several local and international economic, political and social factors that are rapidly changing. However, the decisions to manage the price related risks and the discovery of the optimum price require the farmers to be au fait with all of these complex trends and data. These factors that influence prices of grain commodities in South Africa are addressed in the next section with a specific focus on the maize commodity. Subsequently, the rest of this chapter will explore the requirements of a decision support system for decision making that farmers can use in order to take advantage of all the grain trading alternatives.

### **3.4.2 Factors influencing grain commodities trading**

It was established in Section 3.4 that one of the most important issues for farmers when making decisions about trading their grain commodities relates to their ability to discover the best price that mitigates price-related risks and maximise profits. The overall impact of this is that a country like South Africa can continue to have sustainable production of grain commodities, increased foreign exchange earnings, food security and provision of much needed jobs. It is therefore imperative that research, innovations and the overhaul of policies that will make it possible for the farmers and the entire industry to have more security continue to receive increased attention. Part of such research is the attempt to understand the factors influencing grain prices that has been ongoing for decades and provision of improved decision-making support for farmers and other stakeholders.

There is a general consensus among economists, academics, government policy makers, producers and other stakeholders in the grain commodities industry that there are several variables that affect grain prices (Abbott, Hurt and Tyner, 2011; Wright, 2011; Venter, Strydom and Grové, 2013; Khamis, Nabilah and Binti, 2014; Trostle, 2008). Generally, the factors that have emerged in the discourse on the factors that influence grain prices can be grouped under the following categories:

- Demand, supply and storage;
- Macroeconomics; and
- Political factors.

Several variables fall under each of these themes with varying degree of influence. It has also been noted that different schools of thought exist as to which of the themes should be the focus of understanding the volatility of grain prices (Irwin, Sanders and Merrin, 2009; Wright, 2011, 2014). But the volatility of grain prices, especially during market shocks that create outliers in the price data, might be as a result of a different combination of factors at different times (Abbott, Hurt and Tyner, 2011; Wright, 2014). Once again, this highlights the complexity of the market and the need for innovative solutions that the farmers can use to understand the markets. The variable under the identified themes are explored below.

## **Demand, supply and storage**

Economic theories suggest that prices will go up when there is an increase in demand for any commodity especially when the supply of such commodity does not increase with demand (Burda and Wyplosz, 2009). In reverse, the price of commodities is forced downward when there is over-production, reduced demand, or a huge stockpile of commodities. This summarises the impact that the local and international utilisation of grain commodities for domestic and industrial use have on the grain commodities price.

Variables under the theme include factors that influence the ability of farmer to supply or those factors that cause over-supply and the calming or panic effect that the level of grain stockpile has on the volatility of grain prices (Wright, 2011; Abbott, Hurt and Tyner, 2011; DAFF, 2014; Trostle, 2008). Others include the demand for grain commodities as an important source of calories for human consumption and industrial demand for animal feeds and biofuel. The prominent variables under the demand, supply and storage theme that influence grain prices are described below.

- Domestic utilisation;
- Industrial utilisation;
- Utilisation of major importing countries;
- Production level in major exporting countries;
- Influence of weather on production;
- Input costs;
- Local stockpile;
- International stockpile;
- Price, demand, supply and storage of substitutes; and
- Level of utilisation compared to stockpile (stock-to-use-ratio).

## **Macroeconomics**

Macroeconomic factors have also been identified as influencing the changes that occur in the price of grain commodities. Like the previous theme of demand and supply, there are several variables which influence grain prices that fall under this category. However, studies show that some of the macroeconomic variables influence

the prices because they are linked directly to the factors of production (Trostle, 2008). The influence of the other macroeconomic factors, however, are simply a reflection of the state of the local or global economy (Abbott, Hurt and Tyner, 2011). Although there are suggestions that the use of macroeconomic variables for understanding grain commodities prices requires further research (Wright, 2011), it remains an important part of the discourse on the price of grain commodities (Abbott, Hurt and Tyner, 2011; DAFF, 2014; Wright, 2014; Trostle, 2008). The macroeconomics variables that influence the price of grain commodities identified from the literatures that has been cited above include:

- Currency exchange rates (especially US Dollars to other currencies);
- Price of crude oil;
- Local interest rates; and
- Consumer price index.

### **Political factors**

The influence of government policies, political interactions and international trade, which is largely driven by politics, cannot be separated from the swing in the prices of grain commodities (Abbott, Hurt and Tyner, 2011). Although, the grain commodities markets are deregulated in many countries, political influence on economic, social and trade related issues is a reality. Moreover, to a large extent, the economic growth or decline of a country can be attributed to the actions or inactions of politicians. In many instances, the trade data from derivative instruments is used as an indication of such political activities. An example of this is reflected in the bid of politicians to find a balance in the trade between two countries by imposing import/export trade barriers or trade sanctions as a result of political fallout between countries.

There is general understanding that issues relating to politics affect the prices of grain commodities (Abbott, Hurt and Tyner, 2011; Wright, 2011, 2014; Trostle, 2008). However, unlike variables under the other themes which have quantitative indicators for which data is collected and analysed, the impact of politics on the prices of grain commodities can be said to have to be largely subjective and opinion driven. The next

chapter of this study will explore the approach and new sources of data for quantitative study of the impact of politics on grain commodities prices.

The themes explored in this segment provide insight into the factors that influence the grain commodities prices. This includes variables for which data in monthly, daily, hourly and in some cases minute by minute data are generated and stored. Based on the identified theme and variable, the next segment of this chapter provides the result of survey conducted to under the perception of some stakeholders on the factors that influence grain commodities prices in South Africa.

### **3.5 Grain Commodities Trading DSS Requirement Survey**

The Design Science Research (DSR) process requires that requirement for the design or development of envisaged artefact be collected after explicating the problem as discussed in Chapter 2. Section 3.4.2 of this Chapter initiated the gathering of requirements for a framework supporting decisions on grain trading by reviewing the literature on the factors that influence the prices of grain commodities prices. In order to contextualise the reviewed literature, some South African grain farmers and traders were approached to determine their perception of the factors that affect the price of grain commodities. The survey also measured the perceived gap in the industry and what is expected from a grain commodities trading DSS. The call to participate in the survey was made open to farmers and traders in the South Africa through [www.landbou.com](http://www.landbou.com) (Appendix B). The website was chosen because of its popularity in bringing together stakeholders of the grain commodities industry in South Africa.

An exploratory study was conducted with 10 farmers and 6 commodity traders who indicated their interest in the survey. A semi-structured approach was adopted for the study resulting in the collection of both quantitative and qualitative data. The questionnaires used were designed to collect demographic information, Likert scale and open-ended questions in order to encourage open contribution from the farmers and traders (Appendices C and D). The data collected was analysed using descriptive statistical methods for the demographic and Likert scale questions. Finally, the content

analysis method was used to identify themes from the answers to the open-ended questions (Collis and Hussey, 2009).

### **Farmer's survey**

All the farmers that participated were male but the age followed a fairly normal distribution with five of them in the 41 – 50 years category, while one farmer was in the 21 – 30 years category, two in the 31 – 40 years category and two in the 51 – 60 years category. The majority of the farmers that participated can be described as experienced with five of them having between 11 – 20 years of farming experience, four with more than 21 years of farming experience and only one with less than 10 years of experience. Among the farmers that participated, the minimum educational qualification was a matric or equivalent which three of them possess, four possess a diploma, one an undergraduate degree and two of them hold master degrees. Moreover, the volume of grains produced by the farmers represented different segments, four of them indicated that they produce between 1,001 – 2,500 metric tons per year while three indicated that their annual production is between 2,501 – 5,000 metric tons. Each of the other three fall into the 1 – 500, 501 – 1,000 and 5,001 – 10,000 metric tons production of grain commodities annually.

The general consensus on the demand, supply and storage theme is that the variables discussed under the theme in Section 3.4.2 have an influence on the volatility of the grain commodities price in South Africa. At least eight out of the 10 were in agreement on the influence of most of the factors identified in the literature. Perhaps most remarkable is that eight of the 10 farmers identified that the price of grains in the United States of America, as a major producer of some grains, affects the price of grain commodities in South Africa.

However, the farmers that participated in the survey had different perceptions on the influence of macroeconomic factors on commodities prices in South Africa. When asked if they agree that the price of crude oil influences grain prices in South Africa, six out of the 10 farmers agreed, but only three out of the 10 farmers agreed that the lending rate or Gross Domestic Product (GDP) figures influence grain prices in South Africa. Moreover, only two out of the 10 farmers agreed the overall performance of the

Johannesburg Stock Exchange (JSE) reflects on the prices of grain commodities in South Africa. However, all the farmers agreed that the US Dollar-Rand exchange rates and the Consumer Price Index affects the price of grain commodities. When asked about the influence of politics, seven out of the 10 farmers agreed that government policies have an influence on the prices of grain commodities in South Africa.

The farmers were also asked open-ended questions that brought about open discussions. The following theme was favoured by most of the farmers that participated in the survey about the perceived gap in the industry and their expectation from a grain commodities trading DSS:

- The need for information on international variables that affect grain prices in South Africa;
- Easy access to reliable information; and
- The need for dependable future outlook.

### **Trader's survey**

Commodities traders are experts that act as intermediaries between the farmers and the exchange or other buyers (Hull, 2012). In many cases, the traders are contracted to make or execute trading decisions on behalf of the farmer after advising them. In this category, eight traders participated in the survey to measure their perception of the factors that influence the prices of grain commodities in South Africa. The trader's survey also focused on trying to determine future outlook of the prices of grain commodities.

Out of the traders that participated, seven were male and one female, four of them have master degrees, three of them have undergraduate degrees and only one possess a diploma. Out of the eight traders, four of them have between 11 – 20 years of trading experience, three of them have between 6 – 10 years of experience and the last one falls between 2 – 5 years. It was also noted that three of the traders manage trade that is between 100,001 – 250,000 metric tons of grain commodities annually and two of the traders manage less than 100,000 metric tons annually. But, two of the



traders manage trades that aggregate between 500,001 – 1 million metric tons annually and the other manager trades between 1 – 5 million metric tons annually.

Generally, the traders agree that the factors that were discussed under Section 3.4.2 of this chapter affect the prices of grain commodities in South Africa. All of the traders indicate that they watch at least some of these variables when they make decisions about grain commodities trading for their clients. The traders also indicated that the traders advise their clients and make grain trading decisions by analysing data collected on these factors either visually or by using different calculations that have been developed based on their years of experience. They also indicate that they use the data collected on these factors to advise grain farmers on the right trading strategy to adopt.

It can be deduced from the literature review in Section 3.4.2 and the result of the survey described in this section that data on the identified themes that affect the price of grain commodities in South Africa is an important component of a decision support solution for the grain farmers. This should include data on local and international factors. The DSS should also combine and analyse the data to present insight about the factors influencing the grain prices. It was noted that DSS for grain commodities decision making should allow the farmers to compare the benefit and possible risks of the different grain trading alternatives. This implies that a grain trading DSS should also provide future outlook of grain prices for the different trading strategies. However, the volatility of the grain commodities prices could mean that the insight and future outlook might be relevant for only a short period. Therefore, a DSS for grain commodities trading that collects real-time data and provides insight and future predictions based on real-time analysis will be beneficial to the farmers.

### **3.6 Conclusion**

This chapter reviewed decision making for businesses and the theoretical background of decision making in organisations. It also reviewed the role of computer-based interventions and Information Systems as Decision Support Systems in improving decision making. The review of literature reveals that a functional DSS is made up of

data, model, intelligence and visualisation components. It was identified that the use of DSS can add immense value to organisational decision making and this applied to the decision-making process for grain commodities trading.

Chapter 3 emphasised the importance of quality decision making in any organisation and it was identified that decision making could have a direct impact on the efficiency profitability and productivity of an organisation. Therefore, there is a need to reduce ambiguity to the minimum and to provide conditions or systems that enhance clarity during decision making. It was found that over the years, Computer Science and Information Systems have played a significant role in decision making by enabling the gathering, processing, presentation and the usage of data and information for supporting decision making. This has necessitated the evolution of different Computer-based tools, concepts and principles the majority of which have their roots in the use of data in decision making.

In the context of this study, decision making for grain commodities trading was addressed. Decisions relating to price discovery and risks were identified as the major decision that faces grain farmers when trading their commodities. Hence, the factors that influence the price of grain commodities and the requirements for a grain commodities DSS were also explored in this chapter.

To address this need, Chapter 3 focused on research objective (RO<sub>1</sub>) *to identify data-related requirements for a system to support decisions on trading grain commodities in South Africa*. Based on the set objective, two research questions were raised and answered in Chapter 3. Attempts were made to answer research question (RQ<sub>1</sub>) - What are the local and international factors that influence the grain commodities market in South Africa? Also, this chapter sought an answer for research question RQ<sub>2</sub> – What strategies in trading grain commodities are available for minimising price-related risks and increasing profitability?

It was identified that there are several factors that influence the price of grain commodities in South Africa. Some of the identified factors are local, while the others are external factors that are from outside South Africa. Furthermore, it was identified

that these factors are dynamic and the degree of influence of each of them on the prices of grain commodities in South Africa varies with time. Besides, it was deduced that there is a need to keep abreast of all of these variables and their changing degree of impact on the prices of grain commodities in order to understand the current trend in the market and the future outlooks.

This chapter stressed the need to have real-time intelligence and predictability based on the data created from these factors as a foundation for a grain trading DSS. It was also deduced that a grain commodities trading DSS for South African farmers should provide real-time intelligence on the factors that influence the price of grain commodities in South Africa. It was proposed that the DSS should present the market intelligence and predictions in a simple way because of the behaviour pattern of the South African grain farmer. This can be achieved by having a grain commodities trading DSS that collects, integrates and analyses data from several sources in real-time to provide market intelligence and predictive analysis that can give an indication of the future.

Furthermore, it was found in this chapter that there are different strategies that can be employed in the trading of grain commodities to minimise price-related risks. These options include the use of spot, futures contracts and options trading strategies. It can be deduced from this chapter that a DSS that provides the future performance of grain commodity for the different trading alternatives could enable the farmers to make better and more informed grain commodities trading decisions.

Within the DSR paradigm, the problems and the opportunities identified in this chapter define the relevance of a framework and possible implementation of the framework to support decisions about trading grain commodities in South Africa. On the other hand, the factors that influence the grain commodities market and the structure of the market that were reviewed in this chapter lays a foundation for of the rigour cycle of the DSR process adopted in this study. The next chapter of this study will review the sources of data for the factors that affect grain commodities trading in South Africa. The next chapter will also address how these data can be collected, cleaned and integrated into a single source.

# Chapter 4 : Big Data for Grain Commodities Trading in South Africa

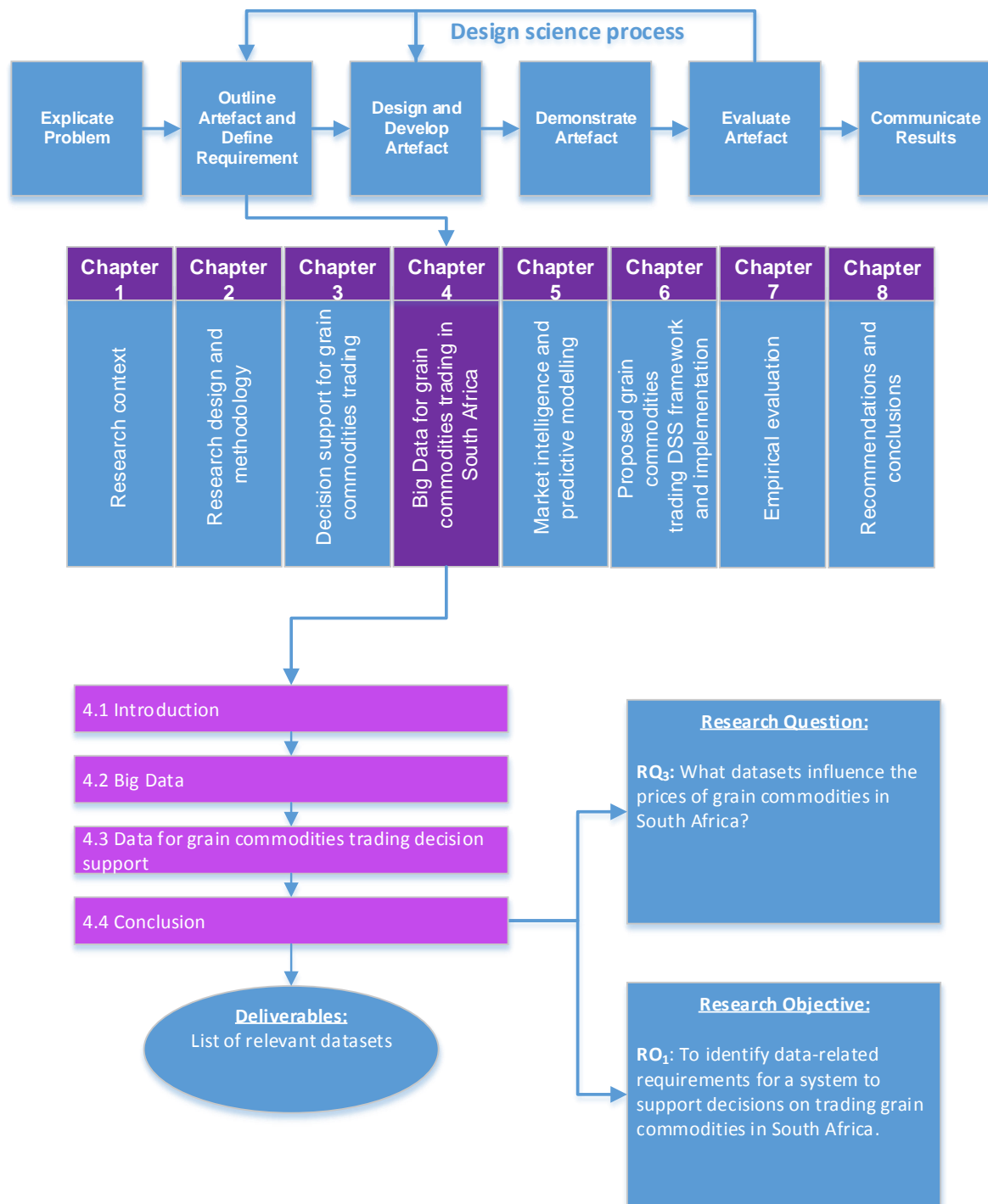


Figure 4.1: Chapter outline and deliverables

## 4.1 Introduction

Chapter 3 addressed the impact of decision making on businesses and organisations. It was identified that the quality of decisions made by decision makers in a business or organisation can generate increased profitability and growth, yet it could also have dire consequences. Furthermore, it was identified that uncertainty, equivocality and ambiguity are the main challenges that confront decision makers. Chapter 3 further discussed the use of Computing techniques and technologies as important components of a Decision Support System (DSS). Prominent among these techniques and technologies are data-driven solutions which have become important for decision makers with the evolution of Big Data and Data Science.

Within the context of making decisions on trading in grain commodities, the previous chapter identified the discovery of optimal price as a key driver for the trading decisions made by grain farmers. Moreover, Chapter 3 explicated the factors that influence the price of grain commodities and proposed the use of data regarding these factors and the grain commodities trade-market statistics as the foundation of a grain trading DSS. This chapter will examine the role Big Data could play in a grain commodities trading DSS.

Big Data has been described as a concept with the potential to influence all aspects of life including work and play (Manyika et al., 2011; McAfee and Brynjolfsson, 2012). This is because of the opportunities to extract actionable insights from large datasets that the concept presents. However, the opportunities presented by Big Data are not without challenges. This chapter will address Big Data as a concept and the role it plays in grain commodities trading decision-making. The aim is to further contribute to the first research objective of this study (RO<sub>1</sub>), which is “*to identify data-related requirements for a system to support decisions on trading grain commodities in South Africa*”.

The aim of this chapter will be achieved by seeking an answer to the third research question of this study **RQ<sub>3</sub>** - *What datasets influence the prices of grain commodities in South Africa?* Chapter 4 will provide the initial components of the grain commodities

decision-making framework. This will form the bedrock of the design/development activity within the design science research process that has been adopted for this study. In doing so, the focus of Chapter 4 will be the consideration of methods, tools, techniques culminating in the definition of requirements and environmental considerations in the development of the eventual artefact of this study. This will be achieved by drawing from the knowledge base through literature review and a study of previous research work that is relevant to this study. Hence, Chapter 4 forms part of the rigor cycle in the DSR process. However, the environmental consideration in this chapter also indicates that there will be an iterative contribution to the relevance cycle.

The grain commodities that are grown by commercial farmers in South Africa are white maize, yellow maize, soybean, wheat, sunflower, sorghum, malting barley and canola among others (DAFF, 2014). Fundamentally, the trading procedure, risk and opportunities of growing each of the commodities are the same. However, agricultural economics literature separates the marketing research of the different grain commodities (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013).

There are suggestions that the economics behind the factors that influence the price of each commodity are different (Irwin, Sanders and Merrin, 2009; Wright, 2011). Therefore, it is expected that the combination and degree of influence of the factors that affect the price of the grain commodities will be different for each of the grain commodities. Because of the constraint of time and resources, the rest of this study will focus on the trading of white maize in South Africa because of its economic importance (DAFF, 2014).

Section 4.2 of this chapter will explore the concept and characteristics of Big Data. The section will also explore the opportunities, challenges and the value of Big Data for decision making in businesses. Thereafter, Section 4.3 will approach the acquisition of data for grain commodities trading decision-making DSS from a Big Data perspective. It will also identify the possible sources for all the datasets required for the grain commodities DSS. Section 4.3 will also discuss how to integrate the relevant datasets into a single data source that can provide insight and intelligence for grain

commodities trading decision making. Finally, Section 4.4 will provide a conclusion of Chapter 4 with an outline of objectives and deliverables that have been achieved.

## **4.2 Big Data**

The impact of data in practice and research are far-reaching. In most cases, data forms the basis of decision making within organisations and is the basis of scientific inference where research is concerned. Some organisations now consider data as one of their most important assets (Brynjolfsson, Hitt and Kim, 2011). This might be irrespective of the quality or the quantity of data available to such organisations. However, in recent years, data has become globally available and the amount of data generated has increased significantly (Manyika et al., 2011). As a result, the expectation of practitioners and researchers about data is intensifying and the concepts of Big Data, Data Science and Business Analytics are emerging in Computing Science fields and several other stakeholders. The new concepts have been described as having challenges and opportunities that will affect productivity, profitability and efficiency (Manyika et al., 2011; Mayer-Schonberger and Cukier, 2013).

The volume of data available globally has grown significantly and the rate of growth is increasing by the minute. This deluge of data generated is now described as Big Data and has changed the way people live and how organisations operate. It has become a defining factor on how business is conducted and the impact is also being experienced in the different fields of research and practice (Manyika et al., 2011; Chae and Olson, 2013).

Traditionally, organisations have collected data resulting from business transactions and internal operations such as sales, marketing, finance, production and human resources management. The collection of organisational related data is as a result of the need to automate processes and systems by using information technology and the accompanying tools to simplify internal systems and provide better service to customers. Organisations soon found the need to integrate the various sources of data into a single repository for the purpose of extracting knowledge and information (Provost and Fawcett, 2013b). This paradigm formed the basis for data warehousing,

which involved the aggregation of data from different systems into a single source of intelligence. Data warehousing evolved with techniques, technologies and approaches that enabled organisations to find answers to questions mostly on “what happened?” and “what is happening?”.

The growth in the amount of data within organisations and external data – that is mostly unstructured, has amplified the value of learning from data and the use of data for supporting decision-making processes. Organisations are now able to predict what is likely to happen in the future and have evidence based scenario planning of future occurrences. These possibilities introduce a new paradigm into the use of data for decision making; hence, the need for an understanding of the concepts introduced by Big Data.

The definition of Big Data has been based on its complexities, sources, storage and management. Big Data was described in earlier studies according to the volume of data created, the velocity of data created and the variety of data that make up Big Data (Manyika et al., 2011; McAfee and Brynjolfsson, 2012). However, recent studies have now included veracity as part of what characterises Big Data (Mayer-Schonberger and Cukier, 2013).

The four main characteristics that have been used to describe Big Data bring the complexities, opportunities and risks associated with Big Data into perspective. According to Manyika et al. (2011), Big Data is described by the fact that it is very large in volume in most cases. Also, the rate of data creation is fast and there are varieties of sources which are beyond the ability and capacity of traditional tools, processes and management practices.

### **Volume**

The ability to generate and collect data has increased across various spectra of our daily lives. A terabyte of storage space for data seemed extravagant for data storage previously, but it is estimated that Walmart generates about 2.5 petabytes of data every hour (McAfee and Brynjolfsson, 2012). This is an equivalent of 2,500 terabytes



of data and that is from only one of many organisations generating such volumes. This is an indication of the volume of data that has now become available.

### **Velocity**

The rate at which data is being created is a characteristic that redefines data. It was estimated that over 2 million searches were requested on Google every minute in 2012, this increased to over 4 million in 2014 (Gunelius, 2014). The report by (Gunelius, 2014) added that about 300,000 messages were sent on Twitter every minute, while users on Facebook shared about 2.5 million posts every minute. These statistics provide an indication of the rate at which data is being generated on these unconventional sources of important data. The rate at which data is generated and captured is also reflected in business applications such as the ability to capture intraday trading data on stock exchanges around the world. Intraday trading data generates data for every change in asset prices and trades that are placed on the market which can be volumes of data in less than a second (Hull, 2012). This suggests that data should no longer be seen and managed from the “warehouse” point of view where data is collected into silos and only used much later. Rather, it is important to recognise the necessity of collecting and using Big Data in real-time or near-real-time for effective management and optimising its value (Davenport, Barth and Bean, 2012).

### **Variety**

The sources and types of relevant data that can be described as Big Data are heterogeneous in nature. It include videos, audios, images, GPS coordinates for mobile device applications, documents, web pages, data created by several business applications and many more. Traditional datasets are structured into rows and columns and their creation is planned in most cases. However, most of the important types of Big Data are those that do not fall into the traditional dataset category. Broadly, Big Data has been categorised into structured and unstructured types. The structured data types are those generated from enterprise systems. Generally, structured data fits the storage and management principles of the relational database management system (Chen, Chiang and Storey, 2012).

The ability to create data from more aspects of life and business activities (Mayer-Schonberger and Cukier, 2013) has resulted in the creation of other data types that cannot be stored or managed by using the conventional databases (McAfee and Brynjolfsson, 2012). These types of data are categorised as unstructured data. Besides the types of data that characterise Big Data, it is also important to take note of the fact that Big Data has broadened the scope of data sources (Manyika et al., 2011). Therefore, to take advantage of Big Data opportunities, it will be important to explore different sources of data (Brynjolfsson, Hitt and Kim, 2011) and also to establish relationships among and with the variety of fragmented data (Mayer-Schonberger and Cukier, 2013).

### **Veracity**

The uncleanness and inaccuracy of Big Data is a result of the other characteristics. An IBM survey indicates that 27% of respondents were not sure of the accuracy of the data used for decision making and more than 30% use data that they do not completely trust (IBM, 2011). When considering Big Data as a source of insight, it is important to consider the integrity of the data. This is because the data might require cleaning, removing of noise and plans to accommodate incompleteness and inconsistency in the data.

All of the characteristics discussed above are what properly defines Big Data. Goes (2014) suggested that it might be erroneous to consider large datasets as Big Data just because of their volume. Large volumes of datasets have existed in fields such as astronomy, studies relating to weather patterns and genomics for much longer. The use of datasets in these fields of study and practice cannot necessarily be regarded as the application of Big Data because the concept of Big Data is beyond the volume of data collected (Goes, 2014). It is, however, the combination of the volume and the velocity and/or variety together that introduce the veracity that necessitates new ways of storing and processing Big Data (Chen and Zhang, 2014).

The fluidity of Big Data makes it possible to generate dynamic and real-time insight, especially because of the volume and velocity of the data (Davenport, Barth and Bean, 2012; Goes, 2014). Therefore, giving attention to Big Data may not be optional as a

tool for predicting the future, the foundation for the emerging digital economy and innovation (Davenport, Barth and Bean, 2012).

#### **4.2.1 Big Data techniques and technologies**

The characteristics of Big Data bring about a new paradigm that suggests a completely new approach to what is considered as data, how data is collected, stored, processed, analysed and used. By its definition and identified characteristics, traditional techniques and technologies for handling data no longer suffice for Big Data (Chen et al., 2013). Hence the need for new techniques and technologies for managing and capturing value from Big Data. Besides, working with Big Data requires that practitioners and researchers think and see things differently (Chen and Zhang, 2014). As a result, new and innovative techniques and technologies that support the demands and characteristics of Big Data are evolving (Minelli, Chambers and Dhiraj, 2013).

Big Data techniques and technologies are mostly multidisciplinary, cutting across Computer Science, Information Systems, Economics, Mathematics, Statistics and other disciplines (Chen and Zhang, 2014). This suggests that dealing with Big Data requires a more scientific approach, even among practitioners that operate outside academia. It can be said that this is important because of the requirement to identify relevant data sources especially from unconventional sources and the need to get the best out of unstructured data.

#### **Big Data Techniques**

Unlike the traditional use of data in business or research, using Big Data and associated concepts such as Data Science to derive optimum value requires a different approach. It needs innovative technologies, creativity, common sense, domain knowledge and systematic thinking. A different way of defining problems or identifying opportunities is fundamental to the paradigm change that exists in the concept of Big Data. This is that the entire solution or project is sub-divided into data driven components (Provost and Fawcett, 2013b). Thus, in making choices concerning the application of Big Data, care should be taken in the choice of techniques that allow for the running of complex components in parallel. Adequate consideration should also

be given to requirements about decision time, which could be in real-time, close to real-time, hourly, weekly, monthly or yearly (Chen and Zhang, 2014; Goes, 2014).

Chen and Zhang (2014) categorised the techniques required for extracting insights from Big Data into the use of mathematical techniques, data analysis techniques and Big Data applications. Each category could require the use of techniques from different fields interchangeably as depicted in Figure 4.2 below. The graphical presentation of the Big Data techniques shown in Figure 4.2 indicates the combination of mathematical and data analysis techniques for leveraging Big Data for different fields of interest.

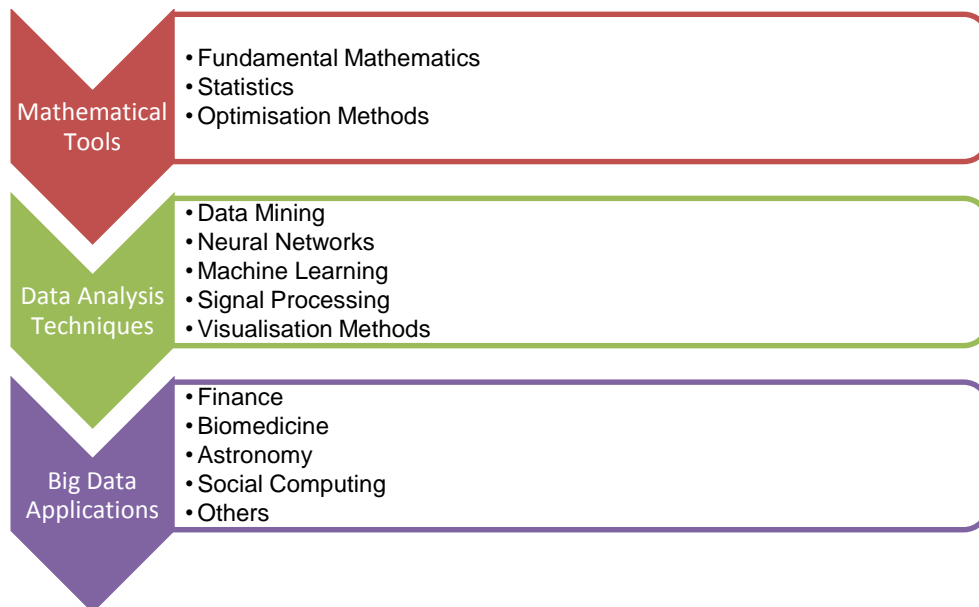


Figure 4.2: Big Data techniques (Chen and Zhang, 2014)

Mathematical tools used for the collection, integration, analysis and extraction of insight from Big Data include fundamental mathematical methods, statistical techniques and optimisation methods (Chen and Zhang, 2014). There are other techniques with their foundation in mathematics that make it possible to deal with the complexities of Big Data. These include econometrics and statistical computing (Goes, 2014).

The data analysis techniques are used mainly for managing and identifying patterns from Big Data. Although some of the techniques have their foundation in the traditional paradigms, many of these techniques are evolving to accommodate the complexities of Big Data. The techniques in this category are data mining techniques, artificial intelligence techniques, visualisation techniques and techniques that are based on analysis of networks.

There is a subtle connection between mathematical and data analysis tools, for example, both statistical and machine-learning techniques can be used for forecasting and clustering. However, the foundation of machine learning is in Computer Science while statistical techniques are from Statistics (O'Neil and Schutt, 2014). Although statistical techniques are more appropriately positioned under mathematical techniques, it has been noted that both can be used together in Big Data projects (O'Neil and Schutt, 2014; Chen and Zhang, 2014). The two types of techniques represent the nature of the relationship that could exist between the use of mathematical techniques and data analysis techniques for Big Data which is determined by the need.

### **Big Data Technologies**

The characteristics of Big Data and the relevant techniques for handling and deriving value from Big Data necessitate a shift from traditional database management infrastructure and technologies. This has brought about a plethora of systems that are designed to provide solutions to complexities associated with Big Data. Some of the fundamental issues identified in the literature as driving the Big Data technologies include the need for scalability, reliability, timeliness and a market shift towards service orientation (Minelli, Chambers and Dhiraj, 2013; Chen and Zhang, 2014). The factors all come together as planning for the unusual flow of Big Data, how processing is done, how users can interact with the data and solutions derived from it.

In designing Big Data applications and solutions, the underlying technologies enabling Big Data include distributed computing, parallel computing, in-memory computing and cloud computing among others (Chen and Zhang, 2014; Goes, 2014). The use of distributed and parallel computing refers to the use of multiple processors for handling

the burden of large datasets and complex analysis associated with Big Data (Minelli, Chambers and Dhiraj, 2013). Distributed computing divides tasks into sub-units which are handled by using different processing resources within the system either at the same time or one task after the other (Chen and Zhang, 2014).

Despite the ability to share the load created by Big Data in order to improve efficiency and reliability, organisations may not be able to adequately predict future requirements of resources. As a result, it has been noted that the processing, storage and other resources used in the Big Data technology stack are designed and planned to scale out rather than scale up as it is with traditional infrastructure planning. Scaling out is the ability to add more units of resources into a system as the need arises (Fernandez, Migliavacca, Kalyvianaki and Pietzuch, 2013). This causes minimum or no disruption to processes whereas the traditional scaling up of computing resources could imply huge investments in new equipment which creates redundant resources.

The large volumes of data and the complexities associated with processing Big Data makes creating copies of data not feasible when running complex algorithms because of resource requirement. Traditional systems are more likely to create multiple copies of the same data across platforms or in between tasks, but new Big Data technologies take the tasks/analysis to the data with in-memory computing. This approach makes use of the data stored in the Random Access Memory (RAM) instead of the disk. This makes it possible for all the necessary tasks such as the running of algorithms to be brought to the data residing in the memory thereby ensuring the efficiency of such algorithms and making it feasible to use them with Big Data (Chen and Zhang, 2014).

The third underlying approach enabling Big Data is cloud computing. This refers to the virtualisation of computing resources, services and infrastructure which are made available over public or private networks. Cloud computing makes the use of supercomputing resources and infrastructure easily accessible and affordable (Chen and Zhang, 2014). It takes away the challenges of administering and maintaining such resources and infrastructure by using shared infrastructures. Cloud computing enables practitioners and researchers to store and carry out intensive Big Data tasks without having to own expensive and difficult-to-manage infrastructure. This is

achieved through a service model that enables users to define and pay only for what is required per time with an option to easily scale out.

Although there are potential disadvantages, such as security-related threats, privacy concerns and dependence on the internet with cloud computing, using it as a basis for sharing large dataset makes it a very important tool for Big Data. This becomes even more important where independent systems need access to the same data or service that is updated in real-time. Moreover, with the need for more self-service orientated solutions that need to deliver real-time solutions for mobile user, cloud computing provides a platform that enables Big Data Solutions to be provided as services (Minelli, Chambers and Dhiraj, 2013) and thin-client applications.

Several systems have emerged in the markets that are designed to support the characteristic nature of Big Data and enable users and organisations to get the benefits. Hardly any of these systems incorporate all of the Big Data technology requirements, instead, the systems have been designed and developed with different areas of focus. Apache Hadoop and SAP HANA are two of such technology frameworks that have emerged recently that are designed specially as technology support for Big Data (Chen and Zhang, 2014; Apache, 2015; SAP, 2015).

The Apache Hadoop Framework has been designed specifically to handle the complexities of Big Data through an innovative, distributed computing approach. It is designed to handle large datasets using a distributed storage and processing approach. It is specially designed to scale out from one to several thousands of computer clusters. It splits data among the computers in the cluster with each of the computers being able to offer localised computing and innovative intelligence that coordinates the operations (Apache, 2015). Apache Hadoop Framework comprises several software packages that are developed to handle different needs which include the Hadoop Distributed File System (HDFS), Chukwa, MapReduce, Pig, Hive and Manhout. Apache Hadoop is an open source suite of applications with several commercial options especially as a service through cloud computing.

On the other hand, SAP HANA is a fully commercial Big Data technology system with more emphasis on in-memory computing, predictive analytics and data partitioning for distributed computing (SAP, 2015). SAP HANA has been designed for acquiring large streams of data. The core focus of SAP HANA is its ability to collect data, perform complex routines and provide insights all in real-time with no latency (SAP, 2015; Chen and Zhang, 2014). SAP is also available as a service through several-cloud computing service providers making it an important Big Data technology in collecting and using data from several sources as the data is being created. There are several other Big Data technologies evolving in the market such as MangoDB and various commercial implementation of the Apache tool. The choice of technologies should be based on needs of the proposed solution architecture.

### **Big Data Approach**

The Big Data tools and technologies described above provide the foundation for approaching a problem from a Big Data perspective. From a problem solving perspective, the Big Data approach has also been summarised into the provision of technologies that can scale infinitely for the acquisition and storage of data. Secondly, Big Data approach requires that large datasets that have been collected should be processed and converted to business intelligence and insights where it sits (Minelli, Chambers and Dhiraj, 2013).

It has also been suggested that the approach for implementing Big Data concept should follow a life cycle of data acquisition, information extraction, cleaning, data aggregation, integration, modelling, analysis and interpretation (Jagadish et al., 2014). From an operational perspective, this aligns with the Data Science process that will be described in Section 5.3. However, the Big Data approach is beyond the tools, techniques, technologies and processes that should be followed. Adequate care should also be taken to ensure that the peculiar benefits of Big Data are exploited and pertinent risks are mitigated (Chen and Zhang, 2014). Such opportunities and challenges and described in the next sub-section.



#### **4.2.2 Big Data challenges and opportunities**

The evolution of Big Data and other supporting concepts open up new ways of managing organisations and have become the foundation of a new type of organisation. This is because of the discovery, insight and application of the new science that is made possible by having access to Big Data (Chen, Chiang and Storey, 2012). The vast amount of relevant data available externally to organisations complements the data generated internally and enables an organisation to know more about its industry, customers, and competitors. Hence organisations are able to do more, plan better, respond to changing customer needs faster and evolve quickly.

Big Data offers organisations an improved decision-making process, improved quality of decisions and the ability to make decisions in record time (McAfee and Brynjolfsson, 2012; Kowalczyk and Buxmann, 2014). Moreover, the ability to use Big Data to predict future outlooks through the use of advanced analytics has also been described as a major opportunity that Big Data offers (Dhar, 2013).

Researchers and practitioners have sought the use of data to determine possible future outlooks for many years before the evolution of Big Data through the use of mathematical, statistical, financial and economic modelling. However, Davenport (2014) suggests that adding datasets from more data sources can add more value to models than just refining such models. As a result, studies have shown that Big Data offers an opportunity to create new products and services that are specially developed, based on the insights and discoveries from Big Data (Patil, 2012; Davenport, 2014).

The emergence and adoption of Big Data also comes with associated challenges that practitioners and researchers need to take into consideration in order to take advantage of Big Data opportunities. The characteristics of Big Data bring about complexities that create different types of challenges. These challenges could be technical in nature; ethics and privacy; talent and leadership concerns; as well as issues of security.

## **Technical**

The volume, rate of flow, variety, inconsistency and incompleteness of the datasets that are classified as Big Data introduces a number of technical challenges. Although, Big Data technologies that deal with these challenges are evolving as described in Section 4.2.1, it is important to understand the challenges in order to have effective plans that maximise opportunities and create value. Different authors have emphasised the importance of unstructured data such as social media data, data from mobile devices and Internet of Things (IoT), as a major source of Big Data that offers important value (Manyika et al., 2011; Chen, Chiang and Storey, 2012; Chen and Zhang, 2014). However, unstructured sources of data contain highly heterogeneous data for which organisations might require different means of acquisition, extraction and cleaning (Jagadish et al., 2014).

It is noteworthy that in most cases, even the structured datasets used in Big Data projects or applications may contain data from several sources. It is possible that in most cases, each of these datasets might have been created for different purposes and systems. Integrating such datasets for a single purpose could be challenging. Moreover, it becomes even more challenging to integrate and analyse such data together with unstructured data. Although several tools and techniques for dealing with the challenges of Big Data have emerged, technical issues of data collection, storage and analysis continue to be a source of concern (Chen and Zhang, 2014).

## **Ethics and privacy**

Organisations now have access to more datasets that are relevant to their business than ever before. In many cases, such data will include data on personal or sensitive information such as health information, location information or data relating to personal finance. Access to such data brings the challenge of ethical considerations, ownership of such data and privacy of people. This is because of the ambiguity that surrounds the ownership of such data especially when the same are available in the public domain. The extent to which organisations can make use of these datasets could also be a source of concern. Some of these datasets are created by users daily without realising the importance of giving away such data. In other cases, some pass across such important data to organisations unknowingly (Jagadish et al., 2014).

Datasets with personal information have a tendency to offer extra benefits because of the level of details they contain. Therefore it becomes an ethical consideration for organisations whether they should use the data, how should the data be used and to what extent it should be used (Minelli, Chambers and Dhiraj, 2013). Organisations that collect such sensitive data are also faced with added responsibilities to ensure that such datasets are kept safe as required by the law. Therefore, it can be concluded that venturing into the use of Big Data requires careful legal and ethical considerations and well defined internal policies regarding the use of data.

### **Talents**

Extracting insight and value from Big Data requires multi-disciplinary skills in fields such as Computer Science, Information Systems, Data Visualisation, Statistics, Machine Learning, Communications and domain knowledge. An emerging role known as data scientist has been described as the closest fit for talents and skills that align with the use of Big Data. But finding such skilled professionals has been described as one of the main challenges facing organisations seeking to take advantage of Big Data (Davenport and Patil, 2012). Whilst training such talents might require universities to implement changes to their curriculum, a major challenge is the rate at which the industry is evolving. This is one of the reasons why the challenges posed by the shortage of skills for Big Data endeavours has become important (Manyika et al., 2011; Davenport and Patil, 2012). Considering forming a team of experts rather than looking for individuals with all the skills could be beneficial (O'Neil and Schutt, 2014), but that also comes with the challenge of increased costs.

The tools techniques and approach of Big Data described above can be followed in sourcing and integrating disparate dataset for the purpose of developing actionable insights. The rest of this Chapter will explore the relevant data types that can be acquired and integrated for the purpose of providing decision support for grain commodities trading.

### **4.3 Data for Grain Commodities Trading Decision Support**

It was identified in Chapter 3 that data will play an important role in developing a DSS for grain commodities trading and that price discovery is an important factor in decision making in trading grain commodities. It was further concluded that there are several factors that influence the price of grain commodities in South Africa. This study gives specific attention to the price of white maize in South Africa and there are various factors influencing its prices. In order to make effective trading decisions, it is important to collect and extract valuable insights from the historical and real-time data on each of the factors that influence the changes that occur in the price of white maize in South Africa.

The remainder of this chapter is dedicated to identifying all the relevant datasets to create an integrated dataset for a DSS grain commodities for trading. Based on the review of literature in Chapter 3, it is expected that a variety of datasets needs to be collected and integrated to form the single source of rich data. These will include datasets that are created at short intervals; some of which will have large volumes. Therefore, a Big Data approach will be followed in the rest of this study in gathering, integrating and extracting insight that can be used in DSS for trading in grain commodities.

#### **4.3.1 Data Sources**

Data can be classified as primary or secondary data based on its sources. Primary data is created, collected and used for the purpose for which it was created while secondary data is data that has been created for a purpose but made available to be used differently (Collis and Hussey, 2009). An example of primary data is the data created during research experiments and used for analysis during the research. However, the same data can be used by practitioners for decision making as secondary data. The evolution of cloud computing and Big Data has led several organisations to make large volumes of data available to the public as sources of secondary data. This is based on the open data concept.

In recent years, the concept of open data has received attention among researchers and practitioners, particularly in fields relating to politics and economic policy development (Janssen, Charalabidis and Zuiderwijk, 2012). However, the concept is broader than making government data available to the public for transparency and accountability. Borglund and Engvall (2014) suggested that the concept of open data is beyond accessibility but that open data also encourages reusability of data. Therefore, data made available for use with minimal restrictions in fields such as geography, weather and business can also be regarded as open data (Janssen, Charalabidis and Zuiderwijk, 2012). The availability and use of such open data have the potential to drive innovations, improve existing business processes and lead to the development of new products and services (McLeod, 2012; Janssen, Charalabidis and Zuiderwijk, 2012).

As outlined in Chapter 3, the grain commodities trading DSS will require the collection of data on the factors that influence the price of the grain commodities in South Africa. The data for most of the factors is available in public domain, although mostly in fragmented, highly unstandardised, semi-structured and unstructured formats. These can be collected, integrated and analysed as secondary data for developing a grain commodities trading DSS. The rest of this section will describe the sources of such data, as well as how they can be acquired and integrated.

### **Market data**

Market data comprises the data collected on the executed and pending transactions on grain commodities trades. There is a need to source local trade data on the trade of the main commodity (white maize for this study) as well as international market statistics for the same commodity in countries where their market affects that of South Africa. This should include trade data like price, volume traded, bidding prices etc as provided by the exchange. The market data should also include the same type of data for other grain commodities that are considered as substitutes. In the case of white maize, the main substitutes are yellow maize, wheat and sorghum that serve as alternative sources of food for human consumption and for animal feeds, hence their prices are interdependent (Wright, 2011). In order to analyse and provide grain

farmers with intelligence on the different trading strategies, data should also be collected for each of the different trading strategies.

### **Demand, supply and storage data**

The data required to represent the effect of demand, supply and storage on the price of white maize in South Africa can be sub-divided into three categories. These are local demand, supply and storage data; the consumption and utilisation data in countries that influence the price of white maize in South Africa; and weather – local and international. Each of the categories are made up of several variables that are collected and made available by government agencies as open data or on request.

The other important factor under this theme is the influence of weather on the production of grain commodities. Several authors have indicated that weather conditions and the outlook of climatic conditions affect the price of grain commodities because of the direct impact that they have on the production and eventual supply of grain commodities (Geyser and Cutts, 2007; Wright, 2011; Trostle, 2008). Hence, weather data such as precipitation as well as minimum and maximum temperatures for areas where the grains are planted, could add value to the DSS.

### **Macroeconomics data**

The required macroeconomics data as highlighted in Section 3.4.2 include data on the South African Rand–US Dollar currency exchange rates, the price of crude oil, the local interest rate and the consumer price index in South Africa. The price of crude oil is a global economic variable that is monitored in different fields of interest because of its overarching influence on the global economy. Mostly, the data is available through stock exchanges and data brokers for a fee and as open data.

The other macroeconomics data required are variables that indicate the state of the South African economy. The South African Reserve Bank (SARB) has the responsibility to “protect the value of the currency of the Republic in the interest of balanced and sustainable economic growth” according to the South African Bank Act 90 of 1989 and as amended in 2003. Therefore, it is the responsibility of the SARB to look after the value of the South African currency which is the Rand. As South Africa’s

central bank, SARB is also responsible for coordinating the growth of the economy by determining the optimal value of important rates such as the interest rate at which banks are allowed to provide loans (Burda and Wyplosz, 2009). As a result, the SARB is the custodian of this economic data. The Consumer Price Index is a measure of the changes in the price paid by consumers for goods and a measure of inflation in the economy (Burda and Wyplosz, 2009). This is handled by Statistics South Africa, a South African government agency that is responsible for the collection, production and dissemination of all official statistics in the country (Statssa, 2015).

### **Political factors**

The changes that occur in the prices of grain commodities cannot be separated from either the direct or indirect influences of the national political climate. An example is market speculation that introduces shocks to the market when there is a change or expected change in the political landscape (Headey and Fan, 2008; Wright, 2011). However, studies on the impact of political activities and events on financial markets such as grain commodities prices are usually subjective. This makes it almost impossible to collect quantitative data for the purpose of analysis.

The developments in the use of Internet technologies for social activities offer new opportunities for collecting useful data about political climates (Kouloumpis, Wilson and Moore, 2011; Cambria, Schuller, Xia and Havasi, 2013). New Internet technologies such as blogs and social media are enabling more people to voice their opinion and make their feelings known regarding political issues (Ayankoya, Cullen and Calitz, 2014). The streams of data obtainable through social media platforms can be analysed for public sentiments regarding any issue (Pang and Lee, 2008). Thus, an avenue is provided for collecting data for quantitative analysis of how political factors influence the price of grain commodities. Twitter is one of such social media platforms that allow users to send text messages that are visible to those they are connected to on the platform. These can be collected and analysed for sentiments on topics relating to political influence on grain commodities prices. However, this will not be included in this study due to limited resources.

### **4.3.2 Integrating disparate data**

The challenge of integrating data from several sources has existed before the dispensation of Big Data. However the difficulties of volume, velocity, variety, the need to carry out complex analysis and extracting value from data in real-time makes integrating Big Data a more difficult problem. The need to integrate disparate data is even more important considering that data is more valuable when it can be linked with other data (Dong and Srivastava, 2013). As a result, the value of Big Data depends on the ability to integrate data from several sources. Therefore, getting the best out of a Big Data implementation requires that the peculiar challenges associated with Big Data integration be considered.

Integrating disparate data, especially when Big Data is considered, requires that the issues of heterogeneity, scope of data, dealing with data inconsistency, scalability and optimisation of how data is used be carefully considered (Chen and Zhang, 2014; Kadadi, Agrawal, Nyamful and Atiq, 2014). Dealing with these issues require the choice of enabling technologies (Jagadish et al., 2014) together with an approach to system design and implementation that takes pertinent issues into consideration. Section 4.3.1 outlined the disparate sources of the data required for a grain commodities trading DSS. The mode of accessing the dataset available from each of the sources is heterogeneous. In dealing with these challenges, adequate consideration was made for the need to collect historical data, update the data storage with new data points when created and the transformation of the data from the source into a suitable format.

In order to collect the historical data for the variables under each of the categories of data required for this study, suitable data-provisioning methods for each of the sources may need to be adopted to ensure that real-time data is collected. Scripts that transform and enable data streaming could be set up to update the datasets at regular intervals using streaming functionalities available in new technologies designed for Big Data.



## 4.4 Conclusion

This chapter explored where the datasets required for a grain commodities trading DSS as identified in Chapter 3 can be acquired. The need to take a Big Data approach was identified, hence this chapter also discussed important techniques, technologies and challenges that should be considered in extracting and using and value from Big Data. Data sources for the factors that were discussed in the previous chapter were identified and broken down to a specific data point. However, data on political factor from Twitter will not be incorporated in this study due to limitations of resources in accessing the data. Also, data on the consumer price index will not be incorporated for the same reason. It was noted that, although the data from these sources is complementary, the structure of the data from these sources was heterogeneous. Therefore, transformation and the integration of the data were identified as imperative in order to extract the best value from the datasets.

This chapter highlights the creation of a data repository that integrates market statistics and datasets of factors that influence the grain commodities trade as the foundational component for a framework a DSS for grain commodities trading. Data from each of the sources can be streamed into a data repository as it becomes available. Each of the datasets contain data which could be collected over successive time intervals – hourly, daily and monthly. These datasets can be integrated as a time series with some of the datasets aggregated or disaggregated in order to create an evenly timed series data.

This chapter complements the research objective RO<sub>1</sub> – to identify data-related requirements for a system to support decisions on trading grain commodities in South Africa. It also provided the answer to research question RQ<sub>3</sub> – What datasets influence the prices of grain commodities in South Africa? The outcome of this chapter provides an initial component of the DSS trading in grain commodities as having access and being able to integrate historical record and updates of relevant datasets as a Big Data repository. It also identified the Big Data approach and considerations as an enabling factor in the acquisition of the relevant datasets and in the setting up of the environment that, by for extracting valuable intelligence, can support decisions about

trading grain commodities in South Africa. Thus, within the DSR process, Chapter 4 provides the building blocks that will be considered in the development of a framework that can be used to develop a DSS that can support grain commodities trading decision making. This forms a part of the rigour cycle of DSR and the environmental issues raised form a part of the relevance cycle of this study within the DSR cycle.

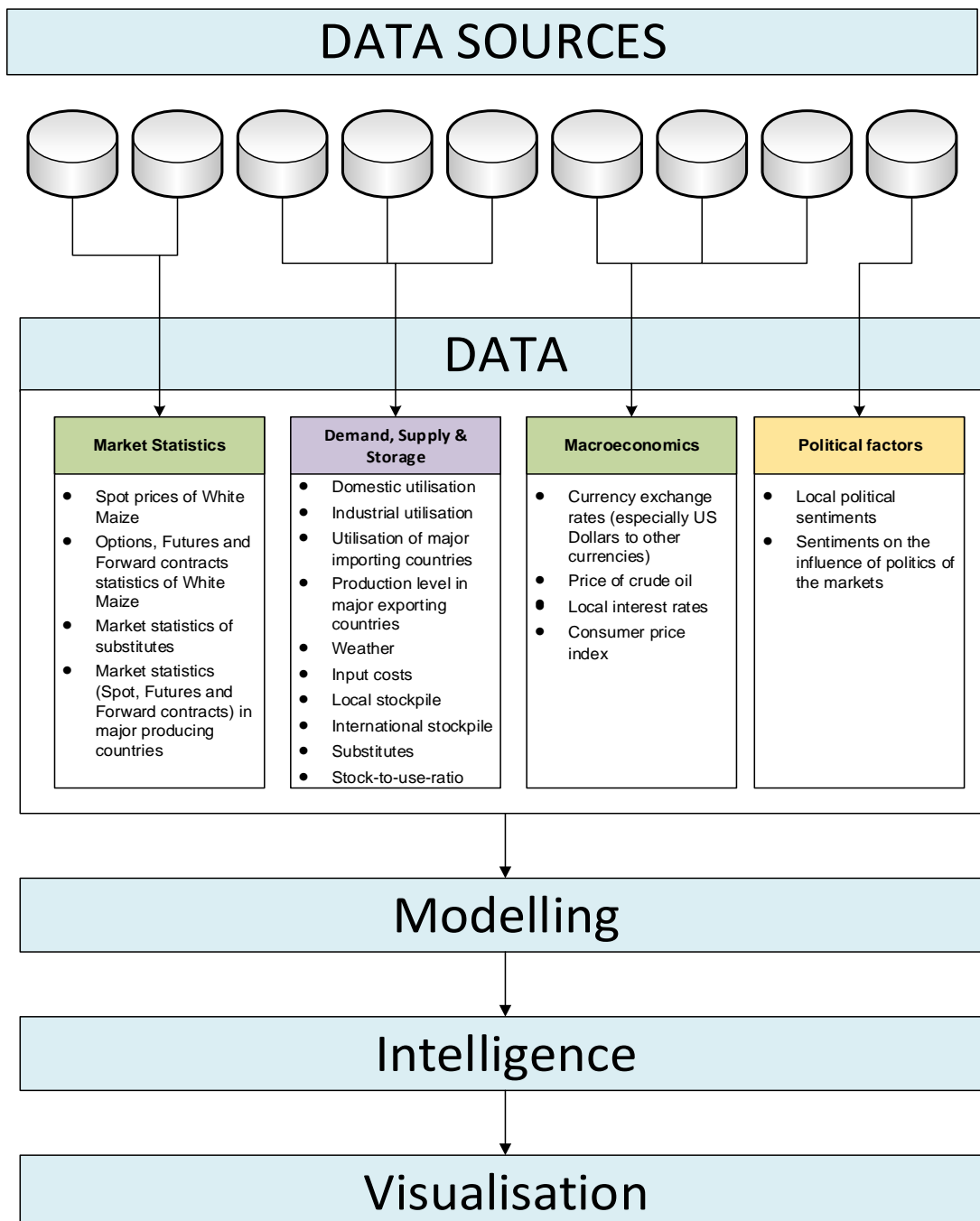


Figure 4.3: Initial framework for grain commodities trading DSS

Chapter 3 identified data, modelling, intelligence and visualisation as the main components of a Decision Support System. Figure 4.3 presents a proposed grain commodities decision support framework showing the main components as the pillars. The data component of the framework shown in Figure 4.3 highlights the data requirements for a grain commodities trading DSS as identified in this chapter. It was also identified that careful technology considerations are important in the acquisition, preparation, transformation and integration of data from disparate sources for a grain commodities trading DSS.

The next chapter will focus on identifying the patterns that exist in the data that has been collected. Attention will also be given to extracting intelligence and valuable insights from the grain commodities trading data.

# Chapter 5 : Market Intelligence and Predictive Modelling

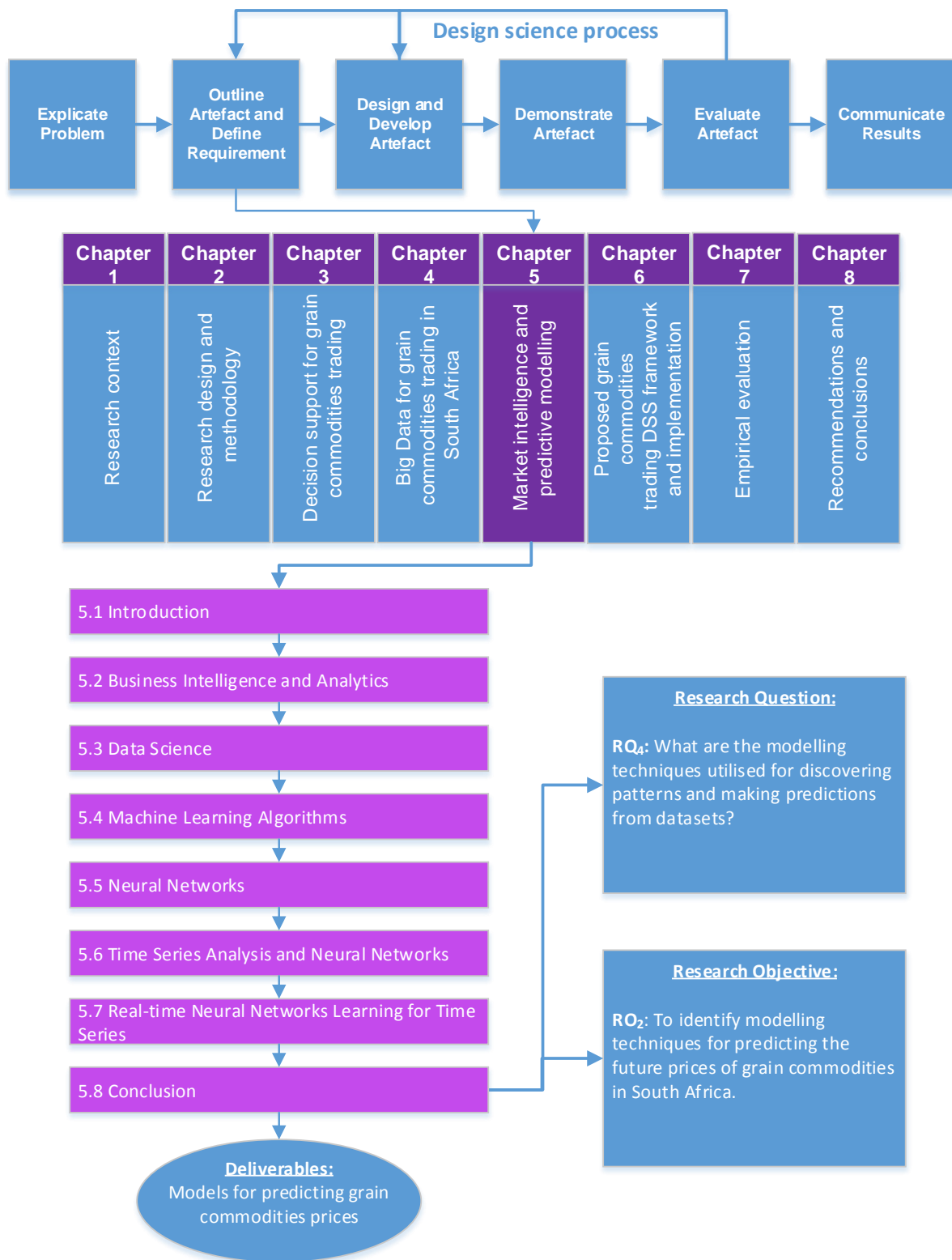


Figure 5.1: Chapter outline and deliverables

## 5.1 Introduction

The previous chapter explored the ubiquitous availability of data that is referred to as Big Data. It was identified that Big Data presents new opportunities for improved decision making especially in making discoveries and insights for decision support. However, in Chapter 4 the fact that the traditional approach of extracting value from data will not suffice for obtaining competitive advantage from Big Data because of the complex characteristics of such datasets. Therefore, it was identified that dealing with large datasets with the characteristics of Big Data require new approaches, tools and technologies. The Big Data issues, tools and techniques that are pertinent to extracting value and decision making were also discussed in Chapter 4.

As part of the requirements for improving trading decision making, in grain commodities it was identified in Chapter 3 that the discovery of optimum price and management of price-related risks are key decisions for grain commodities trading. Hence there is a need to understand the factors that influence the prices of grain commodities. These factors identified in Chapter 3 and Chapter 4 provided a list of datasets on prices of grain commodities and the factors that influence the prices. It was further identified that all the datasets are collected as time series data quarterly, monthly, daily or more frequently in some cases. It was further discussed in Chapter 4 that these datasets can be streamed and integrated as time series data to form the basis for providing support for decisions on grain commodities trading.

Streaming and integrating relevant large datasets on grain commodities trading and the factors that affect the market potentially exhibits characteristics of Big Data. Therefore, a Big Data approach would be beneficial in extracting insight and managing the complexities associated with the streaming and integration of the datasets. As part of understanding the requirements for a grain commodities trading DSS, this chapter will review how insights and discoveries that support decision making can be extracted from large datasets. This chapter will address the second research objective – RO<sub>2</sub> of this study, which is *“to identify modelling techniques for predicting the future prices of grain commodities in South Africa”*.

In order to achieve the set objective, Chapter 5 will focus on providing answers to research question RQ<sub>4</sub> – “*What are the modelling techniques utilised for discovering patterns and making predictions from datasets?*” Specific attention will be given to identifying techniques that can be used for extracting patterns from time series data. It is expected that this chapter will identify suitable predictive analytics methods, tools and techniques by using Big Data that can be used to discover future prices of grain commodities.

Within the Design Science Research methodology that is followed in this study, Chapter 5 will further contribute to the design and development of the framework for a grain commodities trading DSS, which is the expected artefact from this study. The initial part of the rigour cycle of this study in Chapter 3 identified the importance of modelling in the development of a DSS that supports or improves decision making for any practice. This chapter continues with the rigour cycle by researching the relevant knowledge base for methods and techniques for extracting market intelligence and predictive modelling, which can be incorporated into a DSS that supports decisions about grain commodities trading. Chapter 5 will focus on identifying methods and ideas that are scientifically feasible for the building of what could be the modelling component of the eventual decision support framework from this study.

Section 5.2 will review the concepts of Business Intelligence and Analytics, while Section 5.3 will examine the concept of Data Science. It is expected that both sections will provide the foundations for extracting value and making use of data for decision making. The section will examine different types of analysis and the processes that could be followed in order to extract insights that support decision making from data. The use of Machine Learning Algorithms for predictive modelling will be introduced in Section 5.4 and Section 5.5 will focus on Neural Networks algorithms for solving difficult problems and identifying complex patterns. In Section 5.6, the use of Neural Networks algorithms for time series problems will be examined and compared to the use of statistical, time series models for making predictions. Section 5.7 of this chapter will examine how predictive models can be used in real-time, considering the complexities that might be associated with using Big Data for Real-time analytics. It is expected that Chapter 5 will provide an insight into how market intelligence can be

developed from Big Data by using predictive models and how this can be incorporated into real-time analytics for decision support. Section 5.8 will provide a conclusion of this chapter.

## **5.2 Business Intelligence and Analytics**

Business intelligence and analytics are data-centric approaches that complement data with a set of methodologies, processes, technologies and tools for analysing and extracting information from data (Davenport and Harris, 2007; Chen, Chiang and Storey, 2012; Lim, Chen and Chen, 2013). Business intelligence has been the focus of much earlier attention as sets of methodologies and processes. It was used as an enhancement of relational databases for business support and reporting. However, the introduction of business analytics provided opportunities for the application of analytical techniques that allow for data-driven decision making and management of organisations (Chaudhuri, Dayal and Narasayya, 2011; Chen, Chiang and Storey, 2012). While the focus of intelligence is more on the provision of dynamic access and reporting of data, analytics offer an opportunity for extracting knowledge and insight from data (Watson and Wixom, 2007; Chen, Chiang and Storey, 2012).

On-Line Analytical Processing (OLAP) provides analytical functions for extracting information and knowledge within the data-warehousing framework (Provost and Fawcett, 2013b). OLAP enables the multidimensional integration of data from different sources for tasks such as aggregation, summarisation and filtering for better decision making (Chaudhuri, Dayal and Narasayya, 2011). This is preceded by manipulation and transformation of this data using the Extract, Transform and Load (ETL) processing to ensure data integrity and consistency.

In recent years, Business Intelligence and Business Analytics have become more complementary and synchronised and are addressed together as Business Intelligence and Analytics (BI&A) (Chen, Chiang and Storey, 2012; Lim, Chen and Chen, 2013). Overall, the two have become a prominent and almost indispensable set of tools in data management and decision making. Moreover, Business Intelligence

and Analytics have evolved and are influenced by the development of different types, sources and volume of data.

Traditional BI&A efforts are mostly based on structured data that are stored using RDBMS technologies (Chen, Chiang and Storey, 2012). However, the latest developments in the nature, type and management of data have necessitated the use of more advanced analytical tools. In recent times, data-driven decision-support efforts need to consider other sources of data that are mostly unstructured and created as a stream rather than warehoused. This implies that new sources of data such as documents, videos, audio files, user-generated contents on social media, Internet of Things (IoT) and mobile devices/applications have to be considered. Moreover, the rate at which data is being created for many of the new sources of data suggest that if the traditional BI&A processing is followed, the resulting decision could be obsolete.

Earlier applications of BI&A focused on the creation of reports, dashboards, scorecards in order to understand what happened in the past (Chen, Chiang and Storey, 2012; Chen and Zhang, 2014). Although BI&A was also used for some degree of predictive modelling in the past, it has been noted that this was constrained with several limitations (Minelli, Chambers and Dhiraj, 2013). The characteristics of Big Data present new opportunities to use Big Data with new analytical tools and techniques to predict the future and make new discoveries with capabilities to do all that could be achieved with traditional BI&A (Provost and Fawcett, 2013b). This highlights the need for new approaches, tools and techniques for extracting intelligence and analytics that support decision making from Big Data.

### **5.3 Data Science**

Data Science combines a collection of tools and techniques for extracting insights from large datasets (Ayankoya, Calitz and Greyling, 2014). It utilises mathematical methods, statistical computing and various scientific optimisation methods in the collection, integration, analysis and extraction of insight from Big Data. This makes it possible to deal with the complexities of Big Data (Chen and Zhang, 2014). Data Science combines the opportunities and the potentials of Big Data, BI&A, advanced



analytics and the understanding of a particular field to extract value from large volumes of data.

Although Data Science involves the application of statistical methods, it is important to emphasise that Data Science is beyond the application of statistical analysis of Big Data. It is a combination of Computer Science, Data Visualisation, Machine Learning, Statistics, Mathematics, Communication and Domain Expertise skills for extracting meaning for data-driven decision making from complex data (Patil, 2012). Furthermore, Data Science involves principles, processes and techniques for understanding different phenomena and for improving decision making (Provost and Fawcett, 2013b). Thus, the focus is making discoveries and providing answers to difficult real-life questions by using Big Data (O’Neil and Schutt, 2014). The different perspectives on Data Science indicate that it offers opportunities for data-driven decision making, predictions, discoveries, recommendations and a different approach to providing solutions both in research and in practice.

The role of traditional BI&A in decision making has mostly been the discovery of “what happened?” mainly from historical/warehoused data. Moreover, some of the traditional BI&A systems make use of models for some degree of predictive analysis such as the what-if analysis, but Data Science enables the discovery of “what is likely to happen in the future?” and “what will lead to competitive advantage” by using a combination of historical and real-time datasets from several sources (Davenport and Patil, 2012; Dhar, 2013). This involves the use of advanced analytics such as predictive analysis, modelling and machine learning for prediction, recommendation and discovery (Davenport, 2014).

Using Data Science involves an iterative process that combines multi-disciplinary tasks. Figure 5.2 presents a suggested schematic flow of the Data Science process (O’Neil and Schutt, 2014). The process includes data acquisition which involves the collection, processing and cleaning of raw data. These tasks are iteratively linked to exploratory data analysis to properly define the problem and ensure that the right data is collected. This task requires knowledge of the domain of interest. Thereafter, statistical computing/modelling and visualisation tasks are carried out to provide

insights, discoveries and recommendations that can be used for decision making. However, the execution of the Data Science process requires that enabling technology be considered as an important factor. This is due to the nature of the input datasets and the need to provide real-time insight in most cases (Minelli, Chambers and Dhiraj, 2013).

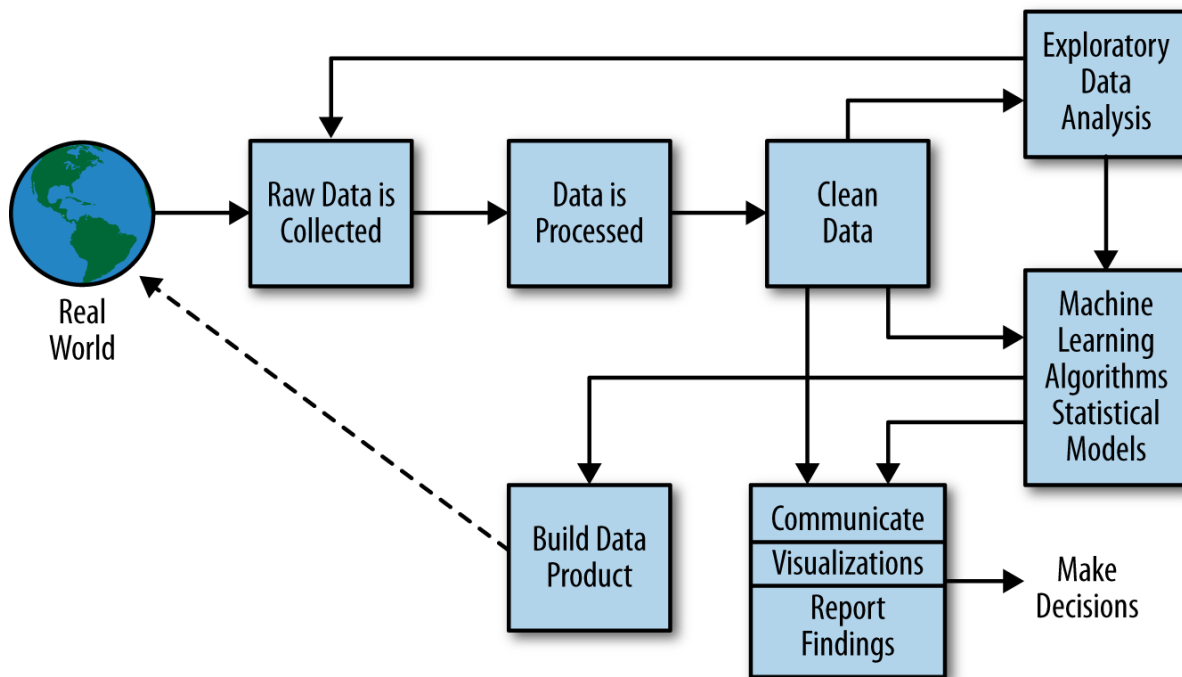


Figure 5.2: The Data Science process (O’Neil and Schutt, 2014)

The process illustrated in Figure 5.2 further highlights Data Science as a set of principles, techniques and processes for using data to understand different problems. Moreover, this process is synonymous with the Cross Industry Standard Process for Data Mining (CRISP-DM) framework that has been popularly adopted for traditional Data Mining efforts (Provost and Fawcett, 2013b). Data Science, however, goes beyond the application of Data Mining algorithms for extracting knowledge from data. It draws from several fields of interest to create a “data-analytic” approach to solving problems right from the identification and definition of the problems through the process of developing and implementing the solution (Provost and Fawcett, 2013a). This also involved the ability to deal with larger datasets and automation of the entire process to deal with a changing problem definition and desired solutions.

The Data Science process presented in Figure 5.2 indicates the acquisition of data from different real-world activities such as real-time air travel data, social media data or real-time market data from a stock exchange. Some of this data can be collected as historical data, near real-time data such as hourly or end-of-day data or it can be streamed as real-time data as it is being created (Chen and Zhang, 2014). Depending on the characteristics of the data that is being collected and what the data will eventually be used for, the data is pre-processed and cleaned to make the data suitable for use. Pre-processing and cleaning could include normalisation where possible, transformation (e.g. logarithmic conversion), removal of outliers and handling of missing values (Engelbrecht, 2007; O'Neil and Schutt, 2014).

After data pre-processing, Exploratory Data Analysis (EDA) can then be performed to understand structures that are embedded in the data and to suggest the best modelling strategies for extracting valuable decision support from the data. This is followed by the use of algorithmic models such as statistical models, machine learning or a combination of both to extract intelligence that supports decision making. The results of modelling can then be presented for the benefit of decision makers using visualisation techniques or as data products. Data pre-processing, EDA, modelling and presentation of results will be explored in more detail.

### **5.3.1 Data pre-processing**

Data does not always come in formats that are ready for analysis. This is particularly true with Big Data where some or all of the data could be unstructured, in several formats, prone to inconsistencies and from multiple sources. These characteristics lower the quality of data which in turn reduces the quality of results that will be produced if such data is used (Han, Kamber and Pei, 2012). Therefore, it is important to understand these risks by identifying the weaknesses of the data or potential weaknesses where the data is collected and used in real-time (Provost and Fawcett, 2013b). The quality of data used for extracting intelligence that supports decision making can be ensured by pre-processing the data before use. Data pre-processing techniques help in addressing potential issues that can affect the outcome of BI&A or

Data Science efforts. The following data pre-processing techniques are identified from the literature:

- **Data Cleaning:** Big data is susceptible to quality issues because of heterogeneity, inconsistency, incompleteness, outliers, duplicate and noisy data (Chen and Zhang, 2014). This can overshadow the potential opportunities and even lead to counter-productive results. Data-cleaning techniques provide routines and rules that can be used to deal with data-quality issues. These could include filling missing data with most probable options that are based on other attributes of the dataset or the use of statistical techniques to identify and correct outliers (Han, Kamber and Pei, 2012).
- **Data Transformation:** Many of the statistical and computational intelligence techniques that are used for analysing and extracting value from data require that the input data be presented in a specific format. The need to ensure that data is in a unified format is of more relevance with Big Data that is a result of the integration of a variety of data. Without the application of appropriate data transformation techniques, the ability to extract insight from data becomes diminished and whatever such data is used to discover, becomes questionable (Williams, 2011). Data transformation techniques can be applied to datasets that are collected or available in different units or frequencies to ensure logical integration. Examples of transformation techniques include the transformation of datasets to make the range fall between 0 and 1 or between -1 and 1 in order to deal with the issue of multiple measurement units, known as normalisation (Han, Kamber and Pei, 2012). Other techniques include aggregation, disaggregation, standardisation, scaling, coding and logarithmic transformation.
- **Data Integration:** Datasets from several sources that are integrated to become a single dataset potentially become more valuable than combining the value of each of the datasets individually. This is evidenced in the use of more attributes to increase the accuracy of data driven-models (Davenport, 2014). However, there is also the possibility of introducing bias or quality issues in the resulting data. Proper attention needs to be paid to the understanding of the datasets to

be integrated, so that the best-matching attribute can be used to establish relationships among the datasets. It is also important to ensure that redundancy is avoided in the resulting dataset by ensuring that unnecessary attributes are ignored (Han, Kamber and Pei, 2012). This can be achieved through the use of a correlation matrix, test of independence or other such tests depending on the type of datasets that are being integrated.

- **Data Reduction:** After integrating several datasets, the resulting dataset is likely to be made up of many variables. Big Data technologies such as distributed computing, parallel processing and in-memory computing have emerged as solutions for working with datasets that have many variables within a reasonable processing time (Chen et al., 2013). However, besides the constraints of processing time and power, datasets with a large number attributes can become too complex to manage and understand. There are several mathematical techniques that have been used in traditional data mining for reducing the dimensions (no of variables/attributes) of a dataset without necessarily losing useful information (Han, Kamber and Pei, 2012). These techniques are specifically relevant to Big Data analytics. One mathematical technique utilised is Principal Component Analysis (PCA) that is used to reduce the dimension of a larger dataset by combining highly correlated variables (O'Neil and Schutt, 2014).

Several of the technologies that support Big Data have the data pre-processing tools as standard features. It has been suggested that these tools should be implemented as a separate application layer interfacing directly with the data-collection layer (Chen and Zhang, 2014). This will make it possible to be able to collect/stream data and make use of the data for real-time processing. Despite the possibility of implementing data pre-processing as an application, it is also important to continuously re-evaluate the implemented pre-processing system for the unexpected event (Chen et al., 2013). Hence, it is suggested that data pre-processing and exploratory analysis should be combined and possibly be implemented in iteration.

### **5.3.2 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is used to get a better understanding of data and to determine the best approach for extracting useful information from data. Before building models that will form the foundation for extracting patterns, discoveries and recommendations from data, EDA can be performed to confirm intuitions about the value or limitations of the data (Provost and Fawcett, 2013b). Performing EDA reveals the structure, relationships and patterns that may exist in the data. This could uncover useful concepts or attributes that might be embedded in the data that might have normally been overlooked.

Before developing a model that represents or explains the relationship that exists in a collected dataset, it is suggested that the EDA should be performed for a better explanation of each of the variables and to compare each of the variables in the data and thereby build a case for choices that will be made during modelling (Larose, 2005). This can be achieved by examining the descriptive statistics for each of the variables such as the mean, median, minimum and maximum values, for continuous variables. The use of graphs and plots have also been found to be very useful for providing visual insight and diagnosis of the data.

How to determine the input variable to be included and/or not included is an important decision (Engelbrecht, 2007). Furthermore, when building models, the presence of two or more variables that are highly correlated could negatively impact the outcome of the study. This challenge can be dealt with by using correlation analysis to determine the strength of the relationship that exists between each of the variables (Larose, 2005). The purpose and possibilities of EDA show that it should be used together with data pre-processing iteratively to ensure that high quality data is used as input for modelling or analysis (O'Neil and Schutt, 2014).

The Data pre-processing and Exploratory Data Analysis components within the Data-Science processing are similar to what is obtainable with traditional Data Mining/data analysis. However, the Data Science approach, as a whole, differs from the traditional approach of dealing with data. Beside the fact that the Data Science process caters for the characteristics of Big Data described in Chapter 4, in most cases the emphasis

is on the discovery of hidden patterns in the data that provide actionable insights. These are particularly useful because of the predictive insight that is offered by using the historical and real-time data to predict future occurrences with a high level of confidence (Dhar, 2013).

Data Science also seeks to present the outcome of the process in several formats other than what was obtainable with data mining or statistical data analysis. In several cases, it seeks to make use of large datasets as they are being created to provide insights in real-time or near real-time. This is achieved by enabling the use of Application Programming Interfaces (API) to expose the result of knowledge discovery tasks such as the use of algorithmic modelling, such as machine learning or other techniques (Loukides, 2010).

## **5.4 Machine Learning Algorithms**

Machine Learning (ML) is the use of computer systems and algorithms to learn from data and adapt in order to provide a continuously relevant solution for problems in a dynamic environment (Alpaydin, 2010; Bell, 2015). It involves the use of algorithms to train the computer systems to be able to make intelligent decisions, as the need arises, based on complex patterns from previously collected data. ML has been used for decision making in areas such as fraud detection, risk management, classification of SPAM emails, facial and voice recognition as well as in several areas of medical research such as cancer research and DNA testing (Lantz, 2013).

In several of its applications, the end-goal is to extract useful knowledge for decision making in complex environments. This is achieved by making use of mathematical models as the basis for decision making (Alpaydin, 2010). Various statistical models that are also used in decision making also have their foundation in mathematics. However, Machine Learning is primarily part of Computing Sciences while statistical modelling forms part of Statistics. Hence the perspectives of ML may be different from those of statistical modelling. Statistical theories are complementary to ML in several areas (Alpaydin, 2010; O'Neil and Schutt, 2014).

Statistical modelling techniques are constrained by the need to find real life explanations for the meaning of input parameters and relationships that exist among the input parameters and the output variables (Dhar, 2013; O'Neil and Schutt, 2014). Machine Learning techniques however, are not only able to create models that understand embedded patterns in complex datasets, they are designed from a software perspective to iteratively seek the most efficient solution to problems considering factors such as time and space (Alpaydin, 2010). The techniques used in Machine Learning can be broadly categorised into three categories;

- **Supervised Learning:** This is used for problems where there are labelled input variable(s) together with the output variable(s) for each of the observations in the data that the Machine Learning Algorithm is learning from (known as the training set). This provides the algorithm with the pattern in the training data. Based on the training set, the algorithm is expected to be able to suggest the output when presented with a set of input parameters. Supervised learning techniques are better suited for solving regression, classification or causal modelling problems (Provost and Fawcett, 2013b). Examples of supervised learning techniques include Decision Trees and K-nearest neighbour.
- **Unsupervised Learning:** In this case, the input variable is unlabelled and there are also no output variables. The goal of unsupervised learning is to allow the algorithm to identify the different groups that might exist in the data without any supervision (Bishop, 2006). This is applicable in cases where the required solution is to identify different clusters or dimensions that might exist in the data.
- **Reinforcement Learning:** There are problems where there are neither input variables nor output variables, but the problem is about making a sequence of decisions in a dynamic environment in order to achieve an end-goal (Alpaydin, 2010; Osman, El-Refaey and Ayman, 2013). Therefore the algorithm is trained to know what decision to make at every step and in the process there are no right or wrong choices until the final decision is made. An example of this will be the use of Machine Learning algorithms in the design and decoding of computer games.



The identified possibilities of Machine Learning complement the opportunities that Big Data offers. The characteristics of Big Data, however, also introduce new challenges in the implementation of Machine Learning techniques for extracting value from Big Data. New studies suggest that the ability to learn from complex data and have intelligent agents that make decisions could be beneficial in several areas because of the possibilities to collect and integrate several large datasets (Condie, Mineiro, Neoklis and Weirner, 2013; Osman, El-Refaey and Ayman, 2013). This includes the ability to utilise, learn and make decisions from both historical data, real-time and near real-time data, thereby making it possible to make intelligent decisions while eliminating the constraint of time.

### **The Machine Learning process**

Several types of algorithms based on the principles of learning from complex patterns that might exist in data for the purpose of decision making, have evolved over the years based on needs and advances in different technologies. Some of these technologies have evolved with particular strengths, abilities to perform better in some areas or simply more suitable for some kinds of problems. One such is Neural Networks which has also found use in several areas such as medicine, finance and engineering, especially for recognising patterns and making predictions. Others include Decision Trees, Clustering Algorithms, Support Vector Machines and so on. However, there is a generic process that can be followed in a Machine Learning project irrespective of the chosen.

A typical Machine Learning project should start with a clear understanding of the problem at hand, identification of the source of relevance and the choice of a suitable Machine Learning algorithm. Thereafter, the following steps can be followed in any Machine Learning project to provide solution to the identified problem (Lantz, 2013; Bell, 2015):

- **Data collection:** The first step in a Machine Learning project is to collect relevant data in an appropriate format. The collected data will serve as the input and learning material for the algorithm that will be used for providing actionable insights.

- **Data preparation and exploration:** The success of any Machine Learning algorithm is dependent on the quality of the data used for the learning process. Therefore, it is important to ensure that the data collected is cleaned, transformed, summarised and aggregated as required in order to get the best result. This should be supported by an exploration of the characteristics of the data in order to know how to use the data to obtain an optimum result.
- **Model training:** The next step after preparing the data is to use the appropriate Machine Learning algorithms to train a model to learn from the data. The model is an encapsulation of the patterns that exist in the data which can be followed in making new decisions or providing insights.
- **Evaluation of model performance:** In order to ensure that accuracy and the ability of the model to generalise, a test data can be used to evaluate the degree of accuracy of the developed model.
- **Improving model performance:** The previous step will provide an indication of the weakness or the strength of the developed model. Based on the result, the model can be improved using different methods such as changing the parameters and retraining the model, improving the data used and so on.

When a satisfactory model has been generated from the steps above, the results can then be presented in the desired format. The generated model can then be used for generating actionable insights such as making predictions. It should be noted that the process could involve the use of multiple Machine Learning algorithms in some cases, in order to achieve the desired result.

## 5.5 Neural Networks

Neural Networks is a branch of Machine Learning that is able to learn complex patterns from data for the purpose of solving difficult problems and making decisions. They are founded in the biological research in the facility of the neural system of the human/animal brain to learn, recognise, store information, generalise and make decisions based on prior knowledge. The human and animal brains are made up of millions of interconnected neurons with which the brain learns complex patterns, processes and stores information. It is based on this ability that the brain is able to

make intelligent decisions in nanoseconds (Larose, 2005). In the same manner, Neural Networks uses mathematical models to establish a relationship between a set of input values/signals and outputs (Lantz, 2013). Neural Networks are made up of layers of interconnected neurons comprising an input layer, hidden layers and the output layer as shown in Figure 5.3.

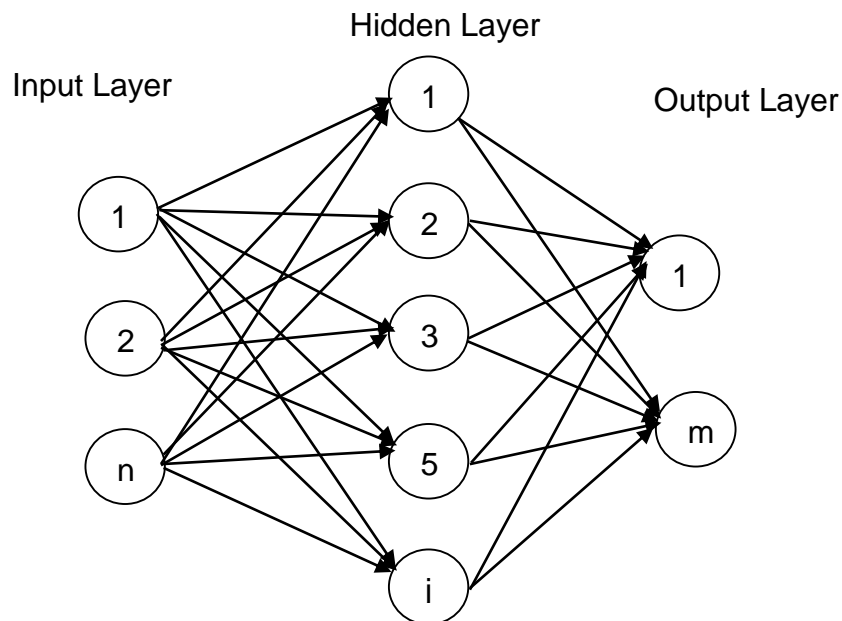


Figure 5.3: Simple Neural Network (Engelbrecht, 2007)

Figure 5.3 presents a diagrammatic representation of a simple Neural Network showing the input layer having input neurons 1 to  $n$ . The hidden layer contains neurons 1 to  $j$  with the output layer containing neurons 1 to  $m$ . The connections that exist between the neurons from each of the layers facilitate the learning process through the use of mathematical functions depending on the type of Neural Network and its setup (Engelbrecht, 2007). Therefore information and learning are done by sending signals between neurons from different layers. Each of these operations leads to the transmission of a signal to the neurons between the input layer and the hidden layer(s) and from the hidden layers eventually to the output layer.

During the learning process, how information is processed and transmitted throughout the network is determined by the activation function (Lantz, 2013). This transfer of

signals from the neurons across the different layers could either be weighted or not weighted depending on the architecture of the chosen network (Kriesel, 2007).

The connections from one neuron to the other, also known as signal is given weights that determines the level of importance of each signal in between the different layers (Lantz, 2013). The weights are allocated randomly during the learning process based on the error function until optimum outputs are identified. Hence the output in a Neural Network is a function of the weighted input signals received as transmitted by the activation function. A subset of the of Neural Network presented in Figure 5.3 can be used to provide further illustration by considering the neuron 1 in the output layer as receiving signals from neurons 1,2,3,4,5,...,j in the hidden layer. Figure 5.4 shows that the signals from each of the neurons 1,2,3,4,5,...,j have associated weighted values  $w_1, w_2, w_3, w_4, w_5, \dots, w_j$ , which are passed onto the output according to the activation function  $f$ .

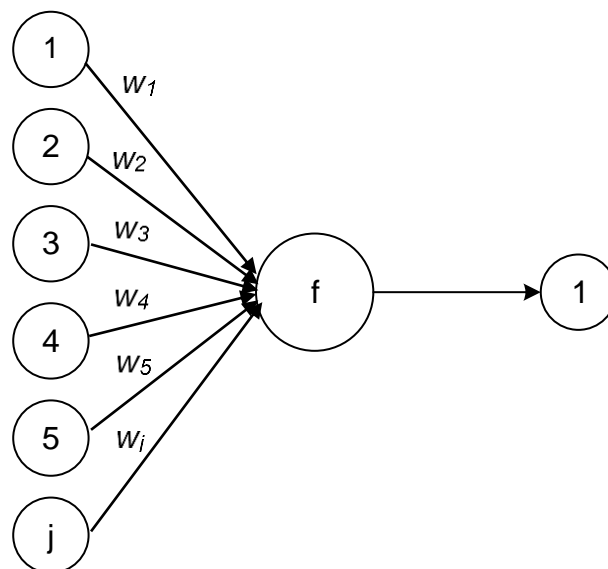


Figure 5.4: Diagram showing Neural Networks weighting (Lantz, 2013)

In order to ensure the best result, it is important to have an understanding of the problem in order to choose the right activation functions and other parametric settings that determine the behaviour of a Neural Network. Engelbrecht (2007) identifies the main activation functions that can be used in a Neural Network as:

- Linear function;

- Step function;
- Ramp function;
- Sigmoid function;
- Hyperbolic tangent; and
- Gaussian function.

Apart from choosing the right activation function to connect the neurons from each of the layers in a Neural Network, there are several other factors that must be taken into consideration in the design of a Neural Network that is efficient and suitable for the problem at hand. These include:

- **Learning rate:** The end-goal of a Neural Network is to identify an optimum learning option that minimises the error in the network (Larose, 2005). The learning rate,  $\eta$ , determines the number of steps that is taken in the search for the output. If the chosen learning rate is too large, the optimum can be missed and when it is too small, the network can take too long to train (Engelbrecht, 2007).
- **Momentum term:** The momentum term,  $\alpha$ , determines the degree of influence that the weights of previous learning will have on the current learning. It allows the training process to use the identified weights of the previous learning iteration, such that the weights of the past iteration are introduced as inertia in the current learning iteration (Larose, 2005). The momentum term ranges from 0 to 1, meaning that when the momentum term is close to or equal to one the weight of the current iteration will be essentially the same as the previous one.
- **Number of training iteration:** The learning process in Neural Networks is iterative, therefore the number of iterations for the learning process should be set from the beginning as an exit criteria for the network. However, depending on the selected learning rate or the momentum rate, it could take much longer to achieve the set number of iterations. In such cases, a target error level can also be set for which the learning process will be terminated when achieved (Larose, 2005). The selection of the optimum learning rate, momentum term

and exit criterion is a balancing act considering the implications that each of the parameters has on performance of the network.

Important choices that need to be made in setting up a Neural Network are those of features selection, the number of hidden layers and number of neurons for each layer that will deliver the optimum performance. Essentially, the determination of features that will make up the input neurons can be based on data pre-processing techniques as discussed earlier. On the other hand, the choice of the number of hidden layers can be discovered by experimenting, although it is suggested that one hidden layer should be sufficient in most cases and it is not advised to have more than two hidden layers (Larose, 2005). Finally, the number of neurons in the output layer also needs to be determined. This should be based on the expected outcome of the network and whether the learning process is supervised, unsupervised or is reinforcement learning. Selecting the number of neurons for any of the layers in a Neural Network has been described as a difficult task. The suggested rule of thumb is to keep the number of neurons in a network as small as possible so that the network can generalise and adjust easily to new patterns (Wilamowski, 2009).

It is possible to have different architectures of Neural Networks based on the type of learning (supervised, unsupervised and reinforcement) and the different topologies (designs) that are available. The architecture of a Neural Network determines what sort of application and the complexity of the problem it can handle. A Neural Network could have a feed-forward or a recurrent topology. A feed-forward network is such that the signals from the neurons propagate from the input layer through the hidden layer or layers and finally to the output layer with any signal feeding backwards as depicted in Figure 5.3 (Kriesel, 2007). On the other hand, recurrent Neural Networks are those able to have feedback signals besides sending signals forward in order to make provision for complexities of the problem being learnt (Engelbrecht, 2007; Alpaydin, 2010).

All architectures/topologies of the Neural Networks can be designed to handle problems from basic levels to more complex problems. But it has been suggested that Neural Networks should be kept as simple as possible with the minimum possible

number of neurons for optimum results (Wilamowski, 2009). The remaining part of this chapter will explore how Neural Networks can be used to learn the complex patterns that exist in prices of agricultural commodities traded on stock exchanges. The review of literature will look at extracting insight from commodities prices using the factors that affect the prices such that traders can make the right decision at the right time when trading grain commodities.

### **5.5.1 Backpropagation Neural Networks**

Several modelling algorithms exist for the different Neural Networks architectures for making approximations such as making predictions or classification. Wilamowski (2009) alluded that the choice of algorithm should be based on the type and complexity of the problem for which a model is being trained. The Backpropagation Neural Networks (BPNN) which are based on the feed forward Neural Network have been found to be widely suitable for problems requiring prediction from data such as a time series.

BPNN follows the multi-layer learning networks system where there is an input layer that is made of neurons representing the independent variable. They comprise one or more hidden layers with neurons which carry weight that determine the degree of influence during the learning process; these hidden layers enable the network to use a non-linear function to model complex patterns (Alpaydin, 2010). Finally, Backpropagation Neural Networks also contain an output layer with neurons representing the estimated variables. During the learning process, BPNN sends a signal about the error from the output back to the hidden layer. This ensures that subsequent learning produces an output with a lesser error until an optimal output is discovered (Alpaydin, 2010).

Other Neural Networks include the Radial Basis Function (RBF) Networks, Counter Propagation and Learning Vector Quantization, some of which have been reported to require a larger number of neurons to train (Wilamowski, 2009; Alpaydin, 2010). There are suggestions that BPNN might be slower than some other available Neural Networks (Wilamowski, 2009). Previous studies, however, have shown that BPNN is suitable for making predictions that are based on historical data even when they

involve complex patterns (Ghwanmeh, Mohammad and Al-Ibrahim, 2013; Tsadiras, Papadopoulos and O’Kelly, 2013). Hence, many time series-related studies such as financial forecasting, engineering and medical research have successfully implemented BPNN. Thus, this study adopts BPNN for the implementation of the Neural Network modelling of grain commodities prices component of the proposed framework.

Zhang (2003) and Qi and Zhang (2008) both suggested that the relationship that exists between the input variables and the output variables in a feed-forward Neural Network can be represented mathematically as:

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g \left( \beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t \quad (1)$$

where  $y_t$  is the expected output,  $\alpha_j$  ( $j = 0,1,2, \dots, q$ ) and  $\beta_i$  ( $i = 0,1,2, \dots, p$ ) represents the weights for the connections between the neurons in the hidden layer and the output nodes;  $p$  represents the number of input nodes and the number of hidden nodes is represented by  $q$  in the equation. The transfer function between the hidden layer and the output node is denoted by  $g$  which could be a sigmoid function, expressed as:

$$g(x) = \frac{1}{1 + \exp(-x)} \quad (1a)$$

The mathematical representation of the Neural Network denotes a non-linear autoregressive relationship that exists between the future value  $y_t$  and past observation ( $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ ) (Khashei and Bijari, 2011). Hence, the Neural Network model in equation (1) can be presented mathematically as:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, \emptyset) + \varepsilon_t \quad (2)$$

Where  $f(\cdot)$  denotes the Neural Network model and  $\emptyset$  is a vector of the parameter in equation (1).

However, for a time series model where external variables are considered besides the internal autocorrelation function, the past observations of the external variables can be included in the model as:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, x_{1t-1}, x_{1t-2}, \dots, x_{1t-p}, \dots, x_{rt-1}, x_{rt-2}, \dots, x_{rt-p}, \emptyset) + \varepsilon_t \quad (3)$$

where  $x_{rt-1}$  denote the observation for the external variable  $r$  collected during period  $t - 1$ . This has been included considering that the predicted outcome  $y_t$  is influenced



by past observations of the same series, as well as the past observations of the external variables. Hence, this implementation will consider the prices of previous trading days as input, as well as observations of the past trading days for the factors that influence the price of white maize for each of the trading strategies.

Using the model represented in equation (3) to make predictions for a period  $t + n$  in the time series, where  $t$  is the current time and  $n$  is a positive integer, there is a need to make provision for the fact that data from the period between time  $t$  and  $t + n$  will not exist. Hence, the Neural Networks model can be built to find the pattern between the independent variables as at the current time  $t$  and the associated past observations for predicting the future time for time  $t + n$ . Therefore, equation (3) can be written as:

$$y_{t+n} = f(y_t, y_{t-1}, \dots, y_{t-p}, x_{1t}, x_{1t-1}, \dots, x_{1t-p}, \dots, x_{rt}, x_{rt-1}, \dots, x_{rt-p}, \emptyset) + \varepsilon_t \quad (4)$$

This model can be used as the foundation for predicting future daily prices while taking into consideration the changes in the market dynamics. It also provides a basis for the retraining of the model to ensure that changes in the market dynamics are captured continuously. Thus, technological advancements that come with a Big Data environment such as in-memory, cloud and parallel computing can be leveraged by developing and retraining different models on how the combination of historical and real-time data influence different periods in the future. Hence, at the close of a business day, different models can be retrained based on the historical patterns that include the day's transactions to determine what will happen in the next 1, 2, 3 days and so forth.

### **5.5.2 Features selection for the model**

Selection of the right input variable that optimally captures and explains the patterns in a time series model is considered as very crucial for the degree of accuracy of the model resulting from a Neural Network (Co and Boosarawongse, 2007; Crone and Kourentzes, 2010; Qi and Zhang, 2008). The extant literature shows that deciding on the input variable for a time series modelling using Neural Networks might be an art as much as a scientific expedition. Several authors conclude that there is no generally

accepted theoretical background to follow in deciding the input variables in a Neural Network based time series modelling (Zou, Xia, Yang and Wang, 2007; Khashei and Bijari, 2011; Jabjone and Wannasang, 2014). However, there are several studies that have made propositions on different approaches for the key subjects that arise in making decisions concerning the input variables of Neural Networks for time series modelling (Crone and Kourentzes, 2010; Bukharov and Bogolyubov, 2015).

Time series modelling using Neural Networks could be a univariate or multivariate analysis just as it is for the statistical modelling approach. A univariate analysis considers only the past observations of the same variable as represented in equation (2). While a multivariate model considers not only the past observations of the variable being modelled but also examines the influence of external variables as denoted in equation (3). Thus, in a multivariate analysis, the choice of external variables that will lead to an optimised model is crucial. Several studies on multivariate time series modelling used the analysis of correlation to support the choice of the external variable (Yu and Ou, 2009; Khamis, Nabilah and Binti, 2014; Jabjone and Wannasang, 2014). It is important to highlight that correlation analysis does not imply that these variables are, of a certainty, responsible for the patterns that exist in the price data (Irwin, Sanders and Merrin, 2009; Bukharov and Bogolyubov, 2015). Other studies simply based their choice of external variables on previous knowledge in their field of study (Wiles and Enke, 2014), while some like Bennett, Stewart and Lu (2014) make use of stepwise regression analysis.

There are, however, there are some other strategies such as spectra analysis etc. However, none of the methodologies is without shortcomings nor are there any of these methodologies that are universally accepted (Crone and Kourentzes, 2010). This is because time series data can have linear, non-linear or a combination of both patterns, but the methodologies are founded either on linear or nonlinear methods. Hence, making a scientific choice about the external variables for Neural Network based time series modelling becomes challenging especially when the series is deemed to have complex patterns. Researchers in the field of Neurocomputing have suggested some approaches, such as the combined filter and wrapper approach by

Crone and Kourentzes (2010) and the use of genetic algorithms by Bukharov and Bogolyubov (2015).

On the other hand, for univariate analysis, as well as for multivariate time series analysis after the external variables have been selected, there is still a need to decide how far back to go in including the effect of past observations in predicting future values. One of the major reasons for using the Neural Networks for time series analysis is to identify and capture nonlinear relationships that might be in the dataset (Qi and Zhang, 2008; Bukharov and Bogolyubov, 2015). However, there are empirical and theoretical evidences that a complex time series data with non-linearity patterns can also possess some linear characteristics (Khashei and Bijari, 2011). Researchers with interest in the application of Neural Networks for modelling time series data have studied the use of statistical approaches such as different aspects of the Box-Jenkins methodology to address linear characteristics of time series data (Zhang, 2003; Khashei and Bijari, 2010). This provides leverage because the combination of different models to form a hybrid has been found to increase the accuracy of predictions from such models (Crone and Kourentzes, 2010).

Qi and Zhang (2008) compared a Neural Network-based time series model and a hybrid of the Neural Networks with different strategies from the statistical time series modelling methodology. The authors found that the resulting network from two different hybrids had better outputs than from using Neural Networks only. Qi and Zhang (2008) further supported the use of techniques such as lagging to include the linear effect of past observations in the Neural Network-based time series data. In determining how far back to go in including the effect of past observations (lag length), Zou et al. (2007) and Khashei and Bijari (2011) carried out several experiments to determine the lag length that produced the best model. There are suggestions that Partial Autocorrelation Function (PACF) which is a part of the Box-Jenkins statistical time series analysis methodology can be used to determine the correct lag lengths in Neural Network based time series modelling (Co and Boosarawongse, 2007; Khashei and Bijari, 2010). But Crone and Kourentzes (2010) warned that PACF could lead to misleading results depending on the characteristics of the data that is being used for

modelling. PACF is used in the Box-Jenkins methodology to give the autocorrelation in a series for each corresponding lagged value (Enders, 2010; Tsay, 2010).

### **5.5.3 Overfitting and Generalisation**

The purpose of identifying the correct parameters and topology for training an optimal Neural Network for different problems is so that the resulting model can adequately be used to estimate future occurrences based on historical data. However, care needs to be taken to ensure that the output from a model is not a result of just memorising the historical data (Provost and Fawcett, 2013a). In such cases, the model will have high accuracy when used to forecast a subset of the data used in training the model, but will not produce a reasonable result when used to predict a dataset not seen by the model during training. Hence the model has not learned the patterns in the data, but it has only memorised the observations in the data. This problem is regarded as the overfitting (O'Neil and Schutt, 2014).

Contrary to just memorising the observations in a training dataset, the desired model from a modelling exercise is one that is able to accurately estimate future outcomes based on input data that has not been seen by the training model at all. This is regarded as generalisation (Provost and Fawcett, 2013b). The level of accuracy of a model can be measured by its ability to generalise even when there is a significant change in the input data. There is a risk of overfitting a model to the training data when the model becomes too complex, such as having too many hidden nodes or too few observations compared to the number of input nodes (Alpaydin, 2010). However, the ability of the model is reduced greatly with an overly simple network (Co and Boosarawongse, 2007). Hence, there is a need to strike a balance between generalisation and overfitting.

A common strategy for avoiding overfitting is to split the available dataset into a training and a test set (O'Neil and Schutt, 2014; Provost and Fawcett, 2013b; Alpaydin, 2010), where the test set is kept completely separate and not used in the training process. The performance of the model is then checked by using the model to forecast the series in the test set and to compare the results with the actual data. Statistical measures such as the Mean Square Error (MSE), Root Mean Square Error (RMSE)

and Mean Absolute Percentage Error (MAPE), provide quantitative measures for comparing the result of prediction from the training set and the test set.

MSE is a modelling evaluation statistic that gives an indication of how much a set of values that has been predicted using a model, varies from the actual observations (O'Neil and Schutt, 2014). It represents the loss function between the result of a trained model when compared with the actual value, hence it is regarded as the training error for Neural Networks (Wilamowski, 2009; O'Neil and Schutt, 2014). The MSE is defined as:

$$MSE = \frac{1}{n} \sum_{t=1}^n (Y_t - F_t)^2$$

Where  $F_t$  represent the predicted values,  $Y_t$  the observed actual values and  $n$  the total number of values. However, a more popular measure of the accuracy of a model is the Root Mean Squared Error that is obtained by taking the square root of MSE (Khashei and Bijari, 2011; Bennett, Stewart and Lu, 2014). RMSE is represented as:

$$RMSE = \sqrt{MSE}$$

The Mean Absolute Percentage Error (MAPE) is another measurement of accuracy of a predictive model which presents the predictions error as a percentage of the actual observed values. It calculates the absolute value of ratio of the error to actual values (Tofallis, 2015), and is calculated as a percentage by multiplying it by 100. MAPE is obtained as:

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{Y_t - F_t}{Y_t} \right|$$

These statistics are generally used for measuring model accuracy in time series forecasting (Enders, 2010; Tsay, 2010) and have also been adopted in measuring the accuracy of Neural Networks-Based time series model as well (Zou et al., 2007; Crone and Kourentzes, 2010; Khashei and Bijari, 2011; Khamis, Nabilah and Binti, 2014).

## 5.6 Time Series Analysis and Neural Networks

Market data from grain commodities and other financial assets trading are time-bound observations which could be dependent on the frequency of trading activities or the structure of the market (Tsay, 2010). Financial assets data are available in different

time intervals such as quarterly, monthly, daily and even in finer intervals such as hourly or even by seconds. Hence, financial market data is collected, stored and analysed as time series data. It was discussed in Chapter 3 that grain commodities prices are considered to be volatile. This could explain the evolution of different models that attempt to understand the patterns that exist in the grain commodities prices (Gutierrez, Olmeo and Piras, 2015).

Box and Jenkins (1970) provide the framework that is popularly adopted as the foundation for different time series analysis and models. The framework is based on the permutation of the degree of the auto-regression (AR) and moving averages (MA) in the data. Auto-regression indicates the relationship that exists in a dataset where the observed value  $x_t$  has a dependence on past values  $x_{t-i}$  where  $t$  signifies a time period and  $i$  is an arbitrary positive integer (Tsay, 2010). On the other hand, the Moving Average (MA) component of a time series provides an indication of a trend that can be identified in the time-series data and the characteristics of the trend where it exists (Tsay, 2010).

AR and MA models have been implemented for understanding the patterns that exist in time series data in several ways such as the Vector Autoregressive (VAR), Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) models (Tsay, 2010; Shumway and Stoffer, 2011). The literature shows that these models have been implemented for solving different time series problems including problems of financial markets such as understanding financial asset prices such as stocks or agricultural commodities (Co and Boosarawongse, 2007; Bennett, Stewart and Lu, 2014; Gutierrez, Olmeo and Piras, 2015).

On the other hand, researchers and practitioners in the Computing fraternity have also been involved in studies that make use of Neural Networks to understand and solve problems relating to time series data with complex patterns. Extant literature indicates that there has been a steady growth in the investigation of Neural Networks for modelling the complex patterns in time series data (Kaastra and Boyd, 1996; Qi and Zhang, 2008; Khashei and Bijari, 2011). Research on the application of Neural

Networks for understanding complex time series data indicates that they are suitable for forecasting future occurrences from patterns that can be found in historical time series data (Khashei and Bijari, 2011).

It has been found that using Neural Networks for modelling and forecasting future time series observations is not limited by the constraint of statistical approaches such as seasonal trend and stationarity (Qi and Zhang, 2008). Moreover, Neural Networks are able to deal with complex patterns and significant changes in patterns that might occur in the time series because of the ability to use non-linear learning to detect changes and relationships that might exist in the data (Zhang, 2003). Neural Networks are also considered to be better than statistical techniques in time series analysis because they are able to analyse and forecast qualitative and discrete data types (Bukharov and Bogolyubov, 2015). Therefore, comparative studies from different areas of application have found Neural Networks to be more efficient than time series analysis that is based on statistical techniques (Co and Boosarawongse, 2007; Zou et al., 2007; Bennett, Stewart and Lu, 2014).

Research into the use of Neural Networks for time series forecasting has been extended to include models that are hybrids of the Neural Networks and statistical approach to modelling time series especially the ARIMA model from the Box and Jenkins framework (Khashei and Bijari, 2011; Bennett, Stewart and Lu, 2014). Theoretical and empirical studies indicate that these hybrids help in determining inputs for Neural Networks that cater for all the complexities in the data and increase accuracy of the forecast (Khashei and Bijari, 2010; Crone and Kourentzes, 2010). The primary goal of combining two or more modelling techniques is to leverage the advantages offered by each of the models that are being combined in order to increase accuracy of the results (Zhang, 2003). Examples of such are hybrids of Neural Network and ARIMA models for time series that deals with the linear, non-linear relationships and complex autocorrelation that might exist in the data.

The capabilities of Neural Networks for time series analysis have made it popular in various fields of research and practice for forecasting the future based on historical data. These include the modelling and prediction of energy demand, sales, currency

exchange rate and the stock market among others. Previous research has also been carried out that indicates that Neural Networks can be used for forecasts relating to the trading of agricultural grain commodities. Co and Boosarawongse (2007) demonstrated the use of Neural Networks for predicting rice exports in Thailand and compared the results with forecasting the same thing using other statistical time series techniques. Zou et al. (2007) investigated Neural Networks for modelling and predicting food prices in China.

More recently, the trading of Soybean Complex was modelled by using Neural Networks by Wiles and Enke (2014) using data from the Chicago Mercantile Exchange and Khamis, Nabilah and Binti (2014) modelled the Wheat price by using secondary data collected on the wheat trade in United States of America. However, these studies have been limited to the identification of the structure of the Neural Networks or comparative investigation of models for making predictions. The studies have generally made use of secondary historical data only. Moreover, the models in these previous studies were only used to make predictions for observations in the past. This leaves an opportunity to explore situations in the future, in a situation where the model require future predictions where external variables are considered, for which, future data will be unavailable.

Although previous studies show the capability of Neural Networks for forecasting commodities prices, no work can be found that includes studies on how Neural Networks can be set up for real-time decision support. This research looks at how data can be acquired in real-time, pre-processed and analysed in real-time in order to make decisions based on the current market trend as much as possible. The next section of this chapter will examine real-time learning for modelling and forecasting time series data such as the grain commodities prices by using factors that affect it.

## **5.7 Real-time Neural Network Learning for Time Series**

The availability of large datasets together with new technologies, tools and the ability to incorporate all these into a real-time solution provides a platform for better support for decision makers (Power, 2014). Having more data available in real-time or near



real-time together with sufficient tools, techniques and technologies that can be used to extract insight from such data in real-time or near real-time could help decision makers to make quicker decisions using more relevant information. The financial markets have been a generator of large datasets for many years through millions of transactions processed daily. The availability, however, of relevant datasets in real-time or near real-time creates new opportunities to analyse financial transactions as they take place, which offers improved decision making (Ruta, 2014).

The ability to collect larger datasets in real-time is of less value except when such data can be used with analytics or for training intelligent systems to extract insight and knowledge in real-time. This will imply that the insight extracted will not be based on historical facts only, but current trends are also taken into consideration as they evolve. Bukharov and Bogolyubov (2015) proposed a Decision Support System that is able to deal with the complexities of using large datasets that are collected in real-time. The authors identified that a DSS that will use large datasets that are collected in real-time to train Expert Systems in real-time should cater for inaccurate and extreme randomness in the data. Thus, the use of Neural Networks or other modelling techniques with Big Data present opportunities to learn from the real-time datasets as well as the historical data. Thereby, new patterns are captured that might be introduced by real-time dataset which could influence future perspectives.

Grain commodities prices have been identified as volatile in Chapter 3. The review of literature and the survey that have been carried out in this study have shown that several factors influence the prices of grain commodities. It was further identified that the degree of influence of each of the factors which influences the market varies with time. Moreover, there are suggestions that the patterns between the grain commodities prices and those of the external factors that influence prices could be linear or non-linear. Thus, the application of a real-time learning strategy together with Neural Networks algorithms that are based on the availability of Big Data, could offer new opportunities for extracting market intelligence that is relevant in time. Using Neural Networks algorithms, predictive models for predicting the future prices and outlook of the different trading strategies for each of the grain commodities could be

developed. Thereby, it would be possible to improve processes to make decisions about training in grain commodities.

## **5.8 Conclusion**

The ability to acquire, integrate and understand the complex patterns and relationships in datasets that exist from several sources can be used to make discoveries, recommendations and even predict future occurrences. This chapter examined the process of acquiring datasets on the prices of grain commodities from several sources and the techniques required to pre-process and integrate the datasets into an integrated data source. A Big Data perspective was considered in approaching data acquisition, management and discovery of insight from the data collected. The data formed the basis for a DSS, which also included modelling, intelligence and visualisation components. Due to the volatility of the grain commodity markets, a real-time or near real-time data analysis of the data collected was suggested for the optimisation of market intelligence and predictive analytics extracted from the DSS.

It was identified in Chapter 5 that Neural Networks are suitable for modelling the relationships in the grain commodities market data and the data of the factors that influence the market. Although, statistical tools such as the ARIMA model, based on the Box and Jenkins framework, was also identified, extant literature confirms that Neural Networks based models perform better than statistical models. This is especially important when dealing with datasets with non-linear relationships, such as the grain commodities market datasets. Backpropagation Neural Networks was identified as a specific Neural Network-architecture suitable for modelling economic time series. Thus, BPNN was recommended for modelling the prices and market for grain commodities. Chapter 5 further examined the background of using a BPNN model for predictions. This includes the requirements, set up and topology of the model such as the input, hidden and output layers.

Chapter 5 addressed the second research object of this study (RO<sub>2</sub>) which to identify modelling techniques for predicting the future prices of grain commodities in South Africa. As part of that objective, Chapter 5 provided an answer to the fourth research

question of this study (RQ<sub>4</sub>) – What are the modelling techniques utilised for discovering patterns and making predictions from datasets? Therefore, Chapter 5 formed part of the systematic design of the framework that can be followed to develop a DSS to support grain commodities trading decision making. While this is a part of the design/building of artefact in the DSR methodology, it also contributes to the rigour cycle. It should be noted, that the need for further rigour in this chapter was identified during the evaluation of the proposed artefact which is described in Chapter 7. Hence, an iteration between the rigour cycle in this chapter and the design cycle, specifically, during the evaluation of the artefact took place.

Grain commodities markets data and the datasets on the factors that influence the prices are time bound. Although these datasets are from different sources and have different structures, data pre-processing techniques can be used to integrate the datasets into an integrated time series dataset. However, there is a need to consider the volume, velocity and variety of datasets that will be handled, especially because the data will be collected and analysed in real-time. Hence, choosing the right technologies, tools, techniques and the general perspective need must be different from the traditional ways of Data Mining.

It was identified in Chapter 5 that the Data Science process can be adopted for real-time data acquisition, extraction of market intelligence and predictive analytics from the acquired datasets. This provides a platform for making discoveries, providing recommendations and extracting future outlook of grain commodities such as predicting grain commodities future prices. Predicting the future grain commodities prices for the different trading strategies that are available could assist stakeholders in the industry in making better decisions. Particularly, it could be of more benefit to grain commodities farmers with fewer skills and resources to make the right decision that will help them in effectively managing their exposure to price-related risks. Moreover, the availability of such insights for decision making to farmers could improve their positions from being the price takers in the industry.

In the next chapter, findings from Chapters 3, 4, and 5 will be integrated to develop a proposed framework for developing a DSS for supporting trading decisions concerning

grain commodities. Chapter 6 will provide a conceptual framework for a grain commodities trading DSS. The chapter will also describe an implementation of the conceptual framework. Furthermore, the use of different modelling approaches will be compared and the outcome will be presented.

# Chapter 6 : Proposed Grain Commodities Trading DSS Framework and Implementation

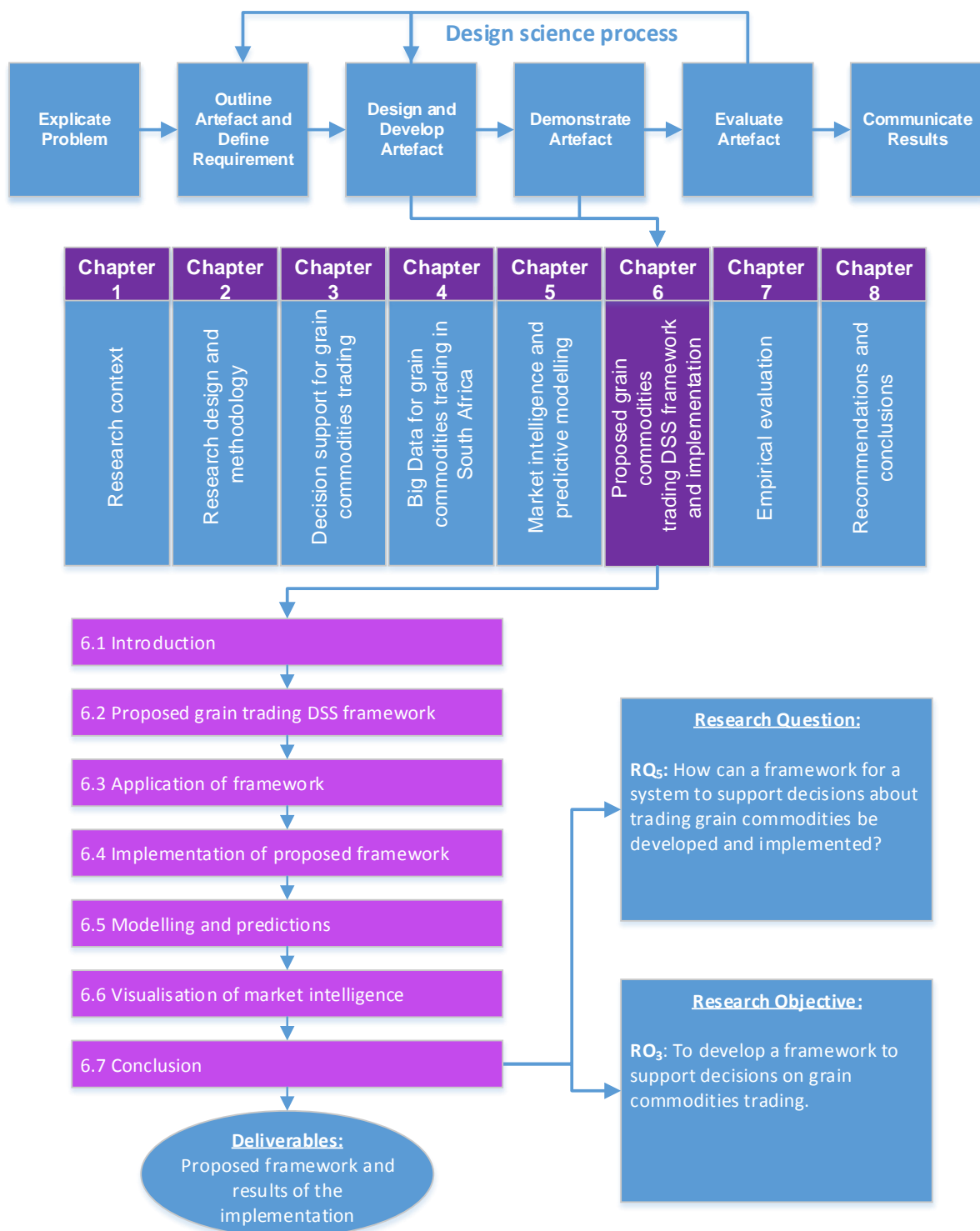


Figure 6.1: Chapter outline and deliverables

## 6.1 Introduction

The review of literature in Chapters 3, 4 and 5 identified the requirements, materials and methods that are required to support decisions for grain commodities trading. The discussions provided a systematic development of the components and important considerations for a framework which can be followed to implement a framework for a system to support making decisions for grain commodities trading. It was identified in Chapter 3 that a DSS framework should combine the data, modelling, intelligence and visualisation components in order to have a computing-based solution that can improve decision making. It was found that the availability of the predicted future prices of grain commodities for each of the different trading strategies would enhance the decision making about trading in grain commodities.

Chapter 3 also identified the variables that influence grain commodities trading in South Africa and Chapter 4 provided a list of datasets for those variables. By taking advantage of the Big Data approach, Chapter 4 discussed how the identified datasets can be acquired and the issues to consider when using such datasets. Chapter 5 described the techniques that can be used for the modelling of the patterns that exist between grain commodities prices and the identified external factors using the spot and the December futures contract prices of white maize as a case study. Therefore, the findings in Chapters 3, 4 and 5 form the basis for a grain support framework for commodities trading decisions.

A framework can be either theoretical or conceptual. The focus of a theoretical framework is in using existing theories and proven abstractions to understand a problem or concept. However, a conceptual framework provides a structural approach for organising ideas into a logical proposition for providing solutions (Shields and Rangarajan, 2013). By its definition, a conceptual framework should be the result of a research rigour and can be used to achieve the objectives of a study such as finding answers to a research problem by synthesising a collection of ideas into a logical sequence (Ravitch and Riggan, 2012). The ideas that were examined in the previous chapters are from different disciplines and this chapter will attempt to present a sequential flow on how these ideas fit together for developing a DSS for grain

commodities trading. This chapter will propose a conceptual framework of trading decisions in grain commodities support.

Based on the findings in Chapters 3, 4 and 5, this chapter will propose a decision support framework that can be implemented to provide a support system for decision making in trading grain commodities. The proposed framework will identify the important components for a DSS in trading grain commodities and how the components fit together. The main objective of this research RO<sub>m</sub> is to design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities.

This chapter will focus on the research objective RO<sub>3</sub> - To develop a framework to support decisions on grain commodities trading. Chapter 6 will also seek to provide answers to the research question RQ<sub>5</sub> - How can a framework for a system to support decisions about trading grain commodities be developed and implemented? As part of the overall objective of this study and to fulfil the set objective for this chapter, an attempt will be made to combine findings from the previous chapters on how relevant data can be collected, integrated into a single and useable source of data, analysed and presented for decision making into a conceptual framework.

As highlighted in Chapter 2 the building of the actual artefact that provides a solution to the identified problem is known as the design/development phase of the DSR process. Within the DSR cycles described by Hevner (2007), the design/development phase is at the centre of the design cycle. The design cycle feeds iteratively from the rigour cycle as the scientific basis for building the artefact. However the building of the artefact could also be iterated internally with the evaluation phase within the design cycle in order to refine and improve the artefact (Hevner, 2007; Vaishnavi and Kuechler, 2015).

In the context of this study, the previous three chapters identified the building blocks which will be joined into the envisaged final artefact of this study. This is a systematic design of the expected artefact. Chapter 6 will present a conceptual framework as the resulting artefact of the research rigour in Chapters 3, 4 and 5. It is expected that the

presentation of the artefact will demonstrate a logical flow of the identified components and propositions. An implementation of the proposed framework will be carried out in this chapter, which will involve iterative experimentation to identify the right input data and architecture of the Neural Network model that will be implemented. It is required that the choices that will be made are grounded scientifically, therefore the iterative tasks during the implementation tasks will link back to the rigour cycle in Chapters 4 and 5. The experimental implementation of this framework will be carried out by applying a framework to develop a DSS which provides a forecast of white maize prices by predicting future prices of white maize for different trading options over the same period. This application of the artefact to address a selected case problem will provide a demonstration of the artefact which is one of the DSR activities (Johannesson and Perjons, 2012).

The next segment of this chapter, Section 6.2, will present and discuss the proposed conceptual framework for grain trading DSS. It establishes a link between the problem, the identified components, propositions and the proposed framework. This is followed by a discussion of the applications of the framework in Section 6.3. The documentation of an experimental implementation and the results of the proposed framework will be provided in Section 6.4. The implementation of the modelling component of the framework will be provided in Section 6.5 while visualisation and market intelligence will be discussed in Section 6.6. The last section in the chapter (Section 6.7) will provide concluding remarks on the proposed framework and the implementation that will be carried out in this chapter.

## **6.2 Proposed Grain Trading DSS Framework**

The need for future price prediction was identified in Chapter 3 as an important factor for decision making about grain commodities trading; this can be achieved if the price of the commodities can be forecast. It was identified that stakeholders in the grain commodities trading industry, especially the farmers, need to compare the future outlook of different trading strategies. The ability to know and compare what the price of a grain commodity will be for each of the available trading strategies will enable sellers, such as the farmers, to decide which strategy to adopt and when it will be more



profitable to sell their produce. Conversely, other stakeholders such as millers, who purchase and add value to grain commodities, will be able to hedge effectively and also set the right price for their products.

It has also been mentioned in Chapter 3 that prices of grain commodities are volatile, perhaps signifying complex patterns, and that several factors influence the price at different times. Some of the factors that influence the prices of grain commodities in South Africa were identified by using the price of white maize as a case study. The review of literature in Chapter 4 pointed out that data for the factors that influence prices of grain commodities can be acquired – some in real-time and others near real-time into an integrated data repository by using the Big Data approach, tools and techniques. Methods for understanding and making forecasts from complex, time series data can then be used to add value to such data for price discovery and decision making.

It was established in Section 3.3.1 that an ideal Decision Support System requires a data component, a modelling component, a knowledge component and the user-interface component. Therefore, a framework for a Decision Support System that supports trading in grain commodities is proposed as illustrated in Figure 6.2. The proposed framework identifies real-time acquisition and integration of datasets from different sources as the data component, together with the use of statistical or computation-intelligence technique for modelling. Key intelligence is categorised as predictions, discoveries and recommendations, while the visualisation component provides user experience. The proposed framework was adapted to include provision for suitable technologies.

### **6.2.1 Real-time data acquisition and integration layer**

The proposed framework suggests that an economic investigation of the factors driving the volatility of prices of grain commodities should form the foundation of the DSS for trading grain commodities. This investigation requires adequate knowledge of the domain to ensure that the process of data gathering has a scientific background. It was noted in Chapter 4 that the datasets for the variables that influence the grain commodities market in South Africa are available from different sources. Some of the

datasets are made available in real-time, others in near real-time, while there are others that are made available at longer frequencies such as daily and monthly.

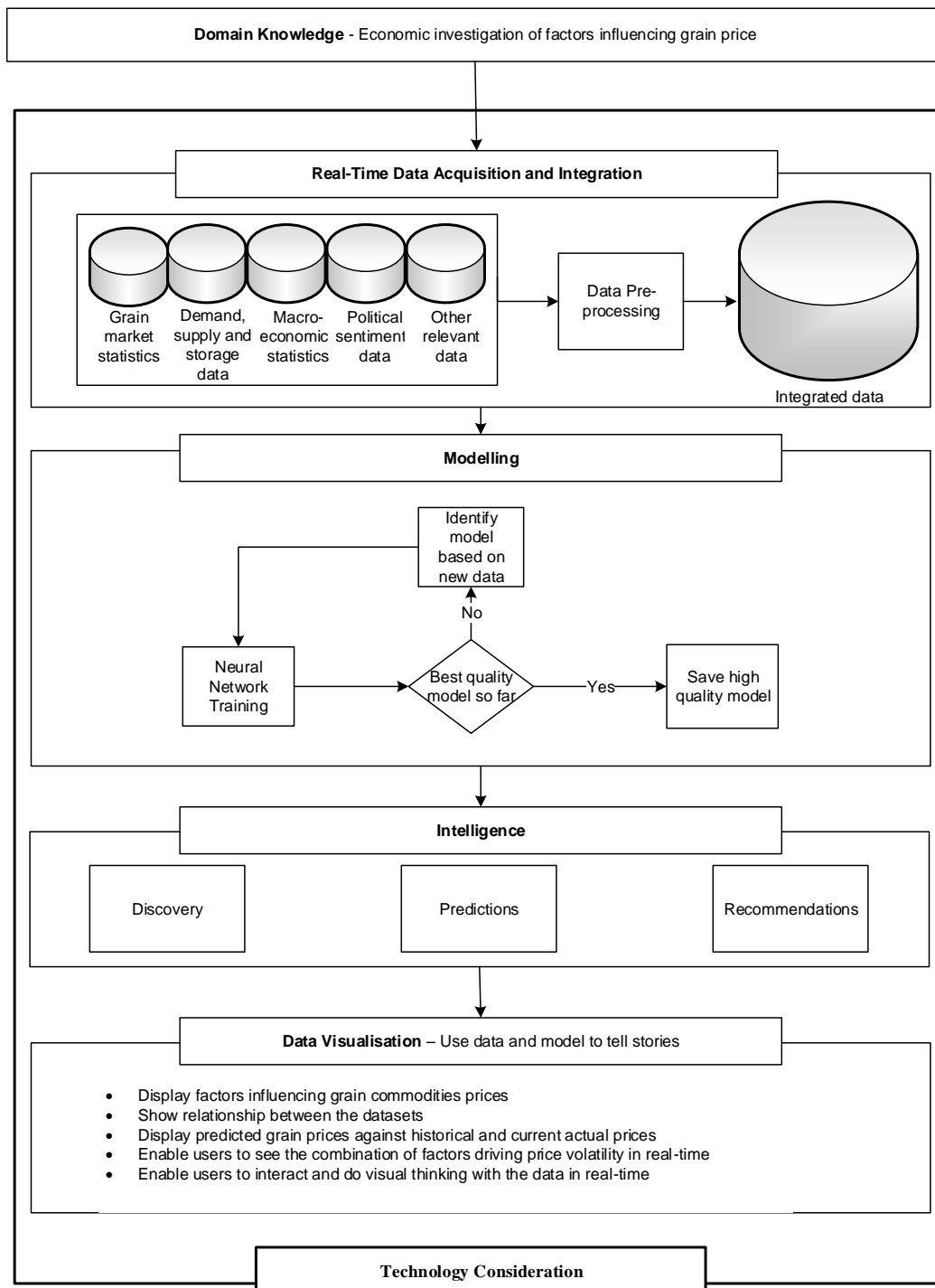


Figure 6.2: Proposed grain commodities trading DSS

A Big Data approach, tools and techniques, were proposed for acquiring datasets. This will ensure that challenges associated with collecting and using data as they are generated; such as inconsistencies, volume and different structures, are dealt with. Therefore, the proposed framework suggests the collection of historical data and the setting of a data-streaming mechanism to collect, pre-process and integrate data in real-time. The choice of a Big Data approach in the acquisition of the data will also make it possible to use the collected data for using the data in real-time despite the associated complexities.

### **6.2.2 Modelling layer**

A modelling layer was included in the proposed framework to understand the patterns and relationships that exist in the data collected. The proposed framework suggests that Neural Networks can be used to model the patterns that exist in the data. Based on the real-time learning strategy described in Section 5.7, it is proposed that the Neural Networks be re-trained periodically as new data becomes available. The application of this strategy will depend largely on the use of technologies that enable Big Data such as parallel and in-memory computing in order to cope with the demand for resources. A separate computing thread can be used for training and testing models and when a model that is proven to be better than the previous model is identified, it can be moved to production for making predictions. By so doing, the DSS will be able to capture the patterns in the market by taking historical data and the current market trend into consideration.

### **6.2.3 Intelligence layer**

The resulting models from the modelling layer can then be used to develop knowledge and information that can be used to support decision making. This layer will extract information, make recommendations and make discoveries based, for example, on predicting the grain commodities prices or making recommendations based on market trends. More importantly, it is expected that this layer of the proposed framework will be able to extract relevant information in real-time due to the availability of real-time data and a modelling technique that is continuously updated. Besides the use of a model to create intelligence, this layer can also be used to extract useful information

directly from the integrated data by showing the relationships that exist among the different variables.

#### **6.2.4 Visualisation layer**

Data-driven solutions are only valuable when they support business or research goals. Irrespective of the volume of data amassed and the complexity of the analysis carried out with the data, the value derived from the solution depends on the experience of the users. It was explained by Minelli, Chambers and Dhiraj (2013) that data visualisation should be used to bridge the gap between analytics and the consumption of analytics. Big Data and Data Science-related efforts are likely to involve large or complex datasets; therefore, the way the results from such projects are communicated will justify such investments. Thus, it is important that the information, insights and knowledge that emanate from analytic efforts be communicated innovatively.

Innovative visualisation of data has advanced recently as an off-shoot of the developments in the Big Data and Data Science concepts. Data visualisation enables users to observe changes in patterns that exist in large datasets over time (Minelli, Chambers and Dhiraj, 2013). Moreover, users are able to interact directly with data to discover and explore relationships between data, thereby generating insights that can be actioned. Developments in data visualisation, especially with Big Data and within the ambit of Data Science, can be described as using data to “tell stories”. It has been noted that new data visualisation should enable users to answer questions like who, what, where, when, why and how when they interact with data (Segel and Heer, 2010).

An appropriate visualisation technique will enable the presentation of the results from the previously mentioned three layers in the proposed framework. The visualisation layer should make the relationship that exists in the data obvious to the users. Also, the framework suggests the need to display the predicted prices and provide a visual indication of the factors influencing the market each time. It should also enable the visual communication of other discoveries and insights from the data in a simple and easy-to-understand way so that stakeholders with minimal skills can make use of it. To cater for mobility and real-time decision making as there are different categories of

users, the visualisation layer should make provision for both web access and access from mobile devices.

### **6.2.5 Technological considerations**

The complexities associated with combining data acquisition, modelling and visualisation require that adequate consideration be given to the choice of technology platforms. The proposed framework includes a technology-consideration component on which data acquisition, modelling and data visualisation depend. It is suggested that the right technology should be carefully selected to support the Big Data and Data Science approach of the framework. A selected technology ecosystem should enable the application of concepts such as distributed computing, parallel computing, in-memory computing and cloud computing. It is also proposed that the technology ecosystem of choice should have modelling techniques as native tools to ensure that the chosen technology has been designed for the desired purpose.

## **6.3 Application of Framework**

The proposed framework can be implemented as a Decision Support System for grain trading in South Africa. The system can be implemented to predict the price of a certain grain commodity for different grain trading strategies so that the decision makers can decide what will be the most appropriate marketing strategy to adopt in trading their commodities. Insights from a DSS, based on the proposed framework can be used to make decisions on whether the speculative trader should maintain or exit a market position. In the same light, organisations that use grain commodities as raw materials can also use the insight from the resulting DSS for supporting their hedging and pricing decisions. Besides, the proposed framework can also be implemented as part of a bigger solution such as algorithmic trading for grain commodities trading. Also, the implementation of the framework can be used for further research that studies the changes in the factors that influence prices of grain commodities. The integrated repository of data also presents an opportunity for more discoveries that can benefit different sectors of the industry.

## 6.4 Implementation of Proposed Framework

The description of the proposed framework for trading in grain commodities in Section 6.2 indicates the design of the artefact for this study as prescribed by the DSR process (Johannesson and Perjons, 2012). The design of the proposed framework is founded on the review of literature on an existing inter-disciplinary knowledge base, tools and technologies. This ensured that the proposed framework is well grounded in ideas, artefacts, methods and theoretical frameworks from previous research described in the rigour cycle of the DSR methodology (Hevner, 2007). Although the artefact is well grounded scientifically, there is also the need to ensure that the artefact is evaluated for its ability to solve real-life problems by using scientific methods (Hevner, 2007). Hence, there is a need for an implementation and evaluation that show how the designed artefact will solve a relevant problem, such as the use of an experiment or case study (Johannesson and Perjons, 2012). These activities could provide a feedback for improvement of the artefact as it is described in the design cycle (Hevner, 2007).

This sub-section describes the demonstration of the proposed framework as a support for decision making for trading white maize (WMAZ) in South Africa. Maize contributes more than 40% of the gross value of field crops and it is considered the most important in the South African grain commodities market (DAFF, 2014). WMAZ can be traded using the spot, futures, forward or options strategies as described in Section 3.4.1. The stakeholders, such as the farmers, always want to know the right strategy to adopt by comparing the expected yield from all the strategies. Moreover, when using the spot strategy, they want to know if the price of the commodity will go up in the near future so that they can hold on longer and decide when to quickly sell their commodities if there is a forecast decline in price of the commodity. Therefore, a system that predicts the future outlook for each of the available trading strategies could assist in making better decisions. This implementation will focus on predicting the spot price and futures contract prices of WMAZ.

White maize is traded on the JSE commodities derivatives markets daily between 9:00 am and 12 noon. During the trading hours, there are thousands of transactions that go

through per minute with a possibility of significant change in prices during the trading hours. On the other hand, there is also an end-of-day price that is captured as the price of white maize for the day. The price data can be collected in real-time as different transactions are recorded, which can be thousands of transactions per second or as end-of-day data which is a single observation per day. The implementation of the proposed framework in this study will make use of the end-of-day data. This is because of the limitation of resources to acquire and manage live data for WMAZ prices and detailed data for the factors influencing the WMAZ prices.

#### **6.4.1 Technology consideration**

Section 4.2.2 described the Apache's Hadoop framework and SAP HANA as some of the technologies that have been developed to support working with Big Data projects. SAP HANA was adopted as the technology of choice to demonstrate the proposed framework because of the availability of a fully functional version through Amazon Web Services for cloud computing. Moreover, the setting up of the SAP HANA on the cloud platform is fairly easy to deploy, making it possible to focus on the core tasks of implementing the proposed framework.

Besides, SAP HANA is specially designed to effectively handle data streaming over the cloud. It provides adapters for different types of data stream for connecting the data sources to the SAP HANA data streaming server which continuously pulls data from the sources based on a pre-defined event. SAP HANA data-streaming services allows for the embedding of business logics in the data streaming. This can be used for pre-processing tasks on the data stream before it is committed to the database (SAP, 2015). This is particularly important because the datasets required for the proposed framework are available on several websites and web-based platforms. Hence, the need to automate the streaming of the data as it is updated from sources without human intervention.

SAP HANA is deployed with in-memory computing and predictive analytics libraries both for Statistical and Neural Networks libraries for predictive analysis. Provision is made within the SAP HANA framework to integrate the R programming platform, which is an open source tool that makes several tools for statistical and computing-based

analysis of data. SAP HANA also provides matured visualisation tools for both web and mobile access.

The other technology tool adopted for the purpose of implementing the proposed framework is the R programming language. R was adopted because it provides scripting functionalities to easily acquire and clean data from different sources. Moreover, R provides a statistical programming capabilities required to execute exploratory data analysis required for the implementation of the proposed framework. Besides, The R platform can also be integrated with other systems such as SAP HANA in the case of this implementation.

#### **6.4.2 Data acquisition**

##### **Market data**

All trading of grain commodities in South Africa is conducted through the South African Futures Exchange (SAFEX), a subsidiary of the Johannesburg Stock Exchange (JSE). Therefore SAFEX is the custodian of trade data on grain commodities in South Africa. In terms of the South African law and as it obtains in other nations, the JSE acts as a regulator of the markets and it is expected to discharge this responsibility with transparency; hence the JSE makes available market statistics of all agricultural commodities among other data (JSE, 2015). Historical data is made available on the website of the JSE and real-time data is made available as a service to registered clients.

For the purpose of this implementation, historical data on grain commodities spot and futures transactions was obtained from the website of JSE with permission to use the data for research purposes (permission attached as Appendix E). End-of-day data for spot prices was captured directly from the newsfeed provided on the website of JSE while end-of-day data was captured from the website of a major grain commodities-storage company ([www.senwes.co.za](http://www.senwes.co.za)). Although the data from JSE was presented in Microsoft Excel files, the data was largely semi-structured because the structure of the files had changed over time.



Historical futures contract transactions in grain commodities are available as daily transaction files from 02-01-2009, making the file to be a total of 1,649 Microsoft Excel file as at 31-07-2012. Each file contains data on the type of contract, such as WMAZ 12-2015 which indicated a white maize futures contract for December 2015. The files also contain variables such as the volume of transactions per day and the minimum and maximum for a particular contract for the day. Included data also provides the number of open interests and market-to-market price which indicates the closing price for the day. Typically, each file contains end-of-day transactions for all the grain commodities traded as futures on the exchange such as white maize, yellow maize, wheat sunflower and so on. A print out of a sample file is attached as Appendix F. The files were downloaded and consolidated into a single file using scripts written in the R programming language before importing them into SAP HANA as a table.

Similarly, the historical data for the grain commodities spot transactions are made available as a single file that is updated periodically on the JSE website. The spot transaction file contains end-of-day prices of white maize, yellow maize, wheat, sunflower and soybeans, linked to a single date column from 02-01-2007. As with the futures transactions, R programming language scripts were used to download the file. A print out of a sample file is attached as Appendix G. Scripts were also written to restructure the data as time series data for analysis, the data were then imported into SAP HANA as a database table.

To include the effect of the international grain commodities market, this implementation will include the effect of the grain commodities market in the USA as a major producer of white maize. The CBOT provides different market statistics such as end-of-day, historical and live-feed data with different levels of details depending on needs, for a fee (CMEGroup, 2015). However, the market statistics from CBOT are also available through different brokers that provide the data alongside market data from several other sources. One such data broker is EODDATA.COM that provides market statistics from 30 different exchanges around the world. A subscription service for historical and end-of-day transaction data of agricultural commodities transactions was purchased from EODDATA. This made available historical transactions from 02-01-1995, in total 5,325 Microsoft Excel files, fortunately, in a structured format that

made it easy to be consolidated and uploaded as a database table in SAP HANA. Each of the CBOT transaction files contained columns of symbols representing different transaction types, dates of transaction, opening, highest, lowest and closing prices. The files also contained the volume of trade per symbol and the number of open interest, which represents the total number of yet-to-be-fulfilled futures contract on the exchange. A print out of a sample file is attached as Appendix H.

### **Demand and supply**

The South African Grain Information Services (SAGIS) is a constituted non-profit organisation which has the responsibility of collecting, analysing and providing data that relates to the economics of grain commodities in South Africa (SAGIS, 2015). The stakeholders in the industry are statutorily mandated to supply SAGIS with data regarding the supply, demand, import and export of grain commodities in South Africa. This data is aggregated by SAGIS and made available periodically to the stakeholders and the public through the SAGIS website – [www.sagis.org.za](http://www.sagis.org.za).

Demand and supply are present on the SAGIS site as Microsoft Excel files that are updated monthly. The file contains the national opening stock, total acquisition – from farms and stock imported, local consumption – human and industrial, exports and closing stock data for each month. A print out of a sample file is attached as Appendix I. Each grain commodity has its own separate file for demand and supply data from May till April of the following year. The data from each of the files was extracted and transformed into time series data with dates and with the code for each grain commodity as a unique identifier and then was uploaded as a SAP HANA table.

Similarly, the Economic Research Services (ERS) of the United States Department of Agriculture (USDA) provides detailed demand, supply and storage data on agricultural commodities in the USA through a web-based portal service. The portal also provides data on the global flow of grain commodities globally which indicates countries that are major producers and those that are major consumers for each of the grain commodities. The ERS provides the data to support stakeholders in their decision-making processes regarding agricultural economic and policy issues (USDA, 2015) which will include the grain commodities marketing decision making.

The data on the USDA portal is updated regularly and access to the data is guided by the Freedom of Information Act of the United States (USDA, 2015). The portal can be accessed by the public via the web at <http://www.ers.usda.gov/data-products/feed-grains-database/feed-grains-custom-query.aspx>. The portal allows the viewing and downloading of the available data in several formats such as Comma-separated value (CSV), Portable document format (PDF), HyperText markup language (HTML) and Microsoft Excel files. Demand and supply data for maize in the USA is made available as quarterly data by USDA. A print out of a sample file of the USA grain commodities demand and supply data is attached as Appendix J. The data was transformed into a monthly, time series data, with the assumption that the demand for the period given is the same figure using script writing in R programming language.

There are suggestions from the literature that weather patterns are considered as a variable that influences the demand and supply of grain commodities. Weather data for the main areas where white maize is planted in South Africa was sought from the South African Weather Service. The organisation made available daily weather data collected for Bethal, Bethlehem, Bloemfontein, Bloemhof, Bothaville, Ficksburg, Kroonstad, Potchefstroom, Secunda, Van Reenen and Vryburg. The datasets made available include daily rainfall, maximum and minimum temperature and wind speed. The data was transformed into a time series in a format that is suitable for the proposed analysis and was uploaded into SAP HANA as a data table.

### **Macroeconomics**

The review of literature in Chapters 3 and 4 suggests that Macroeconomic factors such as exchange rates, interest rate and crude oil price may have a direct influence on the prices of grain commodities. In the experimental implementation of the proposed framework, data for the Rand versus US Dollars exchange rates, prime interest rates, bank repo rates and spot price of Brent crude oil will be considered as macroeconomic data. The historical and updated data on the exchange rate of the Rand to all the major currencies of the world and different interest rates is frequently updated and made available through the research pages of the SARB website – [www.resbank.co.za](http://www.resbank.co.za). On the other hand, data for the spot prices of Brent crude oil is available as open data at

[www.quandl.com](http://www.quandl.com). These organisations make historical data for these macroeconomic factors available dating back to the 1970s. Data for the different variables is downloaded, transformed and imported as an SAP HANA database table with the data and a unique identifier for each factor as keys.

### **6.4.3 Integration of datasets**

All the datasets collected from the different sources of data described in the previous section were transformed and uploaded into database tables as time series data. The grain commodities market data in South Africa from JSE, together with the data market data from CBOT (USA) are end-of-day data. These match the macroeconomics data and the weather data. However, the demand and supply data that could be accessed for this implementation is monthly data for local demand and supply, whereas the demand and supply data for the American market is in a quarterly format. For this data to be in this implementation, it was assumed that the monthly observation provided for the month/quarter is the same for each of the trading days during the period for which both the local and USA demand and supply data was made available. Hence, R programming language scripts were written to disaggregate the observation of each month or quarter over the number of trading days in the period.

Figures 6.3 and 6.4 show a schematic representation of the resulting tables for the spot prices of grain commodities and futures prices of grain commodities respectively, together with the tables of the factors that influence the prices. The schema in Figures 6.3 and 6.4 can be used as the foundation for streaming data from the sources mentioned in the previous section with the initial pre-processing of the data embedded in the data-streaming project that will reside on a separate SAP HANA server. Alternatively, an R server that runs R programming language scripts can also be set up separately from the database server to run the pre-processing script before committing data to the database server. The historical data was originally set up until 31-06-2015 and each of the data tables was updated as new observations became available.

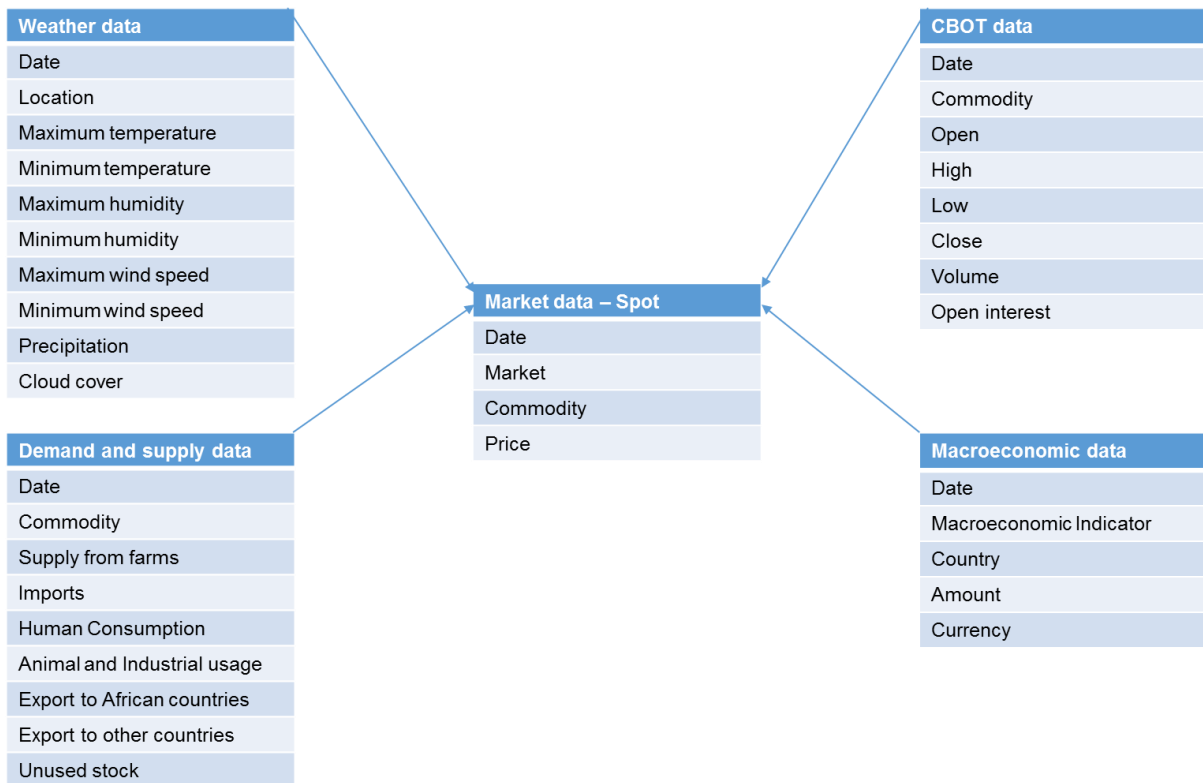


Figure 6.3: Database schema for spot price modelling

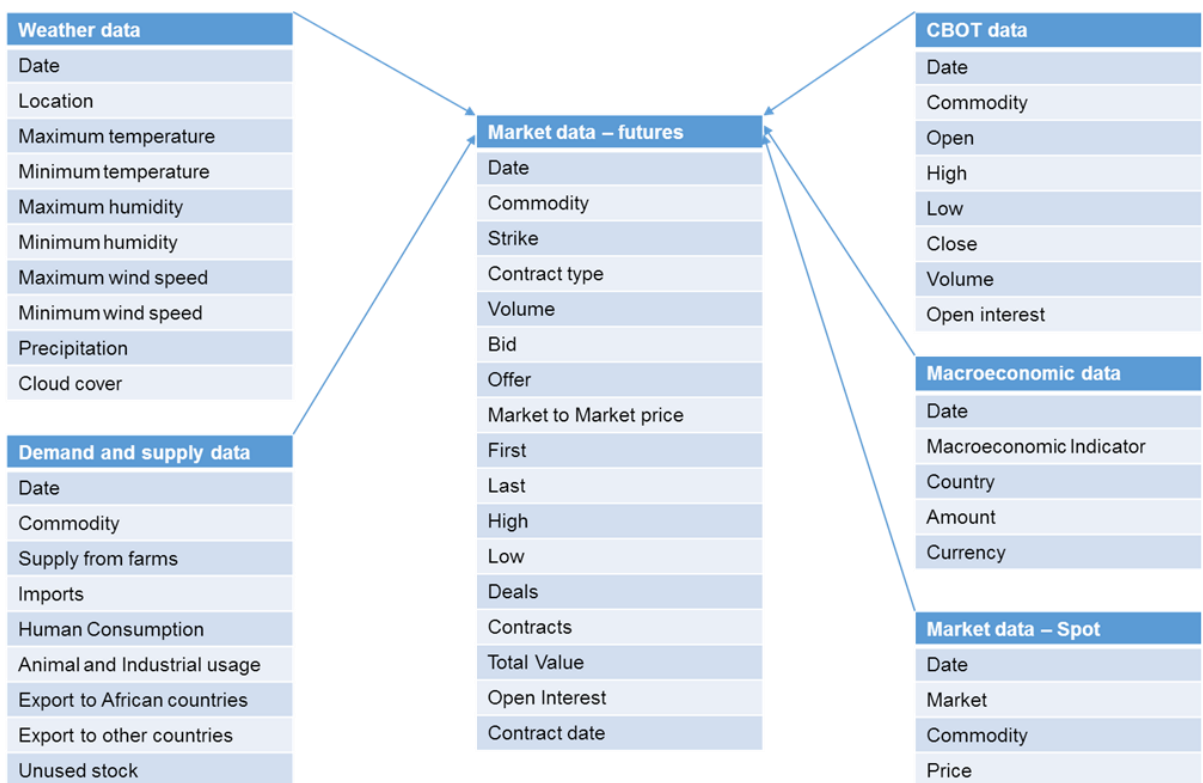


Figure 6.4: Database schema for futures price modelling

#### **6.4.4 Exploratory analysis**

Exploratory Data Analysis (EDA) provides an insight into the characteristics of a dataset. It also gives an indication of the relationship or patterns that might exist in the dataset, thereby providing a basis for setting a hypothesis on which further analysis can be carried out to either confirm or refute. Before the application of modelling techniques to understand the patterns that exist in the datasets that have been collected, an exploration of the data was carried out using the historical data from 01-02-2007 to 29-05-2015. The purpose of this analysis is to make sense of the data by looking at relevant descriptive statistics for each of the datasets. Moreover, visual tools such as graphs will also be used to examine the patterns that might be in the datasets. These visual techniques will also be used to examine the relationships that might exist between the prices of white maize for the two trading strategies that are the focus of this implementation and each of the factors that is presumed to influence its volatility.

##### **6.4.4.1 Spot price of white maize**

Historical data of the end-of-day spot price of white maize was taken from 02-01-2007 till 31-07-2015 resulting in a total and complete 2,149 observations. Table 6.1 compares the summary statistics of the price of white maize in South Africa over the 1, 3, 6, 12, 36 and 60 months with the entire historical data. The summary shows a volatile movement in the mean and minimum price over the different periods while the maximum price of white maize did show a significant volatility within the different periods. However, there is a significant difference between the minimum and the maximum price over each of the periods examined, suggesting that the price of white maize is indeed volatile.

It can be seen on the graph presented in Figure 6.5 that the price of white maize in South Africa can move in different directions over different time periods with no specific pattern that is visible to the eye. The graph shows that the price of white maize has significantly increased from about 2011. Figure 6.5 also suggests that the price of white maize in South Africa might have been at its highest point in 2014 although this might have been a spike because the price also fell drastically within a few months. Figure 6.6 presents a comparative visualisation of the price of white maize in the last 5 years, 3 years, 1 year and 3 months.

Table 6.1: Descriptive statistics for spot prices of white maize over different periods

Duration	Mean Price (R)	Minimum Price (R)	Maximum Price (R)
All	1,954.00	1,019.00	3,765.00
Last 5 years	2,195.00	1,109.00	3,765.00
Last 3 years	2,388.00	1,611.00	3,765.00
Last 12 months	2,354.00	1,629.00	3,338.00
Last 6 months	2,784.00	2,180.00	3,338.00
Last 3 Months	2,956.00	2,570.00	3,338.00
Last 1 month	3,162.00	3,099.00	3,338.00

A line plot of the prices as shown in Figure 6.5 indicates a stochastic time series with evident movement in the price of the commodity daily.

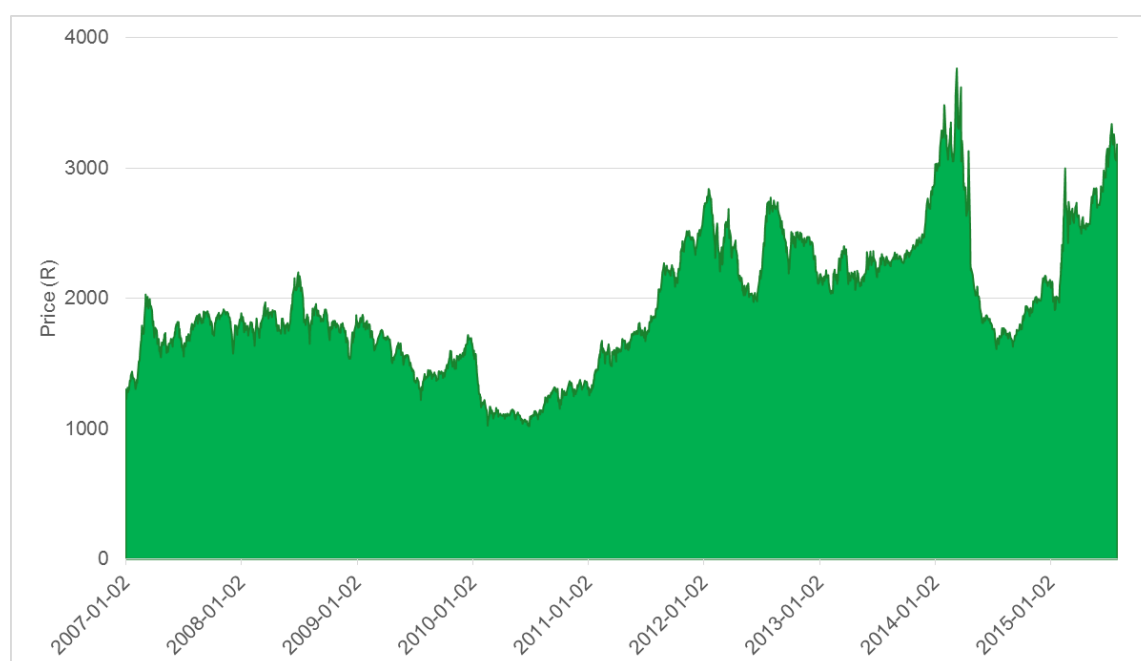


Figure 6.5: Graph showing the spot price of white maize for all historical data

The graphs in Figure 6.6 emphasise the complexities of the price of white maize as a traded commodity. It can be seen clearly that there is a continuous change in the price from year to year, month to month and even on a daily basis. This might be responsible for the difficulty of farmers and other stakeholders in the industry to effectively determine the future outlook of the commodity. On the other hand, the graphs in

Figures 6.5 and 6.6 indicate that the data is void of any anomaly and suitable for use in understanding the volatility in the price of white maize in South Africa.



Figure 6.6: Graphs showing the spot prices of white maize over different periods

#### 6.4.4.2 Relationship between white maize and wheat prices

Products that are alternatives can influence the prices of each other. White maize and wheat are considered to be alternative products, especially for human consumption. Hence, there are tendencies that the changes in the price of one might affect the other. For example, the shortage of one of the commodities can create an increase in demand for the other, while an over-supply of one might bring down the price of the other. Figure 6.7 presents a diagrammatic illustration of a suspected relationship that might exist between the commodities. Although the price ranges for the two commodities are different, the line graphs in Figure 6.7 indicate a subtle pattern in the movements of the prices of both commodities over the year. A further probe indicates that there is a 0.6380 ( $n=2,149$ ;  $p<0.0001$ ) correlation between the prices of the



commodities, indicating a fairly strong and obvious relationship between the two variables.

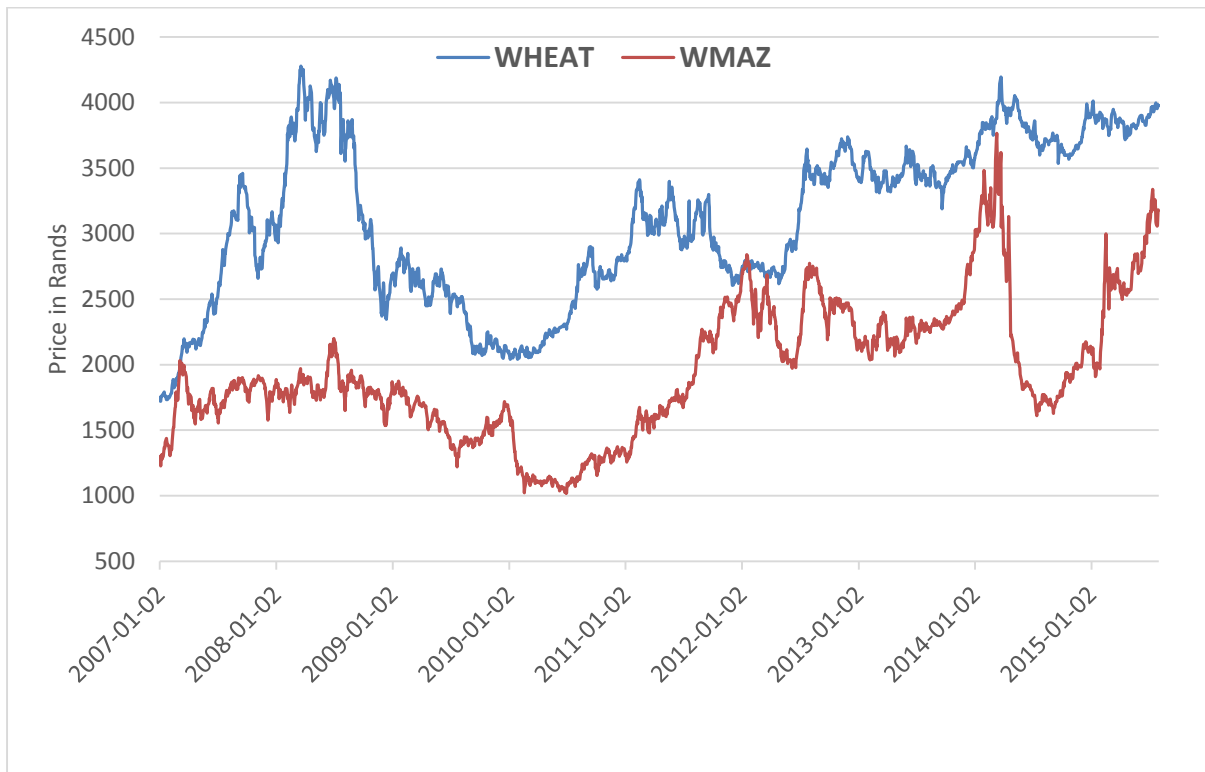


Figure 6.7: Graphs showing the spot prices of white maize and wheat

#### 6.4.4.3 Relationship between spot price of white maize and macroeconomics variables

Visual exploration of the relationship between the prices of white maize and macroeconomic factors is complicated by the difference in the scale of the variables. However, the use of a secondary scale for some of the variables provides a visual indication of the relationship that might exist between the variables, although it is understood this is not a perfect representation of such relationships. Figure 6.8 presents the relationship between the spot price of white maize and the exchange rate between US Dollars and Figure 6.9 presents the spot price of white maize against the spot price of Brent crude oil. A correlation of 0.5885 ( $n=2,149$ ;  $p<0.0001$ ) was found between the spot price of white maize and the US Dollar–Rand exchange rate. However, the spot price of white maize in South Africa and the spot price of Brent crude oil exhibited a 0.3191 ( $n=2,149$ ;  $p<0.0001$ ) correlation. These results suggest relationships that should be investigated.

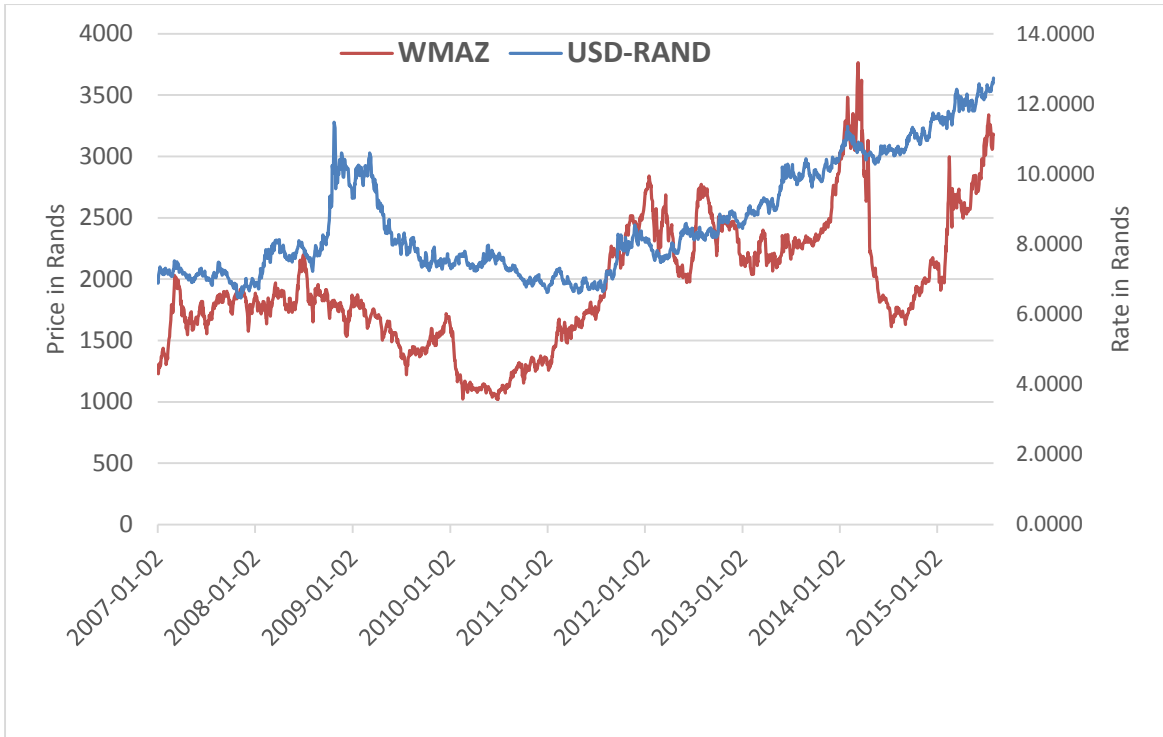


Figure 6.8: Graphs showing the spot prices of white maize and USD-Rand Exchange rates

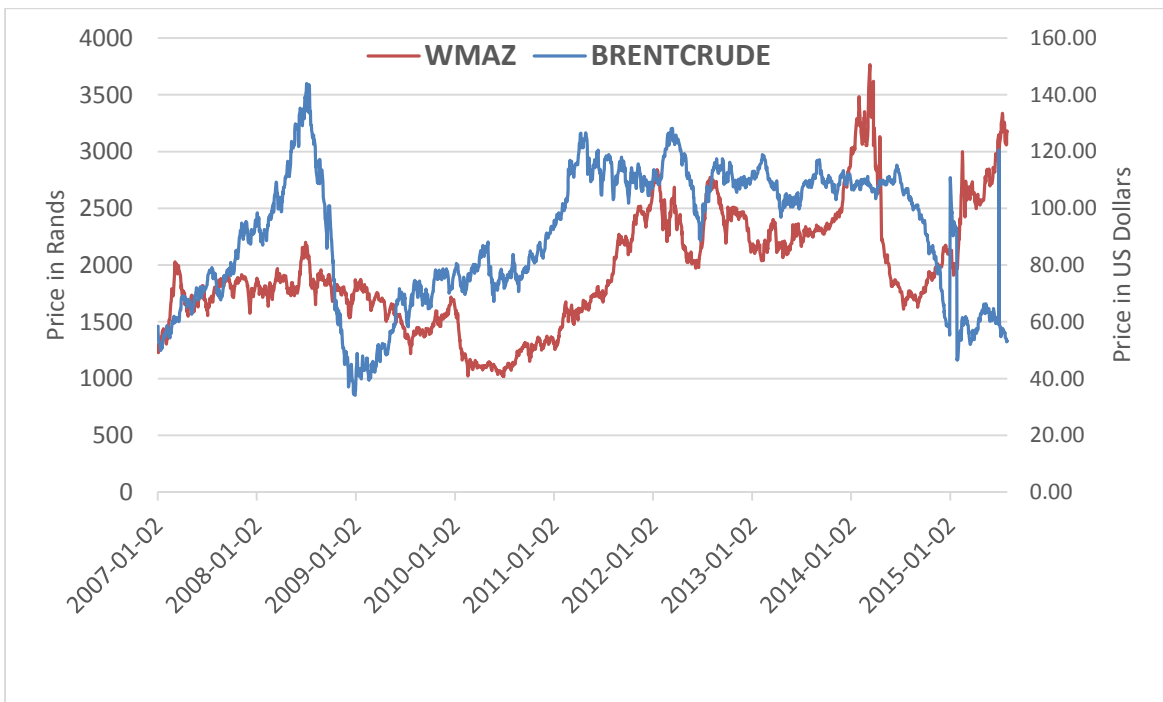


Figure 6.9: Graphs showing the spot prices of white maize and Brent crude oil  
Other macroeconomic factors considered for this implementation are the prime and the bank repo interest rates. The patterns in Figure 6.10 suggest that both interest

rates may be completely dependent on one another; however the relationship between both rates and the spot price of white maize is elusive. Both rates have a weak  $-0.3428$  ( $n=2,149$ ;  $p<0.0001$ ) correlation with the spot price of white maize. This suggests that there is no need to use both rates for understanding the price of white maize. Hence, the rest of this implementation will only make use of the prime interest rate which is the rate at which banks lend or make overdrafts available to their customers. Farmers or other stakeholders that borrow money from banks for their operations may be affected by changes in the prime rates which could be passed on to the selling price of the commodity and perhaps the price that some stakeholders are willing to pay.

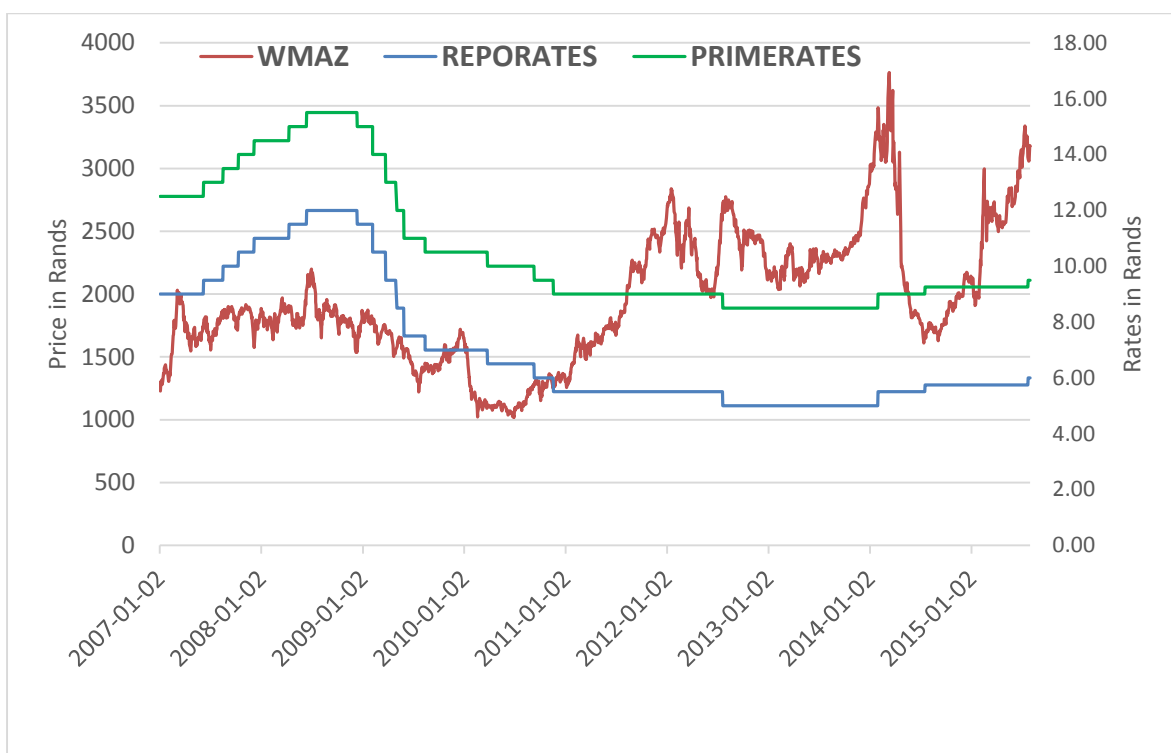


Figure 6.10: Graphs showing the spot prices of white maize and interest rates

#### 6.4.4.4 Influence of the USA market on the spot price of white maize in SA

A correlation analysis between the spot price of white maize in South Africa and corn in the USA shows a weak correlation of  $0.2860$  ( $n=2,149$ ;  $p<0.0001$ ). However, this indicates that there is a relationship between the two commodities that is worthy of further investigation. The volume of trade of corn in the USA can also be considered as an indication of global demand and supply of the commodity. Results indicate a  $0.2848$  ( $n=2,149$ ;  $p<0.0001$ ) correlation between the volume of trade of corn in the

USA and the spot price of white maize in South Africa. Figures 6.11 and 6.12 present a visual representation of the relationship that might exist between the spot price of white maize in South Africa and the price and volume of corn trade in the United States of America respectively.

The results of this exploratory analysis further suggest that there may be influences from outside the country that explain the volatility of the prices of white maize in South Africa and the inclusion of such data in understanding the patterns that exist in the market could be beneficial. Thus, this implementation will include the price of corn and the volumes of daily transactions in the investigation of the patterns in the spot price of white maize in South Africa.

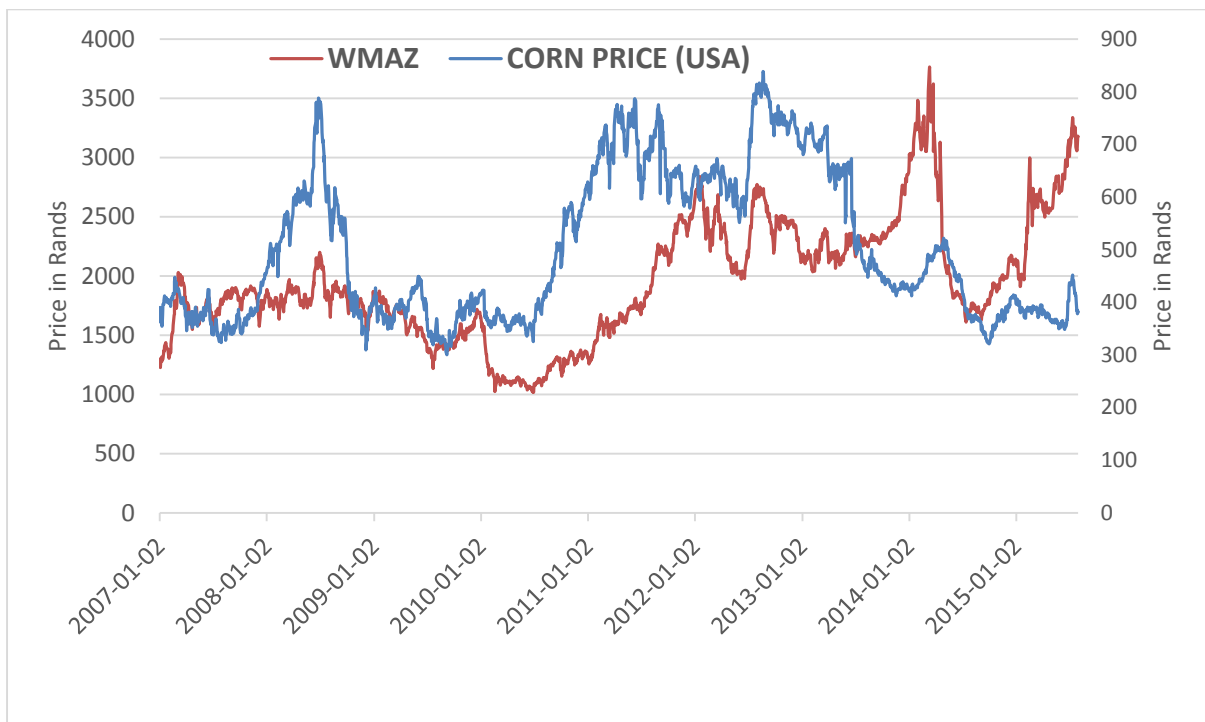


Figure 6.11: Graphs showing the spot prices of white maize in South Africa and price of corn in USA

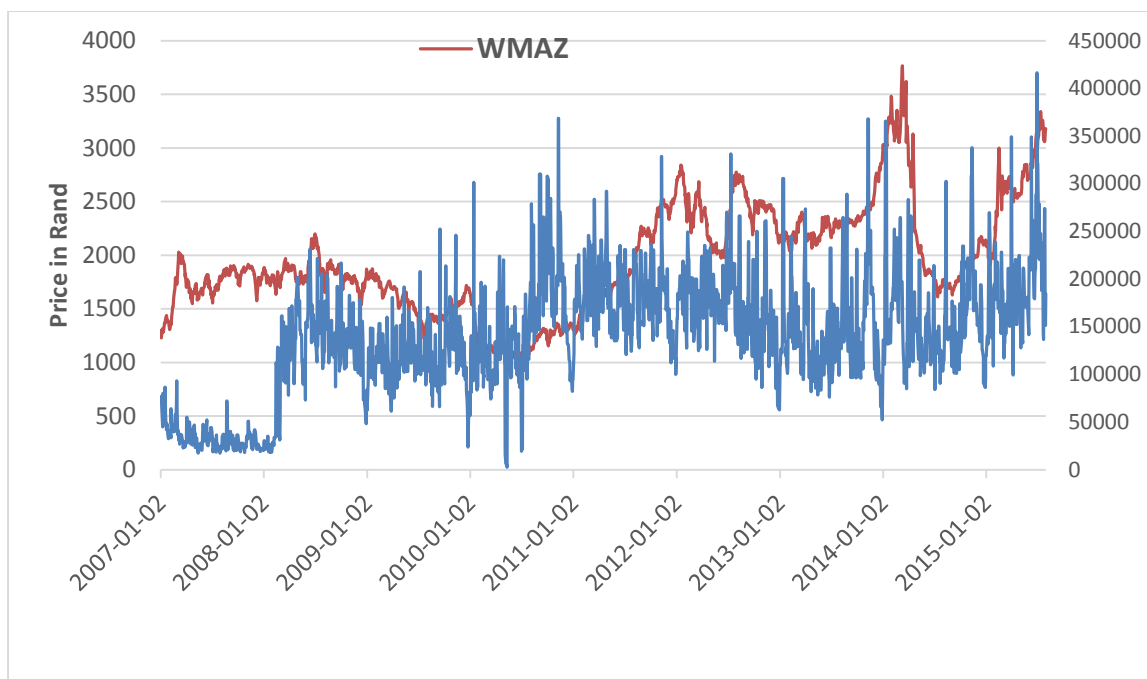


Figure 6.12: Graphs showing the spot prices of white maize and volume of corn trade on Chicago Board of Trade (USA)

#### 6.4.4.5 Relationship between the spot price of white maize and local demand and supply

Theoretically, demand and supply are expected to have a direct and probably strong relationship with prices. However, the non-availability of matching, daily demand and supply data makes understanding the relationship difficult. While daily and even more granular data is available for grain commodities prices, only monthly data is available. In using this data, the demand and supply of commodities for each day of the month is assumed to be the figure available for each month. Therefore, the monthly figures were disaggregated for each of the trading days of each month. The correlation analysis between the spot price of white maize and the disaggregated demand of white maize in South Africa showed a 0.2474 ( $n=2,149$ ;  $p<0.0001$ ) correlation indicating again a very weak and negative  $-0.0057$  ( $n=2,149$ ;  $p<0.0001$ ) supply of white maize in South Africa. Figures 6.13 and 6.14 provide a visual illustration of the relationship between spot price, demand and supply of white maize in South Africa. Based on the correlation analysis and visual exploration, the supply-related variables seemed irrelevant in understanding the patterns in the spot price of white maize. Hence, the variables will not be excluded from the rest of this study.

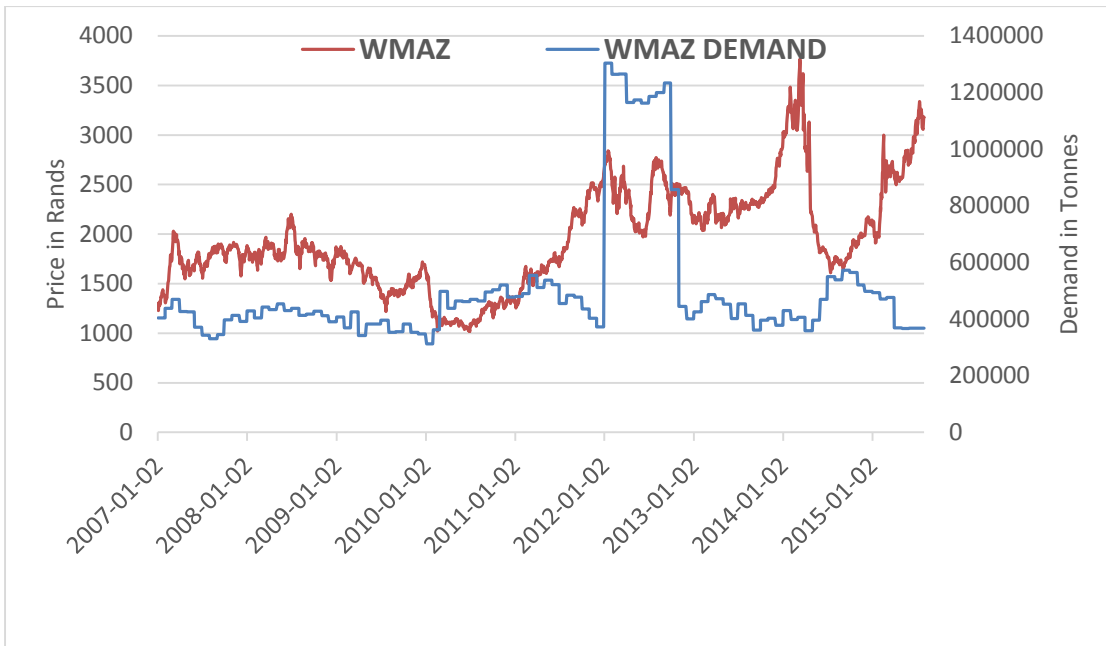


Figure 6.13: Graphs showing the spot prices and demand of white maize

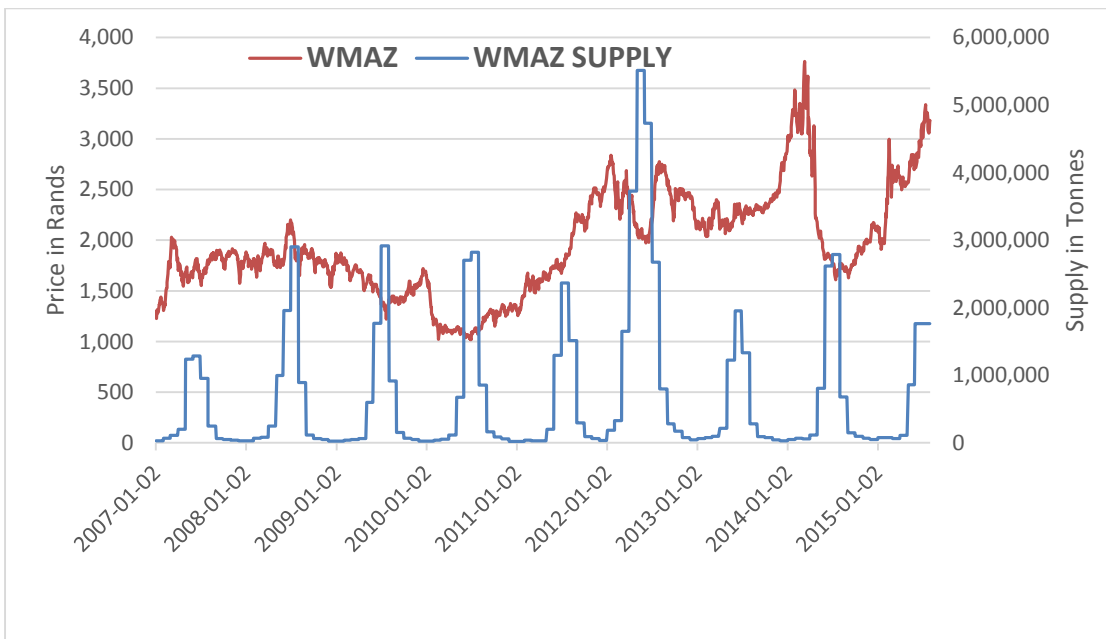


Figure 6.14: Graphs showing the spot prices and supply of white maize

Besides the factors that have been explored earlier in this section, data on other variables that might influence the prices of white maize in South Africa was also collected. The exploratory analysis in this segment provides an indication of which variable to include in the modelling of the prices of white maize as was been seen in

the previous exploratory analysis in this segment. Table 6.2 presents the results of the other exploratory analysis that was carried out on the other variables.

Table 6.2: Correlation between spot price of white maize and other variables

No	Variables	Correlation with spot price of WMAZ
1	Closing Stock of WMAZ	-0.0194 (n=2,149; p<0.0001)
2	Closing Stock of Wheat	-0.0060 (n=2,149; p<0.0001)
3	Supply of Wheat	0.0312 (n=2,149; p<0.0001)
4	Demand for Wheat	-0.3347 (n=2,149; p<0.0001)
5	Average rainfall	-0.1021 (n=2,149; p<0.0001)

Table 6.2 shows the possible relationship that may exist between the spot price of white maize, and other demand and supply-related factors. The results show a rather weak relationship between the spot price of white maize in South Africa and the stockpile of white maize and that of wheat in South Africa. The correlation analysis also suggests a weak 0.0312 (n=2,149; p<0.0001) correlation between the spot price of white maize and the supply of wheat. However, the results show that the demand for wheat in South Africa could have a negative influence on the spot price of white maize with a -0.3347 (n=2,149; p<0.0001) correlation. Daily rainfall in major producing areas of white maize in South Africa was aggregated. The resulting data was to represent average daily rainfall, which resulted in a 0.1021 (n=2,149; p<0.0001) correlation with the spot price of white maize in South Africa. The correlation analyses in sub-section 6.4.4.2 to sub-section 6.4.4.5 that examined the relationship between the spot prices white maize and the factors identified were all found to be statistically significant at 0.05 level of significance, with p<0001 in each case.

#### 6.4.4.6 Futures contract prices of white maize

Four major futures contracts of white maize are available on the JSE based on the expiration date. These are the March, May, September and December futures contracts for every year. This implementation attempts to use the December futures to understand the patterns that might exist in the prices of white maize futures. From the daily transactions, historical December futures contract prices of white maize were extracted from 02-01-2009 till 31-07-2015. When combined with external data where

complete data is available, it resulted in a total of 1,694 observations. An exploratory analysis of the December contracts suggests that the futures contracts might be just as volatile as the spot price. With a very strong 0.9006 ( $n=1,694$ ;  $p<0.0001$ ) correlation between the spot price and the December futures transactions for each day, there are suggestions that the futures prices will respond to the factors already explored in the previous segments.

The graph in Figure 6.15 provides a graphical view of the volatile nature of the future contract prices, while Figure 6.16 gives an indication of the very close relationship that exists between the spot prices and the December futures contract prices of white maize. Therefore, it can be assumed that the prices of the futures contract of white maize will likely respond to the factors influencing the spot prices in the same way.

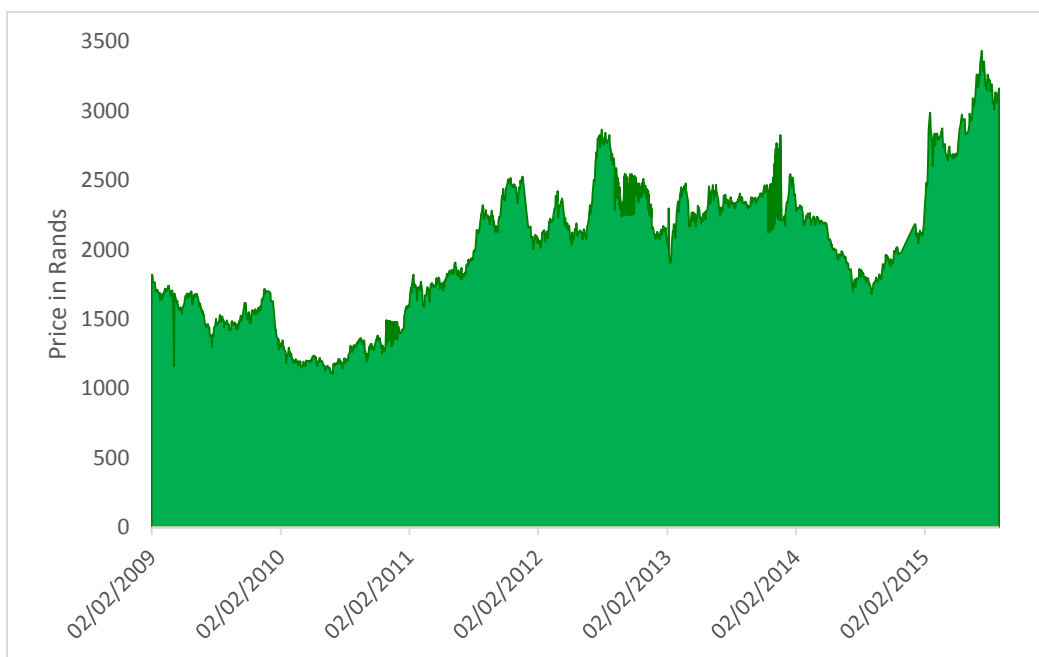


Figure 6.15: Graphs showing the closing price of December futures contract of white maize



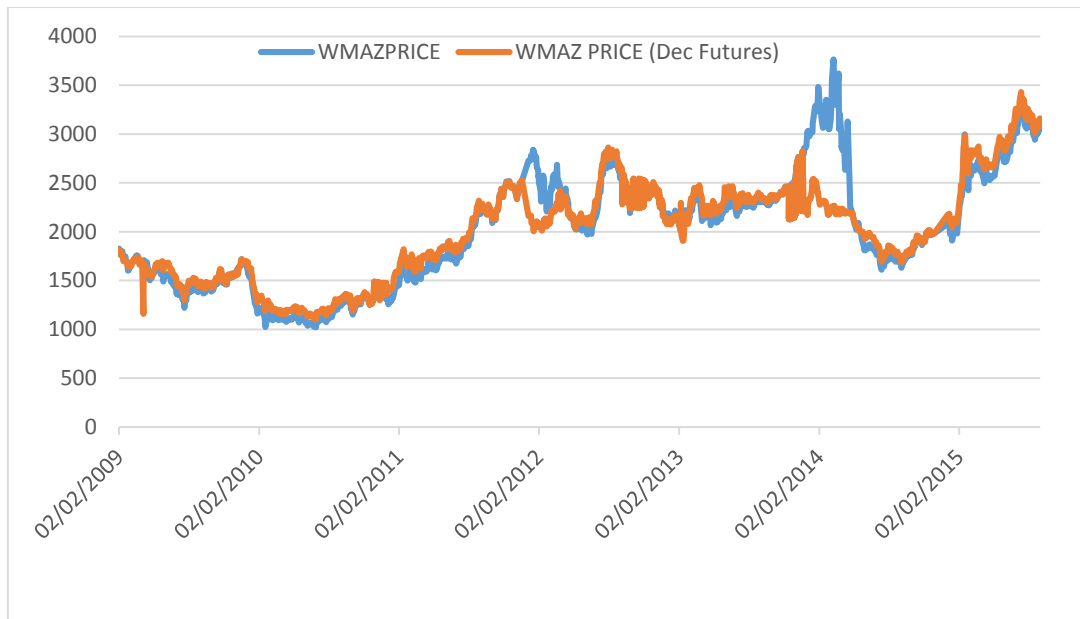


Figure 6.16: Graph showing the spot price against closing price of December futures contract of white maize

Table 6.3 shows the correlation analysis between the prices of December futures contract of white maize and the factors identified as influencing its volatility. The result of the correlation analysis presented in Table 6.3 suggests that the relationship between the identified variable and the price of December futures contracts of white maize is about the same as that of the spot price. The result showed a 0.7436 (n=1,694;  $p < 0.0001$ ) correlation between the price of wheat and the futures contract of white maize while the price of corn in the USA also showed a fairly significant 32.20% relationship. Brent crude oil, US Dollar-Rand exchange rate and the prime interest rate exhibit a 0.2492 (n=1,694), 0.6314 (n=1,694;  $p < 0.0001$ ) and -0.4865 (n=1,694;  $p < 0.0001$ ) correlation with the December futures respectively. Also, the demand for white maize and the demand for wheat also suggest a 0.1986 (n=1,694;  $p < 0.0001$ ) and -0.3938 (n=1,694;  $p < 0.0001$ ) correlation with the price of December futures contract of white maize respectively. But the supply of both white maize and wheat, as well as the stockpile figures of both commodities, seems to have a negligible influence on the price of December futures contract of white maize. Each of the correlation analyses that was carried out between the prices of December futures contract of white maize and the factors identified were found to be statistically significant at 0.05 level of significance, with  $p < 0.0001$  for each of the analysis.

Table 6.3: Correlation between price of December futures contract of white maize and other variables

No	Variables	Correlation with December futures of WMAZ
1	Spot price of WMAZ	0.9192 (n=1,694; p<0.0001)
2	Price of Wheat	0.7436 (n=1,694; p<0.0001)
3	Closing price of Corn on CBOT	0.3220 (n=1,694; p<0.0001)
4	Spot price of Brent Crude Oil	0.2492 (n=1,694; p<0.0001)
5	USD-Rand Exchange Rate	0.6314 (n=1,694; p<0.0001)
6	Prime interest rate	-0.4865 (n=1,694; p<0.0001)
7	Closing stock of WMAZ	0.0871 (n=1,694; p<0.0001)
8	Demand for WMAZ	0.1986 (n=1,694; p<0.0001)
9	Supply of WMAZ	0.0381 (n=1,694; p<0.0001)
10	Closing stock of Wheat	-0.1376 (n=1,694; p<0.0001)
11	Supply of Wheat	0.0476 (n=1,694; p<0.0001)
12	Demand for Wheat	-0.3938 (n=1,694; p<0.0001)
13	BID price of December futures	0.7494 (n=1,694; p<0.0001)
14	OFFER prices of December futures	0.7260 (n=1,694; p<0.0001)

The analysis that was carried out for an exploration analysis of collected data indicates that the historical data that has been collected may be clean enough for modelling and forecasting the spot and futures contract prices of white maize. It further shows the possibility of relationships that may exist between the variables, signifying which variable should be included and which variables are likely to be discarded.

The implementation of the proposed framework of this study was limited to the prediction of daily prices of white maize for the spot and futures trading strategies. Therefore, only the end-of-day price data of white maize in South Africa for the two trading strategies and the corresponding data of factors that influence the prices were required. The implementation did not focus on more granular transactions such as hourly, half-hour, quarter-hour, minute by minute or every second transaction, but the datasets collected for the implementation exhibited some characteristics of Big Data.

The historical market data for the futures contract of grain commodities was a total of 1,649 Microsoft Excel files. This data was unstructured to a large extent because many of the files had inconsistent and different formats that made it very difficult to integrate the files. As mentioned in Section 6.4.2, the corn data from the Chicago Board of Trade exchange also came as a total of 5,325 Microsoft files, although all the files had a consistent structure. On the other hand, the weather data was also received as a total of 685 Microsoft Excel files which were also summarised and integrated. Moreover, this implementation also took advantage of the open data available on different websites that are driven by the Big Data and Open Data concepts.

Besides the unstructured nature of some of the datasets, the variety of the data collected from different sources also identifies with Big Data characteristics. After the data was loaded in SAP HANA, the historical CBOT data had 2,687,971 records while the data table for the futures data had 1,025,774. The table with the data on spot prices of grain commodities had only 18,847 records while there were also 30,195 records for the Demand and Supply data and the collected weather data had 64,284 records. Moreover, Figures 6.3 and 6.4 that provide a schematic flow of the data tables show that several variables were collected for each of the factors out of which the relevant data is being identified. The realities of this implementation highlights one of the challenges of Big Data to sift through large datasets to bring out data that is of high value. The proposed framework can be implemented to provide more detailed support by providing market intelligence, predictions and other insights every hour, minute or seconds. This will imply that much more data in terms of volume, velocity and variety will be involved for such implementation.

After the data collected from different sources had been integrated, the dataset for spot prices of white maize had 50 variables while 68 variables were identified for the futures contract of white maize in South Africa as shown in Figures 6.3 and 6.4 respectively. However, after the exploratory analysis, it became obvious that several of these variables may not be relevant for the modelling and predictions of the spot prices and futures contract prices of maize in South Africa.

## 6.5 Modelling and Predictions

The grain commodities trading DSS framework that has been proposed in this chapter suggests that the use of statistical time series methods or a Neural Network-approach can be used to model the patterns that exist in the grain commodities prices as mentioned in Chapter 5. Thereafter, the models can be optimised based on the changes in the market which will be captured and implemented as soon as they occur by making use of real-time or near real-time data for retraining the models and improving the accuracy of future price predictions. This section will implement Backpropagation Neural Networks for modelling and predicting spot and futures contract prices of white maize in South Africa. The December futures contract of white maize will be used to simulate the futures contract prices.

The review of literature in Chapters 3 and 4 of this study indicates that there has been extensive research into the factors that influence the price of grain commodities globally and also in South Africa. Hence, the choice of input variables for the Neural Network modelling component of the implementation of the proposed framework will be based on the combination of the literature review and the correlation analysis. For both of the trading strategies that are being investigated, the choice of variables that will be used for modelling will be based on a combination of two factors for the purpose of this study. These include the perceived relevance of the variable based on the literature study in Chapters 3 and 4 and the variables that exhibit a correlation higher than 0.2 with dependent variables. Table 6.4 provides the input variables considered for inclusion as input variables for the spot price of white maize while Table 6.5 shows the factors selected as input variables for the modelling of the December futures contract of white maize in South Africa.

Table 6.4: Input variables for Neural Network model for WMAZ spot price

No	Variables	Correlation with spot price of WMAZ
1	Spot price of WMAZ (lagged)	
2	Spot price of Wheat	0.6280 (n=2,149; p<0.0001)
3	USD-Rand exchange rate	0.5885 (n=2,149; p<0.0001)
4	Spot price of Brent Crude oil	0.3191 (n=2,149; p<0.0001)
5	Prime interest rate in SA	-0.3428 (n=2,149; p<0.0001)
6	Price of Corn in USA	0.2860 (n=2,149; p<0.0001)
7	Volume of Corn Trade in USA	0.2848 (n=2,149; p<0.0001)
8	Demand for WMAZ in SA	0.2474 (n=2,149; p<0.0001)
9	Demand for Wheat in SA	0.3347 (n=2,149; p<0.0001)

Table 6.5: Input variables for Neural Network model for WMAZ December futures contract price

No	Variables	Correlation with December futures of WMAZ
1	December futures prices (lagged)	
2	Spot price of WMAZ	0.9192 (n=1,694; p<0.0001)
3	Price of Wheat	0.7436 (n=1,694; p<0.0001)
4	Closing price of Corn on CBOT	0.3220 (n=1,694; p<0.0001)
5	Volume of Corn Trade in USA	0.2848 (n=1,694; p<0.0001)
6	Spot price of Brent Crude Oil	0.2492 (n=1,694; p<0.0001)
7	USD-Rand Exchange Rate	0.6314 (n=1,694; p<0.0001)
8	Prime interest rate	-0.4865 (n=1,694; p<0.0001)
9	Demand for WMAZ	0.1986 (n=1,694; p<0.0001)
10	Demand for Wheat	-0.3938 (n=1,694; p<0.0001)
11	BID price of December futures	0.7494 (n=1,694; p<0.0001)
12	OFFER prices of December futures	0.7260 (n=1,694; p<0.0001)

The accuracy of a Neural Networks model for a time series depends greatly on the ability to identify the lag lengths for the input variables and this requires interdisciplinary skills (Khashei and Bijari, 2011). Choosing the appropriate lags has a direct influence on the performance of the resulting model because it determines the level of the time series components, such as seasonality and trends, that are built into

the model (Crone and Kourentzes, 2010). There is, however, no generally acceptable theoretical basis for choosing the right lag length for time series modelling using Neural Networks, (Zou et al., 2007; Khamis, Nabilah and Binti, 2014). As a result, in the implementation of the proposed framework of this study, experiments are carried out with different lags to identify the optimum lag for modelling the white maize price for both the spot and December futures contract prices as presented later in this chapter.

### 6.5.1 Neural Networks modelling experiments

SAP HANA provides an in-memory computing-based database storage and a predictive analytics library with several modelling algorithms. The predictive analytics library of SAP HANA is designed to enable easy scripting of predictive modelling algorithms as database system procedures using SQL scripting language (SAP, 2015). This makes it easy and flexible to implement dynamic predictive models. The combination of the data streaming, storage, pre-processing, modelling and visualisation techniques built on in-memory architecture makes SAP HANA a candidate for the implementation of the DSS.

Table 6.6: Mandatory parameters for setting BPNN topology in SAP HANA (SAP, 2015)

Name	Data Type	Description
<b>HIDDEN_LAYER_ACTIVE_FUNC</b>	Integer	Active function code for the hidden layer.
<b>OUTPUT_LAYER_ACTIVE_FUNC</b>	Integer	Active function code for the output layer.
<b>LEARNING_RATE</b>	Double	Specifies the learning rate.
<b>MOMENTUM_FACTOR</b>	Double	Specifies the momentum factor.
<b>HIDDEN_LAYER_SIZE</b>	Varchar	Specifies the size of each hidden layer.

Table 6.7: Optional parameters for setting BPNN topology in SAP HANA (SAP, 2015)

Name	Data Type	Default Value	Description	Dependency
<b>MAX ITERATION</b>	Integer	100	Maximum iterations.	
<b>FUNCTIONALITY</b>	Integer	0	Specifies the prediction type: <ul style="list-style-type: none"> <li>• 0: Classification</li> <li>• 1: Regression</li> </ul>	
<b>TARGET COLUMN NUMBER</b>	Integer	1	Specifies the number of target value columns for regression.	Ignored when FUNCTIONALITY is 0.
<b>TRAINING STYLE</b>	Integer	1	Specifies the training style: <ul style="list-style-type: none"> <li>• 0: Batch</li> <li>• 1: Stochastic</li> </ul>	
<b>NORMALIZATION</b>	Integer	0	Specifies the normalization type: <ul style="list-style-type: none"> <li>• 0: None</li> <li>• 1: Z-transform</li> <li>• 2: Scalar</li> </ul>	

The predictive analytics library of SAP HANA provides a function for implementing the Backpropagation Neural Networks (BPNN) that has been selected for implementing the Neural Network modelling of white maize prices as proposed in this study. The BPNN algorithm in SAP HANA is divided into the training and prediction functions. The training function enables the user to set parameters that determine the topology of the network that will be created by using a data table. Tables 6.6 and 6.7 show a list of mandatory and optional parameters that can be set respectively to determine the topology of the network. Both tables show that a network can be trained and optimised by setting the right parameter.

In order to train a Neural Network in SAP HANA, it is required that the input data should also be created as a database table. Thereafter, the name of the tables containing the input data, the network parameters, the model and model statistics are passed as parameters to the BPNN function for creating a model. This function returns the model

in JSON format that is inserted into a model table for future reference. An error value is also returned when the function for creating the BPNN is executed; this error value can be used for evaluating or comparing models.

As discussed in the previous chapter, some of the network topology settings and parameters needed to create an optimised BPNN model for the implementation of the proposed DSS in this study requires an experimental approach for discovery. Iterative experiments were carried out in phases to identify the parameters, which resulted in acceptable models for predicting the spot and futures contract price of white maize. To ensure that the resulting model is able to make generalised predictions after training, the time series data collected was divided into training and testing datasets. There is no generally accepted format for dividing datasets for Neural Networks modelling into training and testing. The general indication is that the training set should be between 60% and 90% of the data available, which means that the test set will range from 10% to 40% of the data available.

In this implementation, the complete and complementary historical datasets available for spot prices of white maize are from January 2007 till July 2015 while the same datasets for the futures contract start from January 2009. However, for the purpose of this implementation, only the datasets from January 2010 will be considered for the modelling of spot prices and from January 2012 for futures prices of white maize. This is primarily to ensure that the knowledge base of the resulting model is based on trends from a reasonable past (Ruta, 2014). The historical data for spot prices will be divided into a training set (01 January 2010 – 31 December 2014) and a testing set (01 January 2015 – 31 July 2015). Datasets from 01 January 2012 – 31 December 2014 will be used as the training set for the futures contract and the test set will be the same as in the case of spot prices of maize.

#### **6.5.1.1 Determination network topologies**

The first phase of experiments was conducted to determine the appropriate structure for a BPNN for the time series-data. The variables identified in Table 6.4 as the factors influencing the spot prices of white maize in South Africa were used to set up an experimental training network. The purpose of this network was to determine the



optimal learning rate, momentum factor and the number of hidden layers that minimise the error. The training of a BPNN is an iterative process during which the network gives weights to the nodes in the network (Alpaydin, 2010). The network calculates the error for each of iterations and sends the error signal back as knowledge that has been learnt so that the network can calculate weights that improve the accuracy of the current iteration (Kabari and Nwachukwu, 2013). This process continues until the network identifies the minimum error. SAP HANA reports the training error of the final iteration of the training process.

During the first phase of experiments conducted to determine the appropriate BPNN structure, the initial learning rate was set to 0.7, momentum factor to 0.001 and the number of hidden layers to 3. This generated a larger training error of 100.1096. The error factor converged by the time the learning rate was set 0.4 with no further significant improvements. The momentum factor was reduced to 0.0001 to get an improvement and the training error converged significantly to 5.1300 with 3 hidden layers.

Activation function for the hidden layer and the output layer was set as sigmoid function as prescribed in the literature (Co and Boosarawongse, 2007; Ghwanmeh, Mohammad and Al-Ibrahim, 2013; Khamis, Nabilah and Binti, 2014), while the functionality parameter setting was set to regression because the analysis was time series. The training style was set to batch, but the normalisation functionality was disabled and coded manually because it generated undesired results.

#### **6.5.1.2 Other structures of the Neural Networks model**

The use of experiments is also suggested as the approach for choosing the other network topology parameters such as the hidden layer, learning rates, momentum factor, and the transfer functions. Hence, this implementation will design experiments to determine the hidden layer, learning rate and momentum function that maximises the accuracy of the model. Also, the sigmoid transfer function will be selected based on previous research on financial and economic time series Backpropagation Neural Network models (Qi and Zhang, 2008; Tsadiras, Papadopoulos and O’Kelly, 2013; Ghwanmeh, Mohammad and Al-Ibrahim, 2013; Khamis, Nabilah and Binti, 2014).

Besides the mentioned network structures, it is suggested that the input data for Neural Network modelling be pre-processed into a normalised format ranging between -1 and 1 or 0 and 1 (Engelbrecht, 2007; Co and Boosarawongse, 2007; Khamis, Nabilah and Binti, 2014). Transforming the input data into a range of 0 to 1 can be achieved by using the equation:

$$y_t = \frac{y_t - y_{min}}{y_{max} - y_{min}} \quad (5)$$

Where  $y_t$  is an observation for time  $t$ ,  $y_{min}$  and  $y_{max}$  are the minimum and the maximum observed values of all the observations of a given variables.

### 6.5.1.3 Number of input and hidden nodes

The number of input nodes required in BPNN for a time series is determined by the number of past observations in a univariate Neural Network model as discussed earlier in this chapter. This number of input nodes is increased when external variables are considered in the model because of the past observation of the external variables. The input variables that are considered for the modelling of spot and December futures contracts prices of white maize are presented in Tables 6.4 and 6.5 respectively. This section provides the result of model-training experiments carried out to determine the number of past observations to be considered for the input variables (lag length) and the number of the hidden nodes.

Iterative model training experiments were carried out with a single lag and a different number of nodes in the hidden layer. These produced results with very high error and no significant difference between the results for each iterations. However, the iterations of the same experiments with increased lag length showed a significant reduction in the error. Furthermore, it was noticed in the subsequent experiments that using more than one lag for the variables representing the demand for white maize and wheat in South Africa produced high training errors. This is perhaps because the variables are made up of disaggregated data, therefore representing false patterns in the dataset. The error introduced can also be interpreted as the fact that the demand data does not contain daily data; therefore, lagged value could likely introduce spurious relationships. Hence, these variables were used as single lags for all the experiments.

Table 6.8: Comparison of BPNN models for spot prices of white maize

Hidden layer	Lags	No Input variable	Average error		Hidden layer	Lags	No input variable	Average error
1	2	15	0.8339		1	5	36	0.8719
2	2	15	0.6829		2	5	36	0.5700
3	2	15	0.6515		3	5	36	0.5622
4	2	15	0.7191		4	5	36	0.5655
5	2	15	0.5260		5	5	36	0.4772
6	2	15	0.6331		6	5	36	0.4800
7	2	15	0.5463		7	5	36	0.4265
8	2	15	0.5980		8	5	36	0.4918
9	2	15	0.6129		9	5	36	0.5231
10	2	15	0.5734		10	5	36	0.5086
1	3	22	0.8331		1	6	43	0.8405
2	3	22	0.6472		2	6	43	0.6310
3	3	22	0.6106		3	6	43	0.5765
4	3	22	0.5541		4	6	43	0.5310
5	3	22	0.5242		5	6	43	0.4872
6	3	22	0.5574		6	6	43	0.5272
7	3	22	0.5389		7	6	43	0.4757
8	3	22	0.5121		8	6	43	0.4662
9	3	22	0.4990		9	6	43	0.4761
10	3	22	0.5316		10	6	43	0.4466
1	4	29	0.8266		1	7	50	0.8510
2	4	29	0.6625		2	7	50	0.6179
3	4	29	0.5747		3	7	50	0.5821
4	4	29	0.5879		4	7	50	0.5175
5	4	29	0.4948		5	7	50	0.4826
6	4	29	0.4887		6	7	50	0.4796
7	4	29	0.5250		7	7	50	0.4552
8	4	29	0.4961		8	7	50	0.5029
9	4	29	0.4945		9	7	50	0.5009
10	4	29	0.5696		10	7	50	0.5287

For each combination of the number of nodes in the hidden and lag length for the input layer, the same experiments were carried out six times. This was done in order to capture the best representation of the error factor for each combination by taking an average of the errors from six experiments with the same criteria. Table 6.8 shows the summary of the results of the 360 experiments that were carried out with the BPNN model for spot prices of white maize in South Africa. The best model is considered to be the model with the least error factor and is highlighted in red in Table 6.8. The result of the experiments indicate that the model created with seven nodes in the hidden layers and five lags could mean that the effect of the market behaviour in the previous week was most relevant for predicting daily prices of white maize. It was also noted that this combination produced the model with the least error factor in the entire 360 experiments and the error factor for each of the six experiments for the optimal combination had very low deviations from the mean value. Hence, this model will be used in this implementation for predicting the spot prices of white maize.

The same experiments were carried out to identify the optimal BPNN model for the December future contract prices of white maize in South Africa. The variables that were identified in the results, displayed in Table 6.5, were used to train BPNN models with different combinations of lag lengths (input variables) and hidden layers. Six experiments were also carried out for each combination and the average of the error factors of the six experiments was captured as a fair representation of the error factor for each combination of input and hidden layers. The summary of the results from the experiments is presented in Table 6.9. The pattern of the results in Table 6.9 for the December futures contract prices of white maize follows the results for the spot prices presented in Table 6.8. This suggests that there are common similarities in the trading strategies for both the spot and December futures contract prices of white maize.

Each of the experiments for the December future contract prices of white maize as, well as those of the spot prices, was configured to exit after a maximum of 50,000 iterations in search of the output with the best generalisation ability based on the training data. As presented in Table 6.9, the result of the experiment for the December futures contract indicates that the resulting models improved as the number of the nodes in the hidden layers was increased. This is evident with the increase in the lag

length. However, an optimal model was identified with seven nodes in the hidden layers and a lag length of 5, following the results that were obtained with the spot prices.

Results of the experiments conducted to identify the optimal lag length and number of hidden nodes suggest that the daily grain commodities trading in South Africa is largely influenced by activities of the previous trading week. However, it is suspected that this will be different with datasets in a more granular format such as for every hour, minute or seconds. However, this implementation will be limited to the daily data. Finally, based on results from both groups of experiments shown in Tables 6.8 and 6.9, a total of 36 and 51 input variables will be used for training BPNN modelling for the spot and December futures contract prices of white maize respectively. This is specific to this study and the same framework can be followed to identify the right network topology when the structure of the dataset used is different.

Table 6.9: Comparison of BPNN models for December futures contract prices of white maize

Hidden layer nodes	Lags	No Input variable	Average error		Hidden layer nodes	Lags	No input variable	Average error
1	2	21	1.2016		1	5	51	0.8749
2	2	21	0.6853		2	5	51	0.6555
3	2	21	0.6072		3	5	51	0.6188
4	2	21	0.5314		4	5	51	0.5016
5	2	21	0.5734		5	5	51	0.4922
6	2	21	0.5412		6	5	51	0.4878
7	2	21	0.5347		<b>7</b>	<b>5</b>	<b>51</b>	<b>0.3668</b>
8	2	21	0.5537		8	5	51	0.4719
9	2	21	0.5345		9	5	51	0.4711
10	2	21	0.4948		10	5	51	0.4917
1	3	31	1.1409		1	6	61	1.0993
2	3	31	0.5649		2	6	61	0.5680
3	3	31	0.5333		3	6	61	0.5635
4	3	31	0.5397		4	6	61	0.5354
5	3	31	0.4964		5	6	61	0.5222

<b>6</b>	3	31	0.4944		<b>6</b>	6	61	0.4802
<b>7</b>	3	31	0.5011		<b>7</b>	6	61	0.4758
<b>8</b>	3	31	0.4996		<b>8</b>	6	61	0.4806
<b>9</b>	3	31	0.4577		<b>9</b>	6	61	0.4704
<b>10</b>	3	31	0.4532		<b>10</b>	6	61	0.5106
<b>1</b>	4	41	1.0800		<b>1</b>	7	71	1.3615
<b>2</b>	4	41	0.6213		<b>2</b>	7	71	0.8067
<b>3</b>	4	41	0.5007		<b>3</b>	7	71	0.6221
<b>4</b>	4	41	0.5098		<b>4</b>	7	71	0.5588
<b>5</b>	4	41	0.5150		<b>5</b>	7	71	0.5606
<b>6</b>	4	41	0.4699		<b>6</b>	7	71	0.6329
<b>7</b>	4	41	0.4940		<b>7</b>	7	71	0.4882
<b>8</b>	4	41	0.4569		<b>8</b>	7	71	0.4874
<b>9</b>	4	41	0.4717		<b>9</b>	7	71	0.5406
<b>10</b>	4	41	0.5013		<b>10</b>	7	71	0.4611

#### 6.5.1.4 Verification of models

Results from the previous phase of the experiments indicate that the model with five lags and seven nodes in the hidden layer is likely to perform better. Using this guideline, BPNN models for both the spot and December futures contract prices of white maize in South Africa were created. Subsequently, a verification process was also carried out to determine the generalisation ability of the models. Each of the models was used to make predictions by using subsets of the training dataset and the testing dataset. Table 6.10 presents the result of the evaluation of the BPNN models for the spot price of white maize in South Africa that was trained using historical data of transactions that happened between 01 January 2010 and 31 December 2014. In-sample evaluations were carried out with subsets of the training data while out-of-sample evaluations were carried out with subsets of the testing data. For both categories, the created model was used to make predictions for 1, 3 and 6 month periods.

Table 6.10: Summary of verification of BPNN model for spot prices

Period	In-sample			Out-sample		
	MAPE(%)	RMSE	R <sup>2</sup>	MAPE(%)	RMSE	R <sup>2</sup>
<b>1 month</b>	1.31	32.97	0.6568	2.26	61.02	0.1412
<b>3 month</b>	0.97	24.61	0.9709	9.20	348.64	0.9598
<b>6 month</b>	1.12	25.91	0.9862	12.08	429.08	0.9480

The dataset for the trading days in the last month of the training data from 01 December 2014 to 31 December 2014 was predicted and compared against the actual prices. In-sample predictions were also made and compared with actual prices over a period of 3 months using data from 01 October 2014 till 31 December 2014. Furthermore, a subset of the training dataset from 01 July 2014 till 31 December 2014 was used for a 6-month period of in-sample training. Figures 6.17, 6.18, 6.19, 6.20, 6.21 and 6.22 present a graphical comparison of in-sample and out-sample predictions and actual spot prices of white maize in short (1 month), medium (3 months) and long term (6 months).

To measure the prediction accuracy of the model, the Mean Absolute Percentage Error (MAPE) statistic for the in-sample and out-sample prediction for the three different periods was compared. The results show a 0.95% difference in the Mean Absolute Percentage Error (MAPE) between the performance of the model when used in-sample and out-sample over a period of a single month. The difference in the MAPE was 8.23% for 3-month predictions and 10.96% for 6 months. However, the correlation between the predicted prices and the actual prices for the 6-months in-sample prediction was 0.9862 and 0.9480% for the out-sample prediction. These results suggest that the model is able to generalise and make predictions for unseen data, although there is room for further research into improving the model.

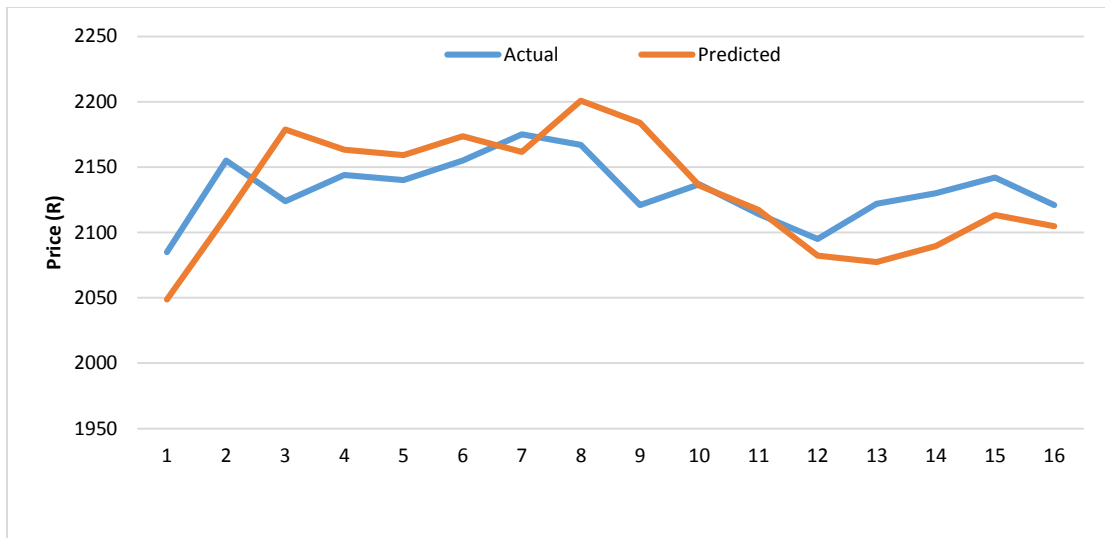


Figure 6.17: Comparison of actual vs predicted spot prices of white maize (1 month in-sample)

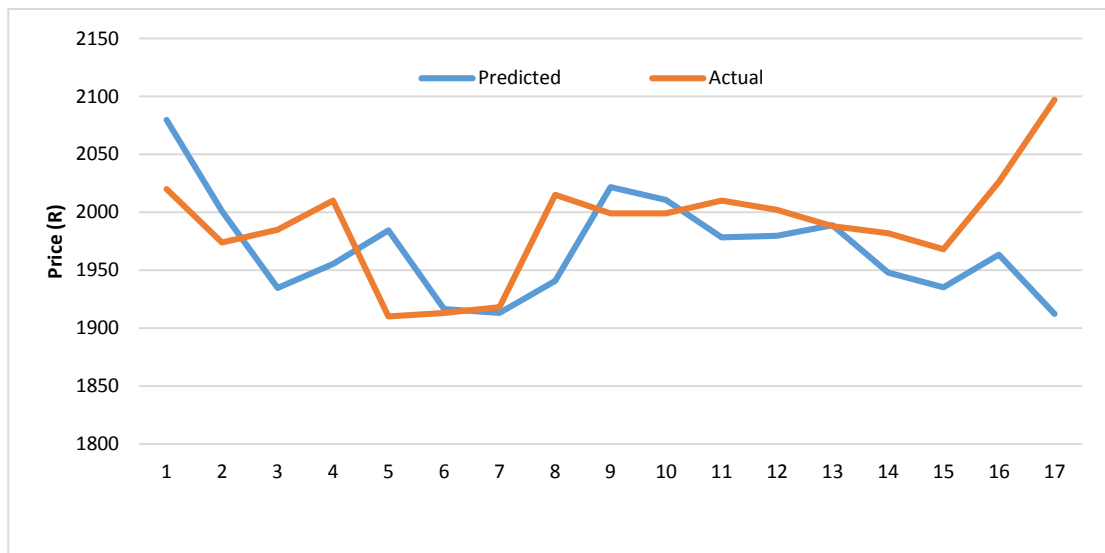


Figure 6.18: Comparison of actual vs predicted spot prices of white maize (1 month out-sample)

Figures 6.17 and 6.18 draw a comparison between the in-sample prediction and the out-sample predictions of the spot price of white maize for the available data over a period of 1 month. The graphs show that the in-sample predictions are very close to the actual values and the figures also indicate that the in-sample predictions followed the trend of the actual values much better than that of the out-sample predictions in Figure 6.18.



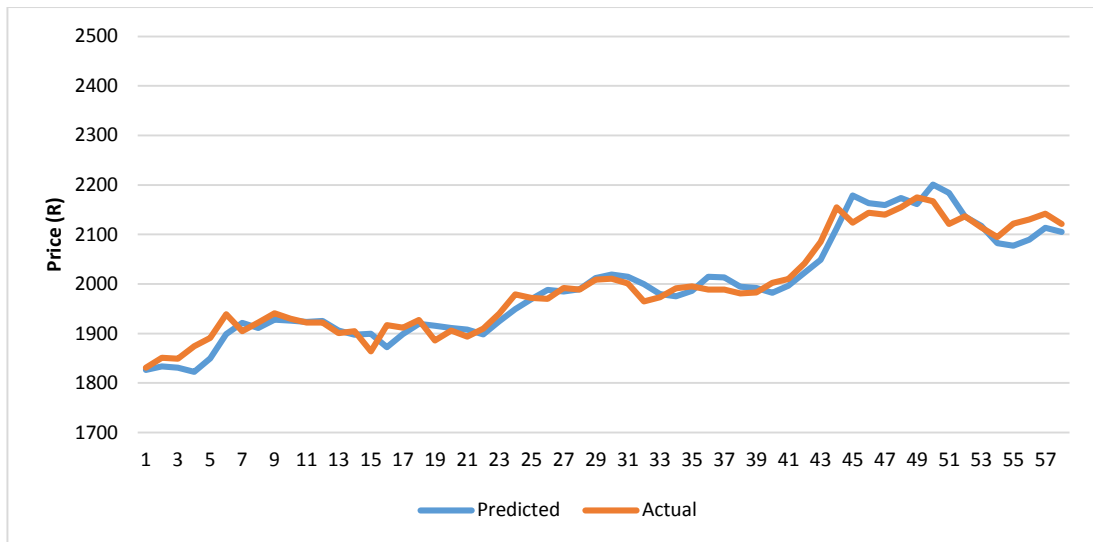


Figure 6.19: Comparison of actual vs predicted spot prices of white maize (3 months in-sample)

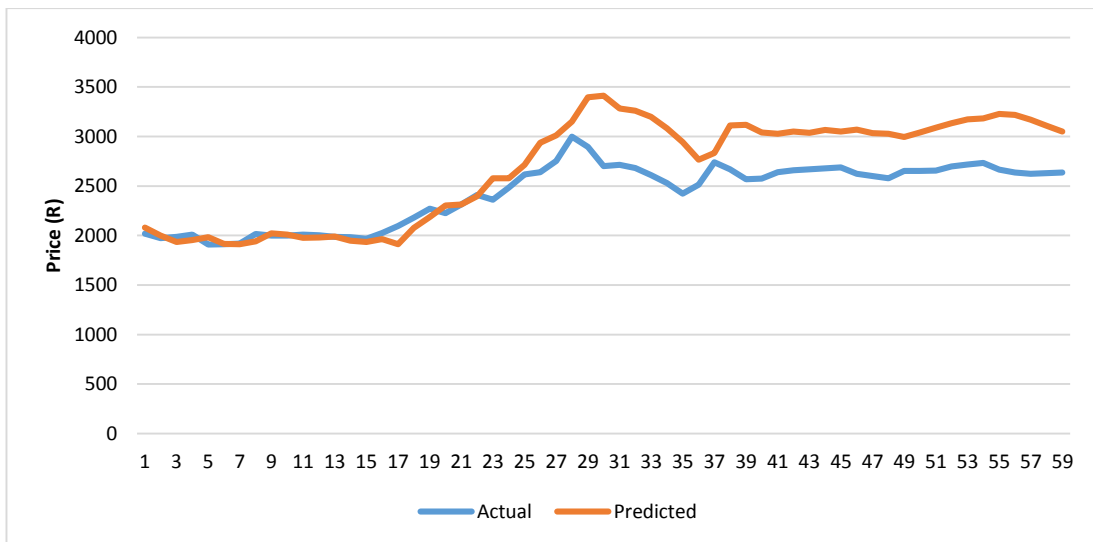


Figure 6.20: Comparison of actual vs predicted spot prices of white maize (3 months out-sample)

The results presented in Figures 6.19 and 6.20 for the predictions over a period of 3 months indicate an improvement in the accuracy of the predictions when compared with the prediction for transactions over a 1 month period. Both graphs show that the predicted values for both the in-sample and the out-sample predictions followed the trend of the actual prices very closely. This indicates that the models performed better when predicting with more data. However, the out-sample predictions deviated quite

reasonably from the actual value in the latter half of the predictions, signifying the need for optimisation of the model.

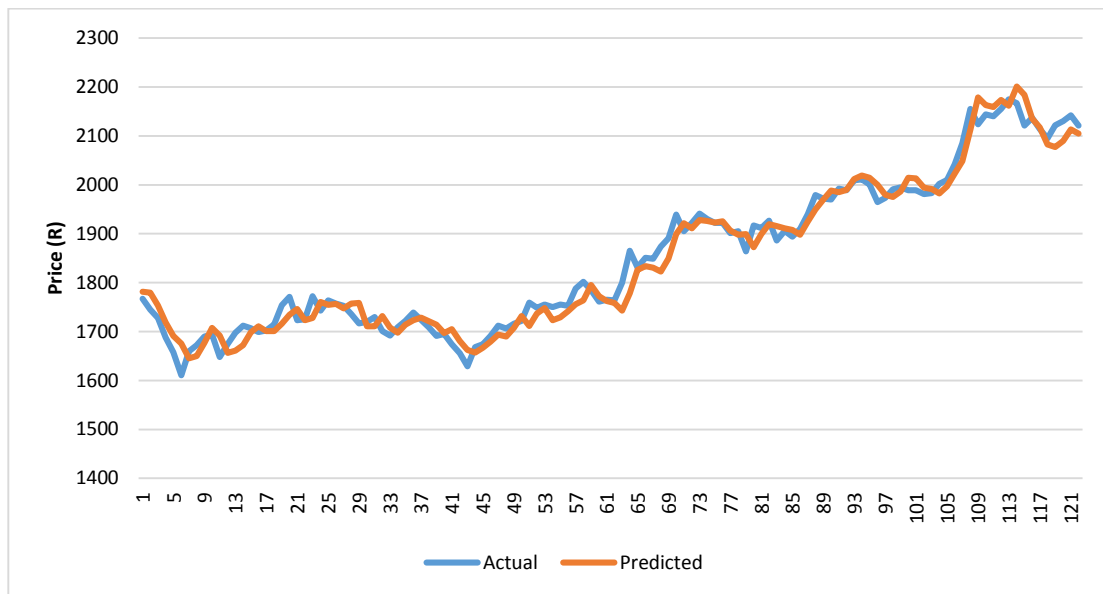


Figure 6.21: Comparison of actual vs predicted spot prices of white maize (6 months in-sample)

The 6-months in-sample and out-sample predictions of the spot price of white maize is represented in Figures 6.21 and 6.22 respectively. Both predictions exhibited the same pattern that was seen with the 3-months in-sample and out-sample predictions. The in-sample and out-sample predictions followed the trends of the actual values very closely, even when the market trend changed direction significantly. However, Figure 6.22 shows that the prediction in the out-sample experiment deviated significantly from the actual value after about the 30-day prediction. It is obvious that this is an issue specifically with the model because the deviation of the predicted value from the actual values in the 3-month out-sample experiment also started at about the thirtieth day, further indicating the need to find ways to optimise the model.

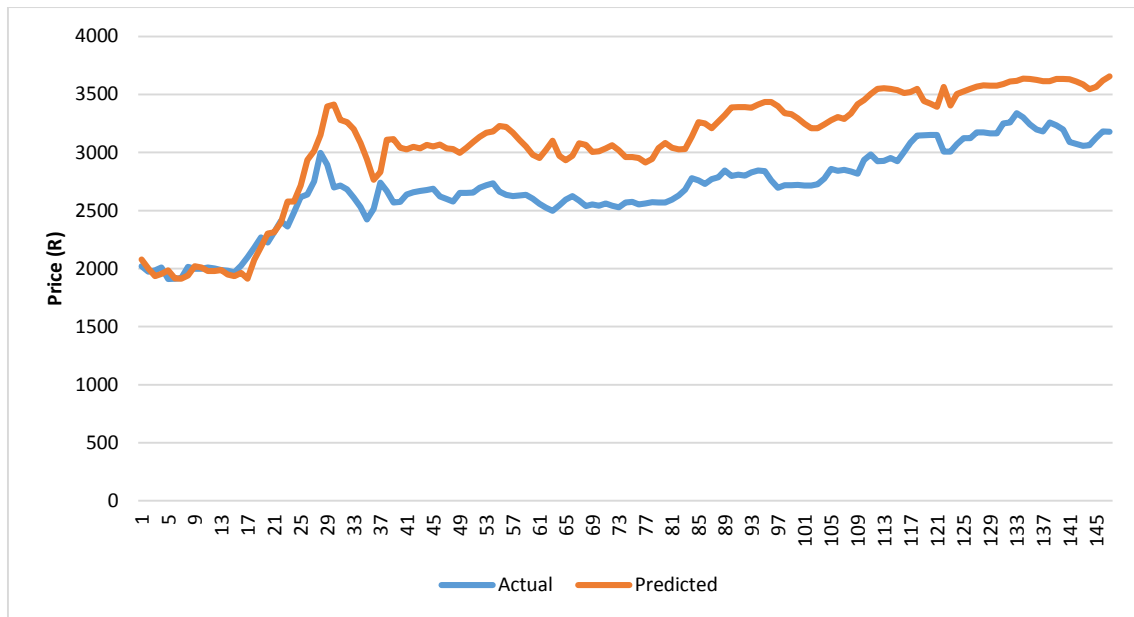


Figure 6.22: Comparison of actual vs predicted spot prices of white maize (6 months out-sample)

The predictions depicted by the graphs in Figures 6.17 – 6.22 show that the model is more accurate with in-sample predictions as expected, especially for predictions over 3 and 6 months. When the same model is applied for making out-sample predictions using input dataset that was not used for the training process, the models were less accurate. However, the results of the out-sample predictions suggest that the models were intelligent enough to recognise the market trend although the deviation between the actual and the predicted price increased significantly with time. This result suggests that the identified topology and architecture used in building the model for predicting spot prices of white maize in South Africa could be used in a DSS designed for grain commodities trading. Although, there is a need to implement strategies that will improve the accuracy of the predictions.

The model verification process above was also followed to ascertain the possibility of obtaining a BPNN model for predicting the future prices of the December futures contract of white maize. Unlike the spot prices verification, data from transactions between 01 January 2012 and 31 December 2014 were used to train a model. This was done to ensure that the model was focused on the recent trend of the market as suggested by Ruta (2014). In-sample testing was carried out by using a subset of the

datasets used in training the model between 01 July 2014 and 31 December 2014 to make predictions for 1, 3 and 6-month periods. The result of the verification exercise is presented in the Table 6.11.

Table 6.11: Summary of verification of BPNN model for December futures prices

Period	In-sample			Out-sample		
	MAPE(%)	RMSE	R <sup>2</sup>	MAPE(%)	RMSE	R <sup>2</sup>
<b>1 month</b>	0.78	18.14	0.9014	1.78	50.58	0.4916
<b>3 month</b>	0.84	19.83	0.9758	2.50	86.74	0.9590
<b>6 month</b>	0.84	19.91	0.9734	2.90	100.49	0.9290

The results in Table 6.11 show that the model performed reasonably well when the Mean Absolute Error (MAPE) for the in-sample evaluation is compared with the out-sample test. Moreover, the predicted prices of the December futures contract from the model also show fairly significant correlations with the actual prices. The 6-month predictions showed a 0.9734 correlation for in-sample dataset and a 0.9290 correlation for out-sample predictions over the same period. The Root Mean Square Error (RMSE) of the out-sample predictions is between 3 to 4 multiples of that of the in-sample for the different categories. This shows that the predictions from the in-sample predictions have a much smaller degree of errors. On a positive note, the results indicate that the model has an ability to reliably make predictions and can produce better results when optimised.

Correlation analysis of the actual vs predicted prices for the 3-month predictions shows a 0.9758 correlation for the in-sample prediction and 0.9590 for out sample testing. This is about the same with the 6-month predictions showing a 0.9734 correlation between the actual and predicted prices for in-sample predictions and 0.9290 for out-sample predictions. But the 1-month predictions showed a 0.9014 correlation between actual and predicted prices in-sample and 0.4916 for the out-sample analysis.

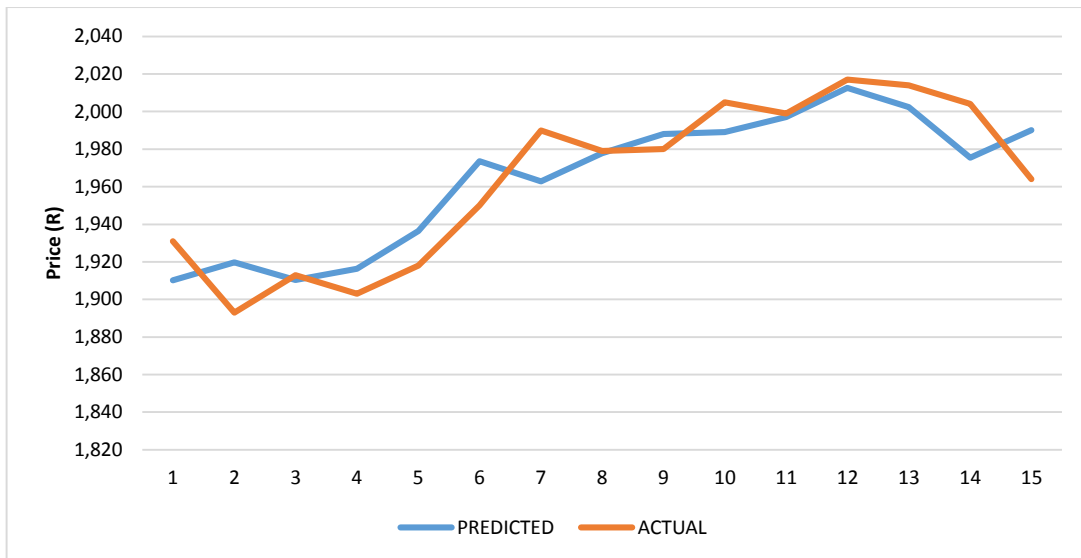


Figure 6.23: Comparison of actual vs predicted December futures contract of white maize (1-month in-sample)

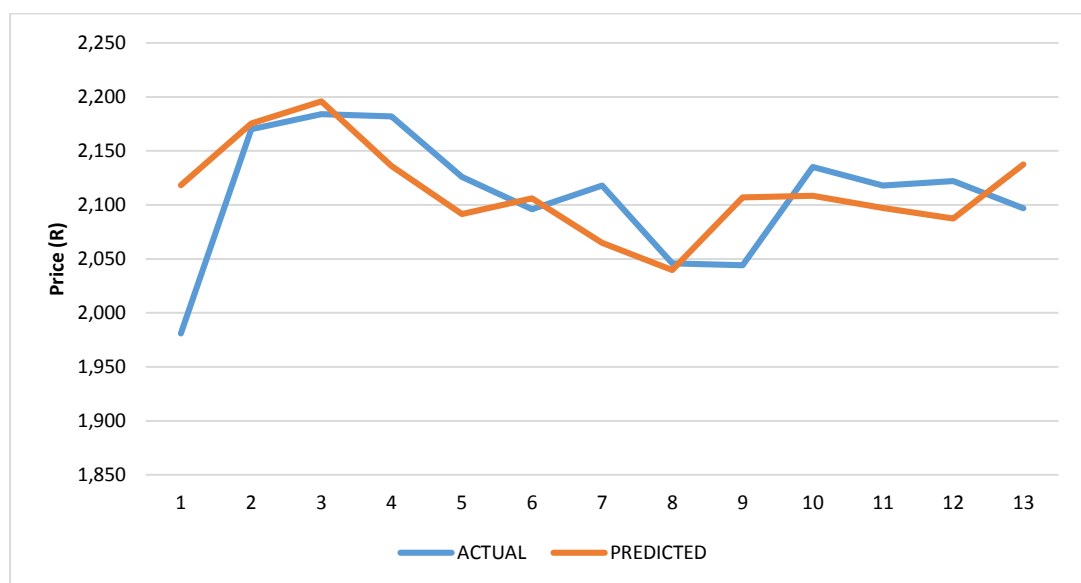


Figure 6.24: Comparison of actual vs predicted December futures contract of white maize (1-month out-sample)

When the in-sample predictions in Figure 6.23 are compared to the out-sample predictions in Figure 6.24, the in-sample predictions are much better in following the trend in the actual values and the predicted values are much closer to the actual values as expected. However, the out-sample predictions in Figure 6.24 showed some degree of accuracy with a few of the predicted values being almost equal to the actual

values. This is quite promising, but then it can also be seen that some of the predicted values seem to be far apart from the actual values, indicating the need to explore ways to increase the accuracy of the model.

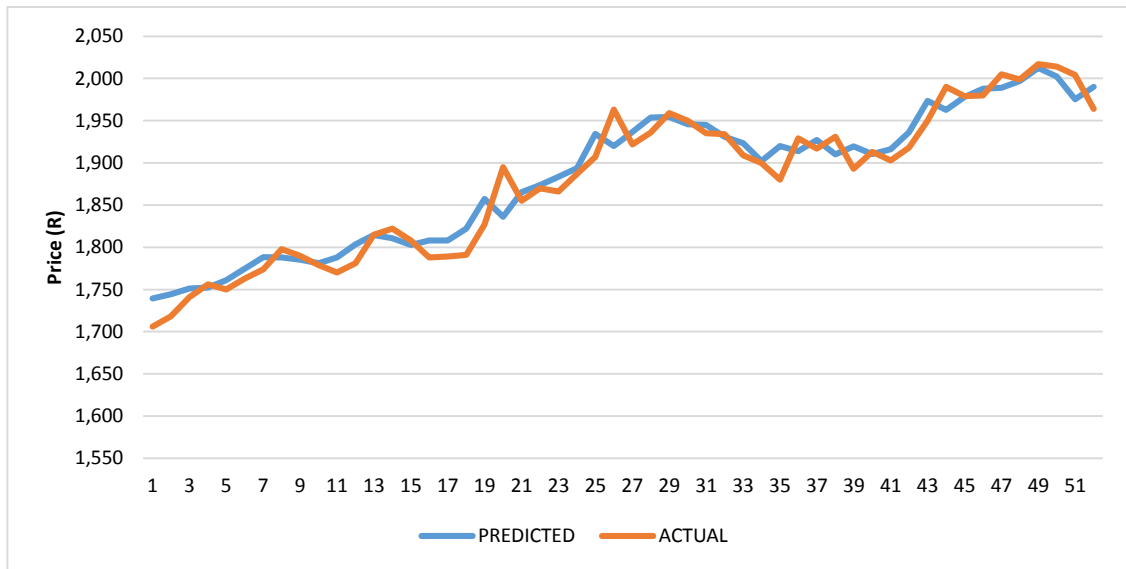


Figure 6.25: Comparison of actual vs predicted December futures contract of white maize (3-month in-sample)

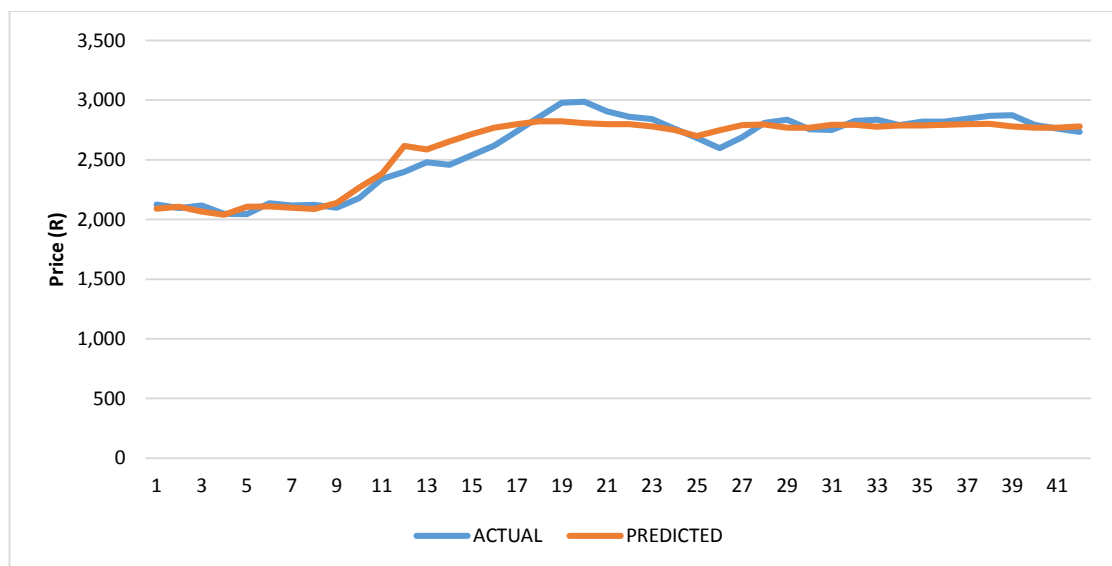


Figure 6.26: Comparison of actual vs predicted December futures contract of white maize (3-month out-sample)

The results of the in-sample and out-sample 3-months predictions for the December futures contracts is presented in Figures 6.25 and 6.26 respectively. The in-sample

predictions are quite close to the actual values and follow the trend quite well. But the out-sample predictions are also quite close to the actual values and follow the trend in the actual values quite closely. This result is much better than that obtained for the 3-months out-sample predictions of the spot prices of white maize conducted earlier. An indication that the model developed for the December futures contract prices might render a better performance than that of the spot prices. This is also evident when the performance measurement statistics in Tables 6.9 and 6.10 are compared. One factor that might be responsible for this could be that data for more relevant factors was identified for the modelling of the December futures contract prices of white maize than what was available for the modelling spot prices of white maize during this implementation. This again highlights the opportunities to extract better insights with more datasets.

Figures 6.27 and 6.28 show the graphical representation of the 6-months in-sample and out-sample predictions of the December futures contract prices respectively. The in-sample predictions in this case also show more accurate predictions than the out-sample predictions. But, the out-sample predictions of the December futures contract prices also showed an improvement when compared to the 6-months out-sample predictions for the spot prices. This supports the suggestion that the model developed for the December futures contract might be better than the model used in predicting the spot prices of white maize. The results of the out-sample predictions shown in Figure 6.26 indicate that the model will be able to recognise and adjust to market shocks. This is a highly desirable characteristic of a DSS to provide early warnings of likely significant changes in the market that can be used to the advantage of the decision maker either to make more profit or prevent huge losses.

The graphical representation of the results above shown in Figures 6.21, 6.22, 6.23, 6.24, 6.25 and 6.26 emphasise the ability of the Backpropagation Neural Network architecture selected to predict the December futures contract prices of white maize in South Africa.

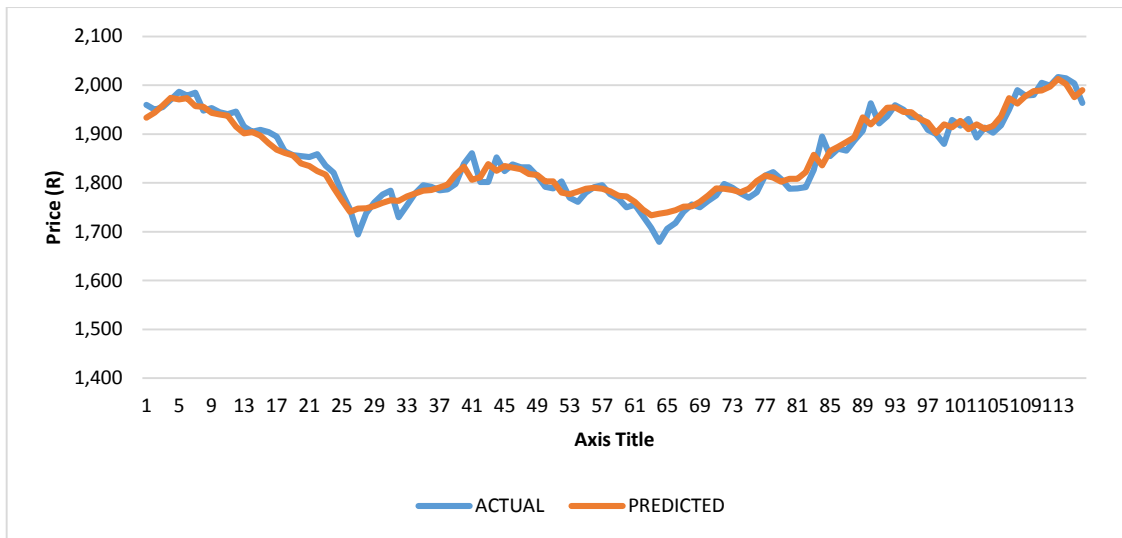


Figure 6.27: Comparison of actual vs predicted December futures contract of white maize (6-month in-sample)

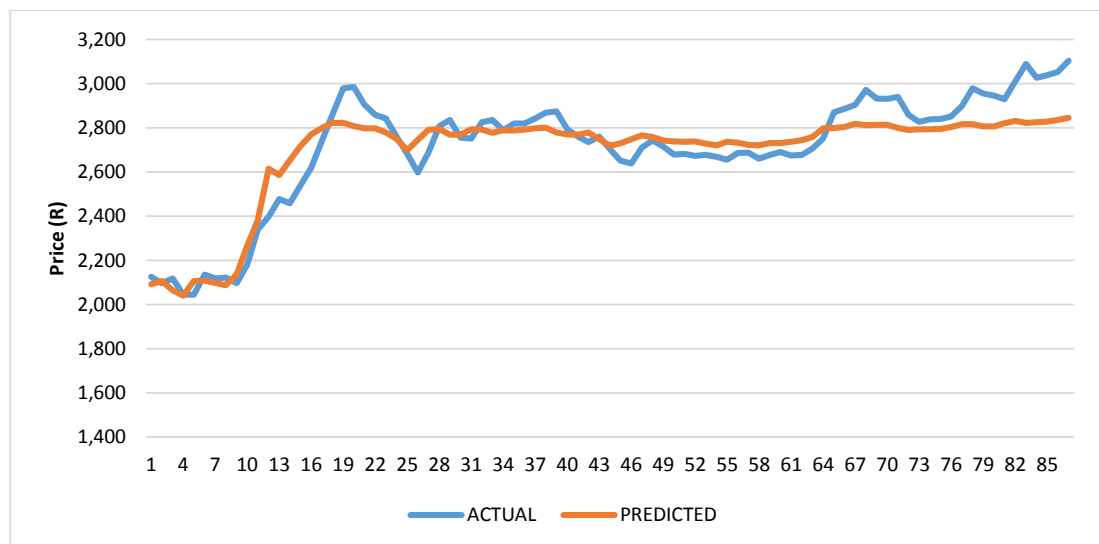


Figure 6.28: Comparison of actual vs predicted December futures contract of white maize (6-month out-sample)

These results indicate that the identified BPNN topology and parameters used can be used for building predictive models for the spot and futures contract prices of white maize in South Africa. The verification process carried out in this section further confirms that relationships exist between the spot and futures contract of white maize prices and the external variables that were identified earlier. Moreover, the results show that a BPNN model can learn these relationships and make future predictions



based on the patterns identified from data within a reasonable time. Chapter 7 of this study will provide further evaluation that further validates the usefulness of the model as an important component of the grain commodities trading decision support system.

## **6.6 Visualisation of Market Intelligence**

Section 6.5 showed that future prices of two different trading strategies can be obtained with the implementation of BPNN models. It was demonstrated that the accuracy of the predictions can be improved by introducing real-time data and dynamic model training. This is an important market intelligence that could be served to farmers or stakeholders in a simple manner for them to understand. As it has been shown in Section 6.5, the use of graphs could be employed to display market intelligence in real-time as trends change.

The results of predictions and data collected on several factors that influence the market are available as data tables that are updated regularly according to the implementation in the previous sections of this chapter. Moreover, the data and all the other processes were implemented by using the SAP HANA instance that was hosted as an Amazon Web Service. It is therefore possible for mobile and web interfaces that serve this intelligence can be developed for a different category of users. The integrated data source and platform provide a platform that can be used for other discoveries and predictions which can be served to farmers and stakeholders for making informed trading decisions about grain commodities. It is expected that further studies could be carried out on how this market intelligence can be served to different category of users that make grain commodities trading decisions.

## **6.7 Conclusion**

The initial part of this chapter proposed a framework for support in making trading decisions about grain commodities which resulted from the critical review of literature and the ideas developed from Chapters 3, 4 and 5 of this study. The proposed framework comprises the domain knowledge component for understanding the factors that influence the grain commodities trading market. It also includes the real-time data acquisition and integration component for sourcing and integrating data from several

sources in real-time or near real-time so that the trends in the market can be captured and used to predict the future. The proposed framework also includes a modelling component for determining the relationships and patterns that exist in the data as it is collected. Statistical time series and the use of Neural Networks were suggested as alternatives for modelling the patterns in the data; however, previous research indicates that statistical time series models such as the Box-Jenkins methodology are less effective for non-linear data like the grain commodities market statistics. Hence, this study is limited to the use of the Neural Network for the modelling.

The proposed framework suggests that an intelligence component can be built on the results of the modelling component for extracting actionable insights such as discoveries, predictions or recommendations. This will be presented to the users through the visualisation component of the proposed framework. It was suggested that a Big Data approach should be taken for a successful implementation of the framework. This is particularly important for the acquisition, integration and execution of other components of the framework. Therefore, the proposed framework suggests a technological consideration as a critical component of the framework that will provide the enabling environment for the other components in the framework.

The second half of this chapter presented an implementation of the proposed framework by using two trading strategies of white maize on the Johannesburg Stock Exchange as a case study. SAP HANA was used as the technology of choice because it provides easy-to-implement data streaming, mining and modelling functionalities through a scripting interface. Historical and near real-time daily data for the two trading strategies was collected, integrated and loaded into SAP HANA. By using the embedded Backpropagation Neural Network library, predictive algorithms were developed and verified as suitable for predicting the spot and December future prices of white maize in South Africa.

SAP HANA as a Big Data platform used in this implementation is built for cloud, parallel and in-memory computing which allows for the retraining and scripting of Neural Network models for improving the accuracy of the predictions. Furthermore, it provides a platform for the collection and integration of data from several sources in real-time

or near real-time with opportunities for the modelling components to capture new trends in the data. These results in predictive models could be incorporated in the DSS to enable grain farmers and other stakeholders to make informed trading decisions on time. Besides, they provide a platform for an end-to-end implementation of the components in the proposed framework.

Preliminary verification of the resulting models for the spot and the December futures contract prices of white maize shows that Neural Networks model can indeed be used to model the patterns in market data and those of the factors influencing them. The verification exercise suggests the need to include observations from the previous one week as input data in training the BPNN models for both trading strategies of white maize. The resulting models were used to make predictions using subsets of the data that was used in training the models and subsets that were not included in the training of the model.

This exercise indicates that the models did not just memorise the observations but are able to generalise. However, it also indicates the need to find strategies that can improve the accuracy of the model. Thus, this chapter demonstrates that it is possible to capture and integrate disparate datasets in order to model the volatility of grain commodity prices based on the proposed framework. The model created can be used to generate market intelligence and actionable insights that provide decision supports when trading grain commodities in South Africa.

Predicting the future prices and performance of the different trading strategies for the same grain could assist grain commodities traders in making more informed decisions about the right trading strategy and timing to adopt. This could reduce their price-related risk and increase profitability. Besides deciding the right strategy to adopt and time to sell, the proposed framework also offers the possibility of the development of other market intelligence and actionable insights that could be beneficial for the trading of grain commodities in South Africa. This chapter fulfils the DSR requirement to demonstrate how the artefact that is built could provide a solution to a scenario or case study of the problem that has been identified.

The framework for support in making trading decisions about grain commodities that was proposed in this chapter fulfils the research objective RO<sub>3</sub> – to develop a framework to support decisions on grain commodities trading. The proposed framework and subsequent implementation also provide answers to the research question RQ<sub>5</sub> – How can a framework for a system to support decisions about trading grain commodities be developed and implemented? The next chapter of this study will provide an empirical evaluation of the proposed DSS. During the evaluation process, experiments that increase the accuracy of the models explored in this chapter will also be conducted. It will seek to validate the ability of the modelling component to predict the spot prices and the December futures prices of white maize in South Africa. Furthermore, the empirical evaluation in Chapter 7 will assess the ability of the implementation of the proposed framework to improve grain commodities decision making.

# Chapter 7 : Empirical Evaluation

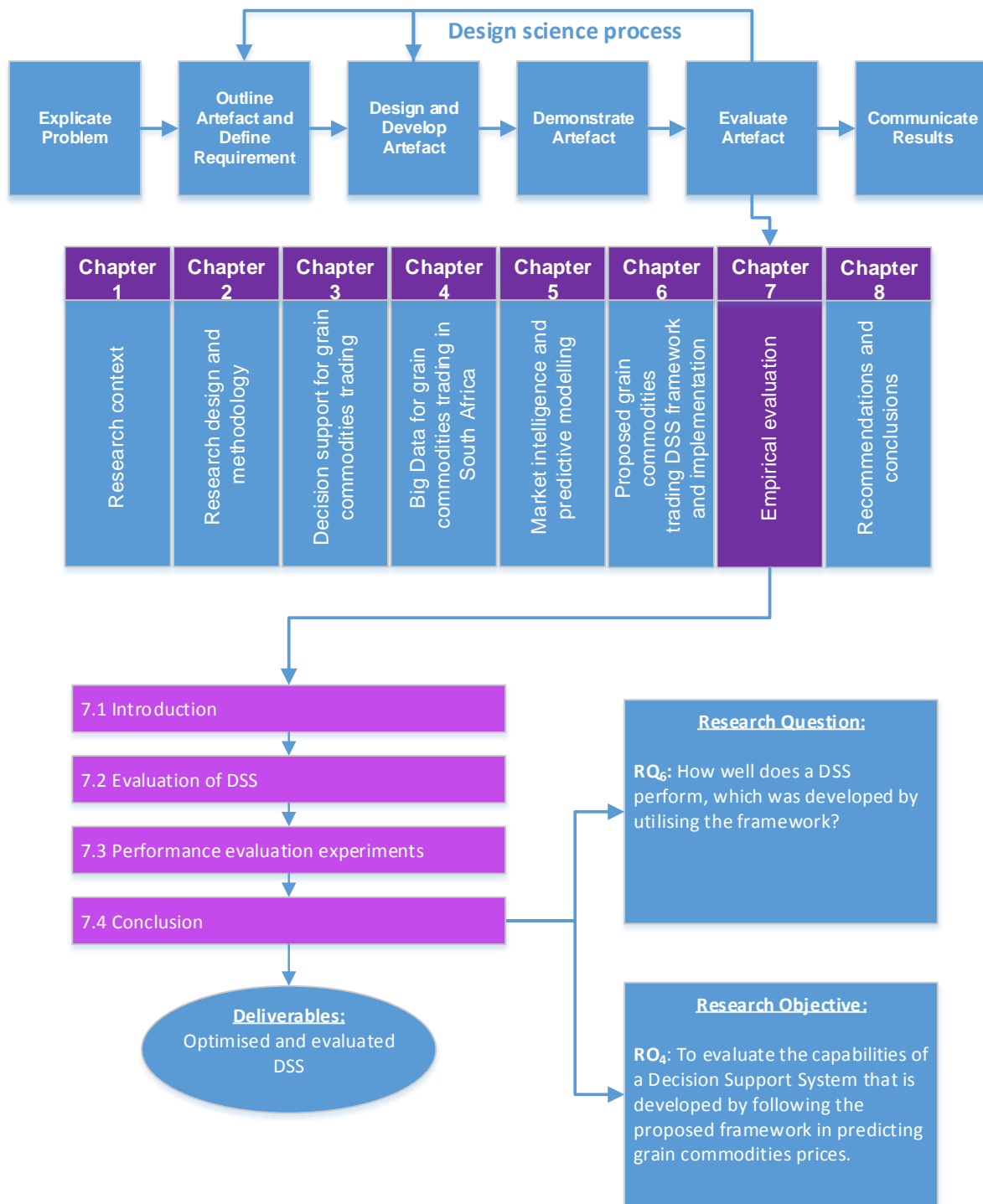


Figure 7.1: Chapter outline and deliverables

## 7.1 Introduction

A framework was proposed to support trading decisions concerning grain commodities for South African farmers and other stakeholders in Chapter 6. The proposed framework was implemented in Chapter 6 using the spot and December futures contract prices of white maize on the Johannesburg Stock Exchange. Components of the proposed framework include domain knowledge, data acquisition, modelling, visualisation and technological consideration. It was identified that the choice of which trading strategy to adopt and the best time to sell poses a big challenge due to the volatility of the markets. Therefore the ability to predict future prices and the performance of different trading strategies could assist farmers to make better decisions. Chapter 3 provided an explication of the problem; price-related risks and price discovery were found to be the major concern when making a decision about trading grain commodities.

Based on the identified need to predict future prices of grain commodities, the factors that influence the prices of grain commodities in South Africa were identified. Chapter 4 presented techniques and principles that could be followed to acquire and integrate relevant datasets that influence the grain commodities market. Analytical methods and principles for identifying patterns that exist in data were discussed in Chapter 5. Thereafter, a framework was proposed in Chapter 6 as a deliverable of the rigour cycle within the DSR methodology that was discussed in Chapters 3, 4 and 5. An implementation of the proposed framework was described in Chapter 6 as a validation of the ability of the proposed framework to solve the identified problem.

Within the DSR methodology, the evaluation of the design artefact can be used to give an indication of how well the artefact is able to address the identified practical problem (Johannesson and Perjons, 2012). After using the demonstration phase of the research to show how the artefact will solve the identified problem, it is common practice to conduct a scientific evaluation of the artefact that was developed by the research process (Hevner, 2007). The purpose of the evaluation in the DSR process is to ascertain the level of performance and how well the artefact, that was built in the design cycle, solves the identified problem (Hevner, 2007; Geerts, 2011; Johannesson

and Perjons, 2012). Therefore the evaluation phase could lead to an iteration with other phases in the research process, especially the design phase, in a bid to ensure that an artefact, the final artefact, addresses the problem properly.

This chapter will evaluate the outcome of the DSS implementation that was described in Chapter 6. Chapter 7 will provide an evaluation of the proposed framework by using an optimised implementation of the BPNN model that has been selected, based on iterative experiments discussed in Chapter 6. Two iterations will be carried out, the first will evaluate the DSS, based on the design in Chapter 6 by comparing the predictions from the DSS to those of a panel of experts that are experience traders in the South African grain commodities trading industry. Based on the result of the first iteration, the DSS will be optimised and used to make second iterations.

The purpose of this chapter is to address the research objective RO<sub>4</sub> – to evaluate the capabilities of a Decision Support System that is developed by following the proposed framework in predicting grain commodities prices. Chapter 7 will address the research question RQ<sub>6</sub> – How well does a DSS perform, which was developed by utilising the framework? In order to achieve this objective and provide an answer to the research question, Section 7.2 of this chapter will provide an overview and motivation for the evaluation strategies that were adopted in this study. Section 7.3 discusses the result of the iterative experiments carried out to validate the performance of the models used in creating intelligence (predictions) in the implemented framework by using real-time datasets. Finally Section 7.4 will provide concluding remarks on the chapter and the evaluation process.

## **7.2 Evaluation of DSS**

A Decision Support System can only be considered to be of any value if it performs in a way that reduces uncertainty and equivocality – ambiguity and lack of understanding during the decision-making process. Uncertainty can be reduced by having access to more information while equivocality can be dealt with by the availability of information that is of a better quality (Kowalczyk and Buxmann, 2014). Furthermore, the implementation of a DSS is considered to be successful if it enables decision makers

to choose correctly, more quickly and easily (Sauter, 2010). Therefore the value of a DSS could be determined by the amount and quality of information that it provides for decision makers. It is important to evaluate the ability of a DSS to effectively provide decision support to ensure that it does what it was built to do, and that it adds real value to the decision-making process, otherwise the DSS will be of lesser or of no real value (Phillips-Wren, Mora, Forgionne and Gupta, 2009).

Over the years, several perspectives and suggestions have emerged on the evaluation of a DSS (Sojda, 2007). The success of a DSS can be measured by evaluating the value it adds to the decision-making process and the quality of the outcome (Phillips-Wren et al., 2009). Broadly, these can be achieved by evaluating its technical capabilities and overall usefulness based on the needs of an organisation or specific decision maker. The technical capabilities and usefulness of a DSS can be evaluated by using the criteria presented in Table 7.1 as highlighted by Sauter (2010).

Table 7.1: Evaluation criteria for DSS

<b>Technical capabilities</b>
<ul style="list-style-type: none"> <li>• Are the features consistent with user information needs?</li> <li>• How many of such features are possible?</li> <li>• Does the DSS have appropriate models implemented?</li> <li>• Do the implemented models deliver on promise?</li> <li>• What is the degree of accuracy of the implemented models?</li> </ul>
<b>Overall usefulness</b>
<ul style="list-style-type: none"> <li>• Does the DSS solve real problems?</li> <li>• Will experienced decision makers find the DSS appropriate and reasonably intelligent?</li> <li>• Are the outcomes of the DSS similar or better than recommendations from experts that did not use the DSS?</li> </ul>

The proposed framework in this study has been crafted to enable the creation of additional features. Based on the study of literature and the study conducted among farmers and traders, discovery of prices for different grain trading strategies is considered relevant to the decision-making process for grain trading. Furthermore, the implementation of the proposed framework in Chapter 6 confirmed that the Backpropagation Neural Networks model and the topologies identified are appropriate



for modelling the relationships that exist in the price data and the identified external variables.

During the implementation of the proposed framework in Chapter 6, it was verified that the implemented models have the ability to generalise and make predictions using data that was used as training input. However, Sojda (2007) proposed that further verification can be carried out by testing the performance of the model against pre-selected standards. Where the DSS involves the use of both real-time and historical data, as in the case of this study, Sojda (2007) suggested that a subset of the dataset can be set apart or dynamically selected for use in a data-driven model. Moreover, the author also suggested that a statistical comparison can be made between the outcome of the DSS and an analysis of the results by a panel of experts who do not have access to the DSS which is being evaluated. The rest of this chapter will explore the empirical evaluation that provides an indication of the accuracy and usefulness of the modelling component of the proposed framework which was implemented in Chapter 6.

### **7.3 Performance Evaluation Experiments**

The implementation that was carried out in Chapter 6 made predictions based on the assumptions that the external data is available for the period for which the price of white maize is being predicted. However, as proposed in Section 5.5.1 with the model denoted as equation (4), models can be built based on all the available data for predicting the spot and futures contract prices for different days into the future. These models can then be retrained periodically as new data becomes available to ensure that new market dynamics are captured in the Neural Network. This will form the basis of experiments for testing the accuracy of the models proposed in Chapter 7. The evaluation of the accuracy of the models was extended to seek an optimisation of the models implemented in Chapter 6.

Beside the use of the measurement of accuracy statistics to measure the technical abilities of the models, the rest of this chapter will also be comparing the output of the models with predictions from a panel of experts as a measure of overall usefulness. This will show the improvements in the performance of the models from one iteration

to the other. In order to do this, a panel of experts that are professional grain commodity traders was approached to participate in the evaluation of the DSS.

Eight (8) experts that are professional and experienced grain commodities traders agreed to voluntarily participate in the evaluation exercise. The panel of experts that agreed to participate are from three different companies that are listed on the Johannesburg Stock Exchange's website as registered to trade in grain commodities in South Africa. This implies that the companies act as brokers on behalf of other stakeholders in the industry such as farmers, market speculators and manufacturers that use grain commodities as raw materials. Moreover, some of these trading companies also buy and sell grain commodities as financial assets on the Johannesburg Stock Exchange. All the experts approached work as grain commodity traders within their organisations over the Johannesburg Stock Exchange and their role requires that they predict future prices.

Out of the panel of experts, two of them provide leadership for other traders in their organisations as General Manager and Chief Operating Officer respectively. Moreover, four of the experts hold a master's degree, three of them hold a bachelor's degree and only one of them has a diploma as the highest qualification. Four of the experts have between 11 – 20 years of trading experience, three of them have between 6 – 10 years of experience and the last one falls between 2 – 5 years. It was also noted that three of the traders manage trade that is between 100,001 – 250,000 metric tons of grain commodities annually and two of the traders manage less than 100,000 metric tons annually. But, two of the traders manage trades that aggregates between 500,001 – 1 million metric tons annually and the other manages trades between 1 – 5 million metric tons annually.

The experts were asked to predict the future prices of the spot prices and the December futures contract of white maize on Johannesburg Stock Exchange for the month of August 2015. All the experts sent in their predictions in the last week of the month of July 2015, which allows them to make predictions based on market trends that are close to the period they predicted. The predictions that were submitted by the

experts were for the spot and December futures contracts prices of white maize from 01 August 2015 till 21 August 2015 during which there are 14 trading days.

While a further evaluation of the model that was implemented in Chapter 6 is the goal of this process, the need for optimisation of the model was also identified. Hence, the evaluation process in the rest of this chapter was implemented as a set of iterative experiments in search for a model with a better performance. As a result, this chapter has an iterative loop with the implementation in Chapter 6. This is based on the DSR process that allows for the iteration of different phases within the research process in the creation of an artefact that solves a practical problem (Johannesson and Perjons, 2012). Vaishnavi and Kuechler (2015) further suggest that application of DSR in the Information Systems related problem, such as the one in this study, often require iterations between the development and the evaluation phases of the research. This will allow for improvement of the artefact by experimentation and learning.

In order to carry out the evaluation process in this chapter, two main iterations between the implementation in Chapter 6 and the actual evaluation in Chapter 7 was carried out. The first iteration was carried out by using the model developed and tested in Chapter 6 as-is to make predictions for the spot and December futures of white maize for August 2015 at the same time in late July 2015 when the experts submitted their prediction. The second iteration was carried out after August 2015 based on the result of the first iterations.

### **7.3.1 Pre-August Iteration**

Based on the equation in Section 5.5.1 that has been adopted for the evaluations in this chapter, new BPNN algorithms for building models for 15 trading days ahead were written. Each of the models was run continuously until 10 different predictions were recorded for each day. Thereafter, the mean value of the 10 predictions captured for each day was taken as the final prediction. Table 7.2 presents the result of the predictions of the panel of experts and that of the models of the proposed DSS of this study that is being implemented and evaluated. The predictions by the experts and that of the DSS were compared with the actual end-of-day spot prices of white maize using MAPE and RMSE as a measurement of accuracy. Table 7.2 also shows the

correlation coefficient of the predictions against the actual prices as an indication of how the predicted prices followed the pattern of the market from day to day.

The result of the experiments presented in Table 7.2 and in a graphical form in Figure 7.2 indicate that the model and the adopted strategy was able to make predictions that are not totally far from the actual prices during the period in review. The MAPE and RMSE measurement of accuracy shows that the predictions from the DSS had lower errors and can be said to have out-performed the predictions made by six experts. However, the predicted prices from the DSS did not follow the trend of the actual value for most of the days. As a result, a correlation coefficient of 0.0323 ( $n=14$ ) was recorded. But prediction from five of the experts had relatively high correlation coefficients. This signifies that five of the experts (A, B, C, E, and F) were better in predicting the direction the market would go during the month of August, although their prediction far deviated from the actual prices. The comparison between the measurement of accuracy of the predictions from the DSS and the experts is portrayed in Figures 7.3. Moreover, Figure 7.4 shows that the DSS did not do very well when the correlation coefficients are compared.

Table 7.2: Comparison between predictions from experts and implemented DSS for spot prices of white maize (Iteration 1)

Day	Expert A	Expert B	Expert C	Expert D	Expert E	Expert F	Expert G	Expert H	DSS	Actual
2015-08-03	3,045	2,950	3,165	3,250	3,250	3,200	3,150	3,190	3,082	3,131
2015-08-04	3,058	2,930	3,140	3,265	3,200	3,225	3,148	3,220	3,188	3,142
2015-08-05	3,035	2,900	3,120	3,280	3,150	3,195	3,155	3,260	3,017	3,138
2015-08-06	3,021	2,930	3,000	3,350	3,080	3,196	3,160	3,230	3,077	3,125
2015-08-07	2,985	2,900	3,130	3,280	3,060	3,190	3,170	3,230	2,981	3,073
2015-08-11	2,985	2,850	2,980	3,240	3,040	3,210	3,190	3,280	3,138	3,124
2015-08-12	2,912	2,820	2,982	3,190	2,980	3,200	3,200	3,330	3,053	3,074
2015-08-13	2,875	2,860	2,985	3,240	3,000	3,180	3,250	3,350	3,159	3,011
2015-08-14	2,901	2,890	2,960	3,260	2,970	3,150	3,240	3,320	3,043	2,987
2015-08-17	2,915	2,850	2,940	3,295	2,940	3,153	3,230	3,330	3,086	2,969
2015-08-18	2,874	2,800	2,950	3,330	2,910	3,180	3,200	3,350	3,148	2,941
2015-08-19	2,877	2,790	2,940	3,290	2,870	3,190	3,205	3,380	2,980	2,960
2015-08-20	2,908	2,750	2,900	3,210	2,890	3,185	3,200	3,400	3,136	3,024
2015-08-21	2,945	2,720	2,880	3,250	2,860	3,187	3,190	3,400	2,927	3,068
<b>MAPE</b>	<b>3.46%</b>	<b>7.11%</b>	<b>2.16%</b>	<b>6.46%</b>	<b>2.26%</b>	<b>4.20%</b>	<b>4.27%</b>	<b>7.50%</b>	<b>2.78%</b>	
<b>RMSE</b>	<b>106.22</b>	<b>212.67</b>	<b>85.78</b>	<b>228.51</b>	<b>87.54</b>	<b>145.40</b>	<b>167.71</b>	<b>280.49</b>	<b>101.79</b>	
<b>R-squared</b>	<b>0.9099</b>	<b>0.5241</b>	<b>0.6454</b>	<b>-0.1554</b>	<b>0.7771</b>	<b>0.7457</b>	<b>-0.7851</b>	<b>-0.7141</b>	<b>0.0323</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>

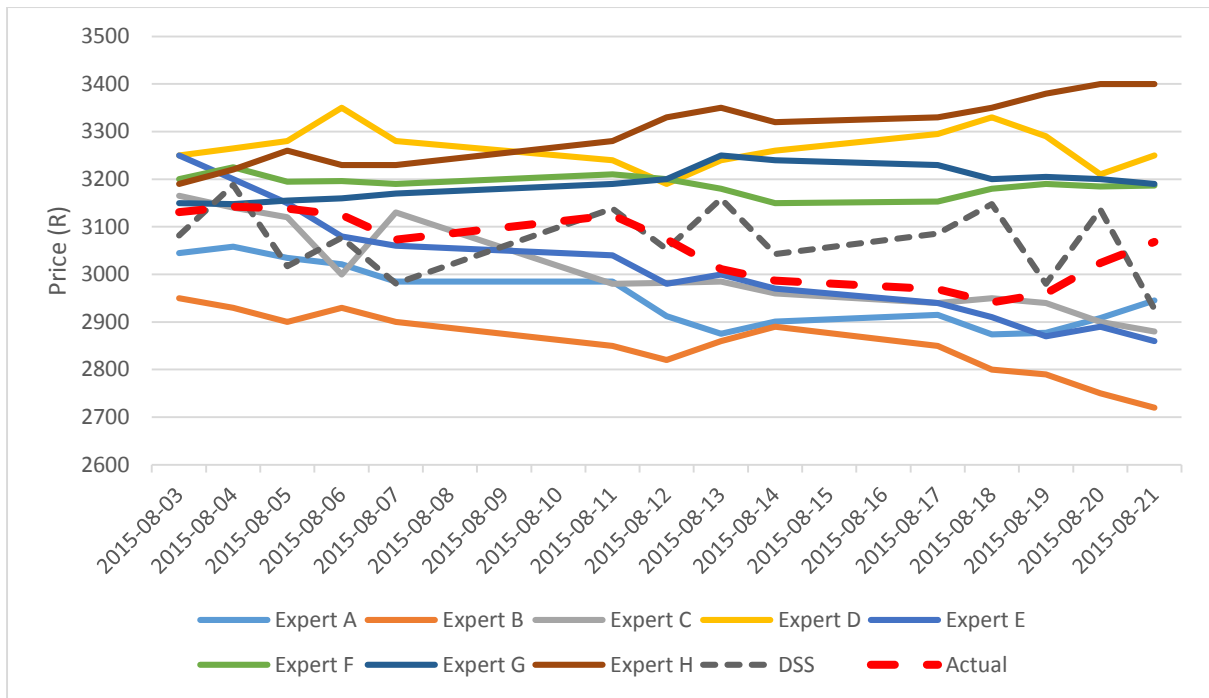


Figure 7.2: Prediction of spot prices of white maize by experts and DSS (Iteration 1)

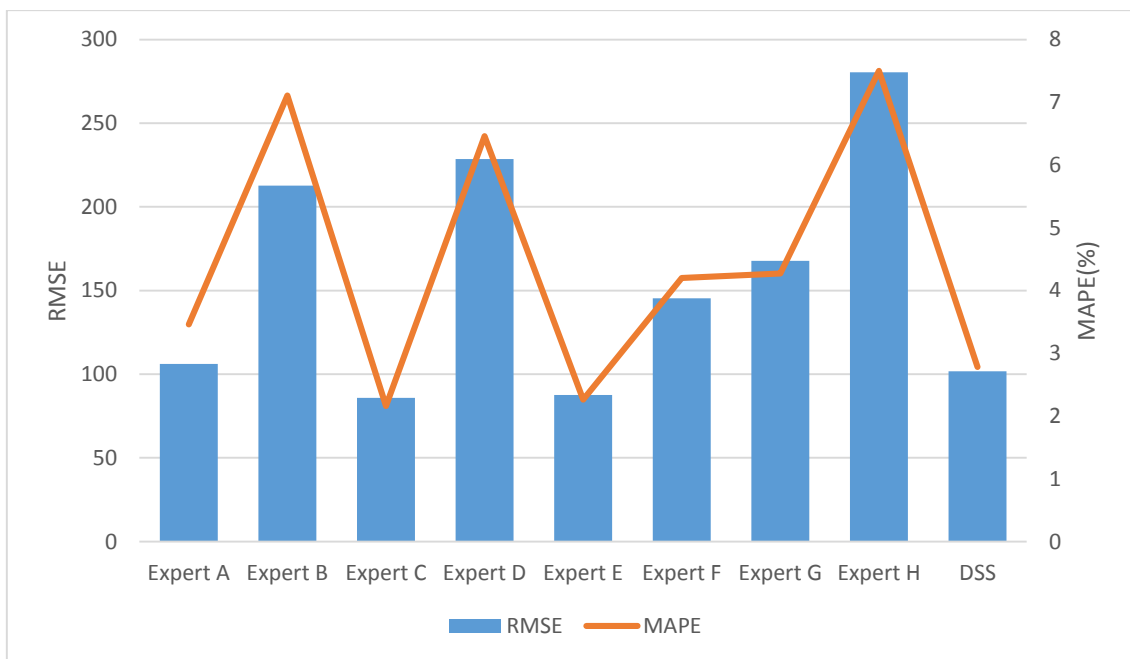


Figure 7.3: Error measurements of experts and DSS predictions for spot prices (Iteration 1)

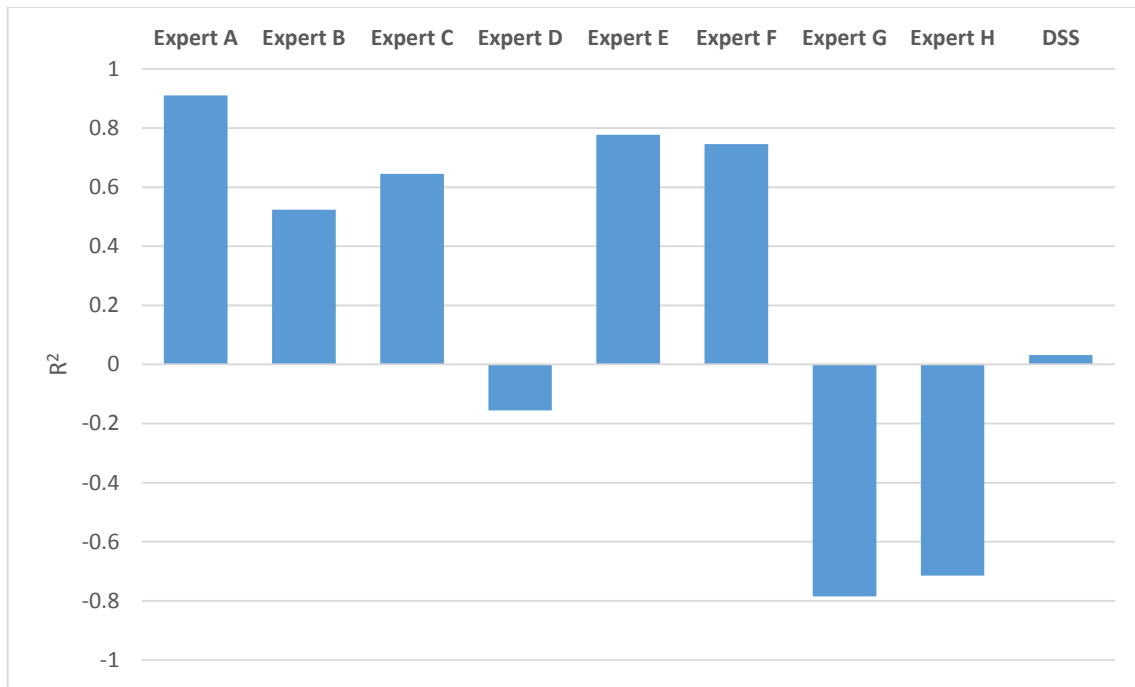


Figure 7.4: Correlation between predictions of spot prices and actual values (Iteration 1)

The futures contract prices for December 2015 were predicted during the month of August 2015 and compared with the predictions of the experts. The data is presented in Table 7.3 while Figure 7.5 shows how the prediction from the DSS fared alongside the predictions of the experts. Moreover, Figure 7.6 shows a graphical comparison of the MAPE and RMSE as measurement of accuracy. Figure 7.7 shows a comparison of the correlation coefficients between the predicted prices and the actual prices for all the predictions. The results portray an unreliable prediction from the DSS. Figure 7.5 indicated that the predictions were only close to the actual value recorded for only 2 or 3 days and predictions for the other days deviated from the actual values progressively. This is further emphasised in the measure of accuracy used with the RMSE for the DSS being 309.30 second only to the predictions of Expert H which is the highest (329.18). It is, however, noteworthy that it was not only the predictions of the DSS that deviated progressively from the actual values, Figure 7.5 shows that predictions by Experts D, G and H followed the same trend suggesting a pattern.

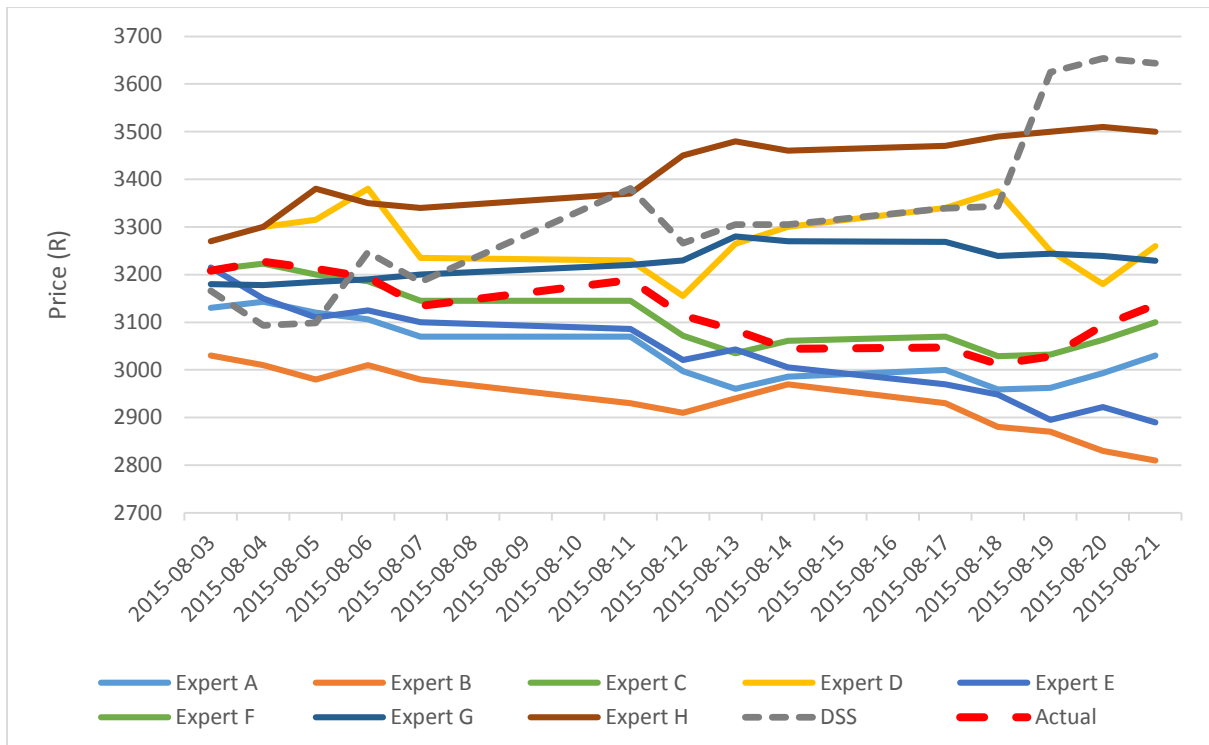


Figure 7.5: Prediction of December futures contract prices of white maize by experts and DSS (Iteration 1)



Table 7.3: Comparison between predictions from experts and implemented DSS for December futures contract prices

Day	Expert A	Expert B	Expert C	Expert D	Expert E	Expert G	Expert H	DSS	Actual
2015-08-03	3,130	3,030	3,210	3,275	3,215	3,180	3,270	3,166	3,208
2015-08-04	3,143	3,010	3,223	3,300	3,150	3,178	3,300	3,093	3,227
2015-08-05	3,120	2,980	3,200	3,315	3,110	3,185	3,380	3,099	3,212
2015-08-06	3,106	3,010	3,186	3,380	3,125	3,190	3,350	3,247	3,194
2015-08-07	3,070	2,980	3,145	3,235	3,100	3,200	3,340	3,185	3,134
2015-08-11	3,070	2,930	3,145	3,230	3,086	3,220	3,370	3,381	3,189
2015-08-12	2,997	2,910	3,072	3,155	3,021	3,230	3,450	3,266	3,114
2015-08-13	2,960	2,940	3,035	3,265	3,043	3,280	3,480	3,305	3,083
2015-08-14	2,986	2,970	3,061	3,300	3,005	3,270	3,460	3,305	3,044
2015-08-17	3,000	2,930	3,070	3,340	2,970	3,269	3,470	3,339	3,047
2015-08-18	2,959	2,880	3,029	3,375	2,948	3,239	3,490	3,343	3,012
2015-08-19	2,962	2,870	3,032	3,250	2,895	3,244	3,500	3,625	3,028
2015-08-20	2,993	2,830	3,063	3,180	2,922	3,239	3,510	3,654	3,093
2015-08-21	3,030	2,810	3,100	3,260	2,890	3,229	3,500	3,644	3,136
<b>MAPE</b>	<b>2.81%</b>	<b>6.45%</b>	<b>0.69%</b>	<b>4.63%</b>	<b>3.00%</b>	<b>3.63%</b>	<b>8.58%</b>	<b>7.28%</b>	
<b>RMSE</b>	<b>88.80</b>	<b>199.34</b>	<b>26.28</b>	<b>180.69</b>	<b>107.62</b>	<b>143.93</b>	<b>329.18</b>	<b>309.30</b>	
<b>R-squared</b>	<b>0.9401</b>	<b>0.5610</b>	<b>0.9421</b>	<b>-0.0206</b>	<b>0.7801</b>	<b>-0.8479</b>	<b>-0.8200</b>	<b>-0.5245</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>

Figure 7.6 shows that Expert C had the best prediction of the December futures contract of white maize for the month of August 2015 followed by that of Expert A and then Expert E. This is also reflected in the correlation analysis shown in Figure 7.7. Predictions by Expert C had a very strong 0.9421 (n=14) correlation with the actual values that were recorded. This is closely followed by the predictions of Expert A, with 0.9401 (n=14) correlation and then Expert E with 0.7801 (n=14) correlation coefficient against the actual values that were recorded. On the other hand, the predictions of the DSS and those of Experts D, G and H had negative correlation coefficients -0.5245, -0.0206, -0.8479 and -0.8200 respectively. These comparisons indicate that it possible to predict the December futures contract as was seen in Chapter 6. However, the results also show that the predictions made by the DSS in the first iteration are technically faulty and will not be useful for decision making if implemented in a DSS.

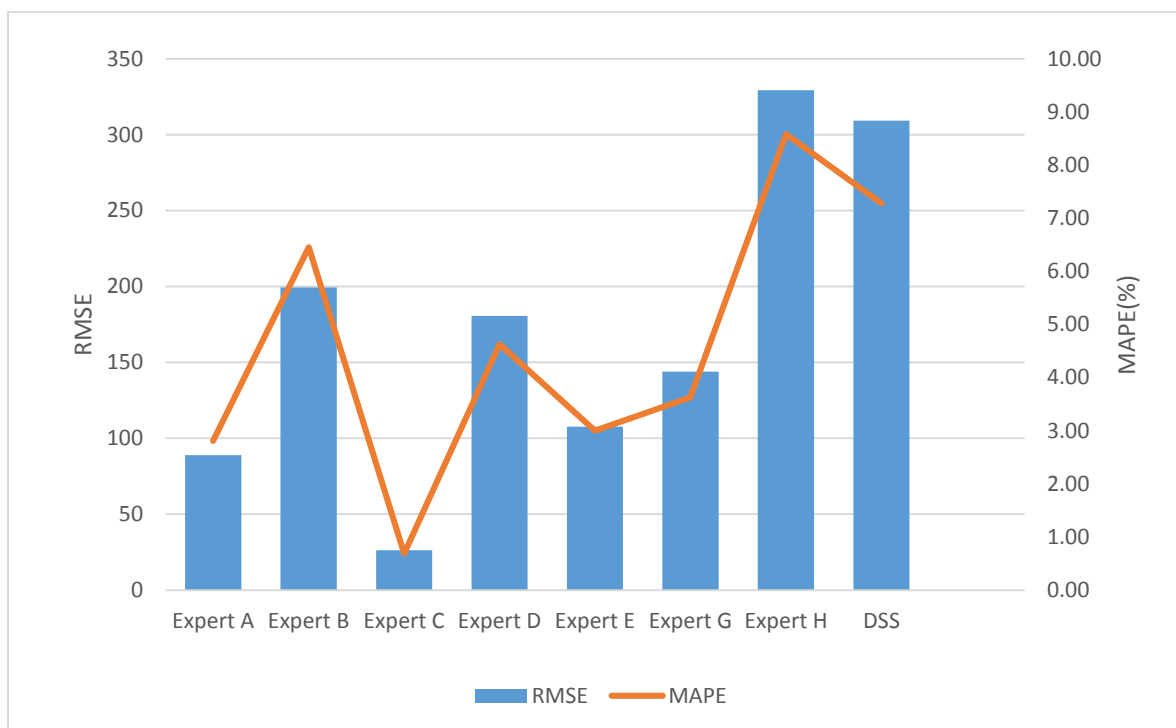


Figure 7.6: Error measurements of experts and DSS predictions for spot prices (Iteration 1)

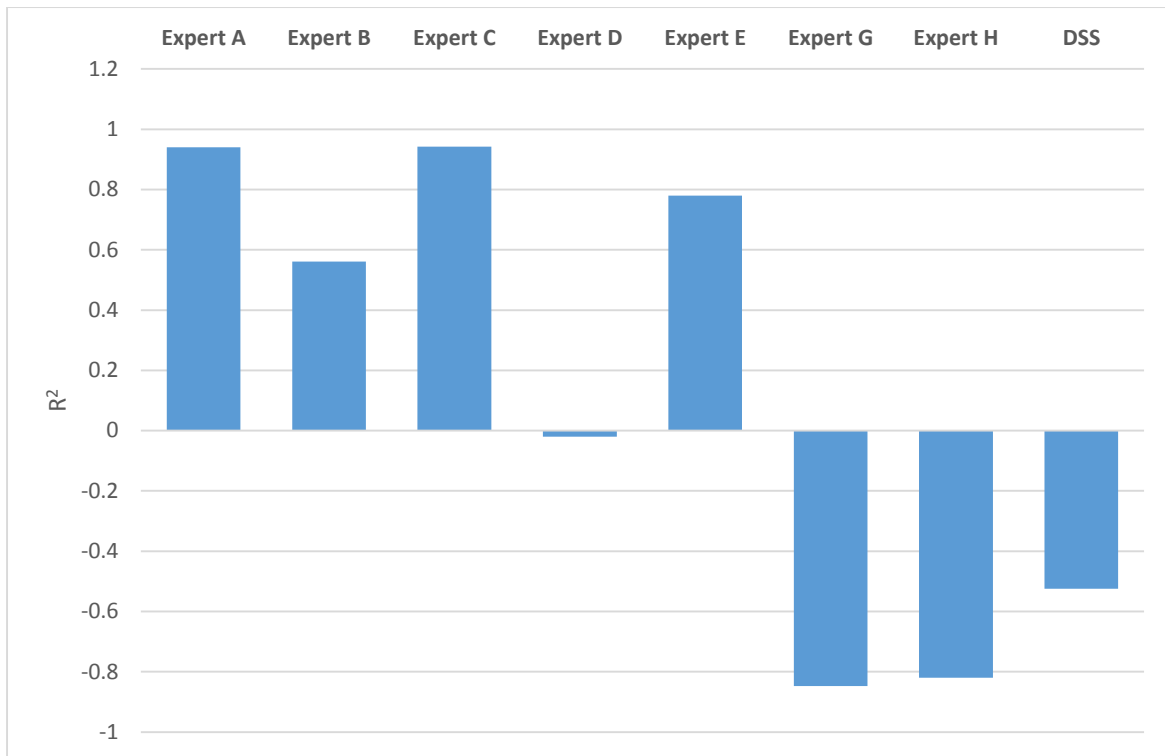


Figure 7.7: Correlation between predictions of spot prices and actual values (Iteration 1)

The experiments, carried out in Iteration 1, were carried out using the BPNN designed in Chapter 6 as the modelling component of the DSS. The data between the periods of 01 January 2010 to 31 December 2014 was used to train the BPNN model for predicting the spot prices of white maize. The data from 01 January 2012 till 31 December 2014 was used to train the model for predicting the December futures contract of white maize. The outcome of the evaluation in Iteration 1 indicates that the predictions from the DSS could still be optimised further to increase performance. The predictions by the DSS for the spot prices out-performed those of most of the experts technically, but the trend is a still major concern. However, the prediction of the DSS for the December futures contract is completely unacceptable when compared to the actual values and the predictions of the experts.

### 7.3.2 Post-August Iteration

At the end of the month of August, the result of the predictions by the experts and the one from the proposed DSS were compared to the actual end-of-day spot and December futures prices of white maize as traded on Johannesburg Stock Exchange. In order to improve the accuracy and performance of the DSS, further literature study was carried out and Chapter 5 was updated by adding Section 5.7 on the real-time learning with Neural Networks. The literature study suggested that the availability and use of new datasets as they become available in real-time, could improve the performance of a Neural Network. Once again, the advantage that the Big Data concepts bring into the design of a DSS for trading grain commodities were highlighted.

Based on the suggestions of Ruta (2014) on the use of Big Data for real-time learning for financial assets trading, three categories of experiments were carried out in Iteration 2 to identify the impact of using different subsets of data as the input data for training the model and making predictions. The same process of making 10 predictions for each day and finding the mean was also adopted in this set of experiments. Those adopted in the previous iteration were also used in Iteration 2. Three categories of experiments were carried out using different subsets of data for the prediction of spot and December futures contract prices of white maize for the first 14 trading days in August 2015. Category A was set up to use datasets from 01 January 2010 till 15 July 2015, while Category B made use of datasets from 01 January 2010 till 31 July 2015. Category C experiments were set up to use a rolling subset of data as the input for the training and the predictions shown in Table 7.4.

The category C experiments made use of datasets between 01 January 2010 and 15 July 2015 as the training set for building the model for the first trading day in the month of August. A prediction was made by using data from 16 July 2015 and 31 July 2015 as the input data. However, for Category C, new daily data was included in the input data for retraining the model at the end of each day. This was also applied to the input data for the predictions, by adding data from the previous trading day as shown in Table 7.4.

Table 7.4: Tables showing the input datasets used in category C modelling

Training		Prediction		
Start	End	Start	End	Results for
2010-01-01	2015-07-15	2015-07-16	2015-07-31	2015-08-03
2010-01-01	2015-07-18	2015-07-19	2015-08-03	2015-08-04
2010-01-01	2015-07-19	2015-07-20	2015-08-04	2015-08-05
2010-01-01	2015-07-20	2015-07-21	2015-08-05	2015-08-06
2010-01-01	2015-07-21	2015-07-22	2015-08-06	2015-08-07
2010-01-01	2015-07-22	2015-07-23	2015-08-07	2015-08-10
2010-01-01	2015-07-25	2015-07-26	2015-08-10	2015-08-11
2010-01-01	2015-07-26	2015-07-27	2015-08-11	2015-08-12
2010-01-01	2015-07-27	2015-07-28	2015-08-12	2015-08-13
2010-01-01	2015-07-28	2015-07-29	2015-08-13	2015-08-14
2010-01-01	2015-07-29	2015-07-30	2015-08-14	2015-08-17
2010-01-01	2015-08-01	2015-08-02	2015-08-17	2015-08-18
2010-01-01	2015-08-02	2015-08-03	2015-08-18	2015-08-19
2010-01-01	2015-08-03	2015-08-04	2015-08-19	2015-08-20

A summary of the prediction for each day and the actual spot prices of white maize in South Africa is presented in Table 7.5 and in a graphical view in Figure 7.8. The result for each day and each category is a mean of 10 predictions.

Table 7.5: Actual and predicted spot prices of white maize

	<b>Category A</b>	<b>Category B</b>	<b>Category C</b>	<b>Actual</b>
<b>2015-08-03</b>	3,153	3,171	3,161	<b>3,131</b>
<b>2015-08-04</b>	3,117	3,164	3,162	<b>3,142</b>
<b>2015-08-05</b>	3,172	3,175	3,094	<b>3,138</b>
<b>2015-08-06</b>	3,131	3,144	3,093	<b>3,125</b>
<b>2015-08-07</b>	3,138	3,167	3,114	<b>3,073</b>
<b>2015-08-11</b>	3,169	3,181	3,075	<b>3,124</b>
<b>2015-08-12</b>	3,118	3,155	3,075	<b>3,074</b>
<b>2015-08-13</b>	3,125	3,150	3,043	<b>3,011</b>
<b>2015-08-14</b>	3,132	3,156	3,059	<b>2,987</b>
<b>2015-08-17</b>	3,119	3,154	3,013	<b>2,969</b>
<b>2015-08-18</b>	3,112	3,144	3,008	<b>2,941</b>
<b>2015-08-19</b>	3,128	3,122	3,014	<b>2,960</b>
<b>2015-08-20</b>	3,136	3,152	3,000	<b>3,024</b>
<b>2015-08-21</b>	3,108	3,127	2,966	<b>3,068</b>
<b>MAPE</b>	<b>2.61%</b>	<b>3.16%</b>	<b>1.43%</b>	
<b>RMSE</b>	<b>99.51</b>	<b>117.09</b>	<b>49.91</b>	
<b>R-squared</b>	<b>0.5169</b>	<b>0.5885</b>	<b>0.7153</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	

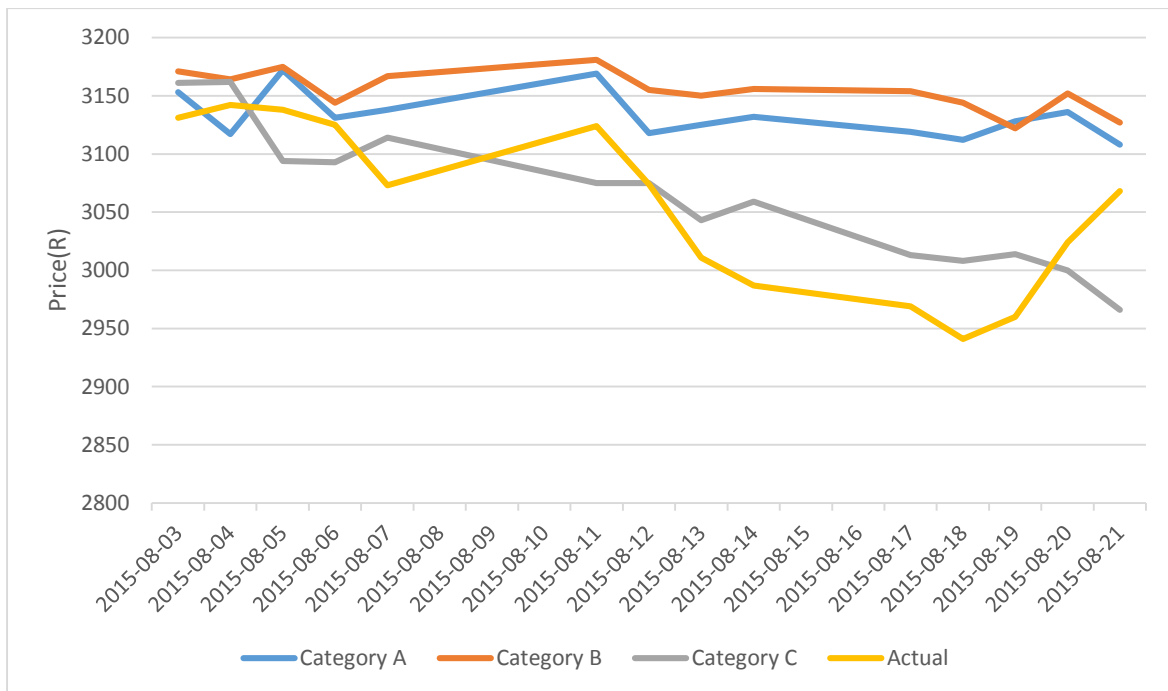


Figure 7.8: Graph showing actual and predicted spot prices of white maize with different models

The accuracy measurement statistics in Table 7.5 show that the models used in Category C performed better than the others with 1.43% of MAPE, 49.91 RMSE and a 0.7153 (n=14) correlation between the actual and the predicted prices. This is only followed by the predictions made with Category A models with almost double the error statistics for both MAPE and RMSE. This is also shown in the degree with which the prediction follows the trend of the actual prices with Category C having about 0.2 more correlation between the predicted and the actual values. The results indicate that re-training the BPNN models periodically as new data becomes available and including new data in both the input data for training the model, as well as input dataset for predictions, can improve the accuracy of the predictions.

The process implemented for the spot prices above was also implemented for the December futures contract prices to determine the optimal sub-sets of data for training the model and for making predictions. Data from 01 January 2012 till 15 July 2015 was used for Category A and the dataset from 10 January 2012 till 31 July 2015 was used as the input set for Category B. Following the structure of the model denoted in

equation (4), a rolling data series as displayed in Table 7.5 was used for the Category C experiments of the December futures contract training and predictions.

Table 7.6: Actual and predicted December futures prices of white maize

	<b>Category A</b>	<b>Category B</b>	<b>Category C</b>	<b>Actual</b>
<b>2015-08-03</b>	3,185	3,215	3,190	<b>3,208</b>
<b>2015-08-04</b>	3,186	3,208	3,158	<b>3,227</b>
<b>2015-08-05</b>	3,184	3,206	3,154	<b>3,212</b>
<b>2015-08-06</b>	3,186	3,214	3,174	<b>3,194</b>
<b>2015-08-07</b>	3,185	2,931	3,139	<b>3,134</b>
<b>2015-08-11</b>	3,198	3,221	3,102	<b>3,189</b>
<b>2015-08-12</b>	3,193	3,211	3,117	<b>3,114</b>
<b>2015-08-13</b>	3,202	3,211	3,152	<b>3,083</b>
<b>2015-08-14</b>	3,191	3,211	3,123	<b>3,044</b>
<b>2015-08-17</b>	3,174	3,217	3,080	<b>3,047</b>
<b>2015-08-18</b>	3,191	3,225	3,045	<b>3,012</b>
<b>2015-08-19</b>	3,179	3,213	3,054	<b>3,028</b>
<b>2015-08-20</b>	3,173	3,225	2,997	<b>3,093</b>
<b>2015-08-21</b>	3,198	3,208	3,010	<b>3,136</b>
<b>MAPE</b>	<b>2.47%</b>	<b>2.97%</b>	<b>1.67%</b>	
<b>RMSE</b>	<b>98.98</b>	<b>117.83</b>	<b>62.91</b>	
<b>R-squared</b>	<b>0.1364</b>	<b>-0.3268</b>	<b>0.5803</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	

The implementation of the Category C model implementation for the December futures contract prices also followed the strategy used for the Category C implementation for the spot price predictions, except that the input datasets started from 01 January 2012. As with the spot prices, models for each of the categories and for each day were run iteratively until 10 satisfactory models were identified and used to make predictions. These predictions were averaged as the prediction for the day. The summary of the results for the December futures contract price predictions of white maize is presented in Table 7.6.



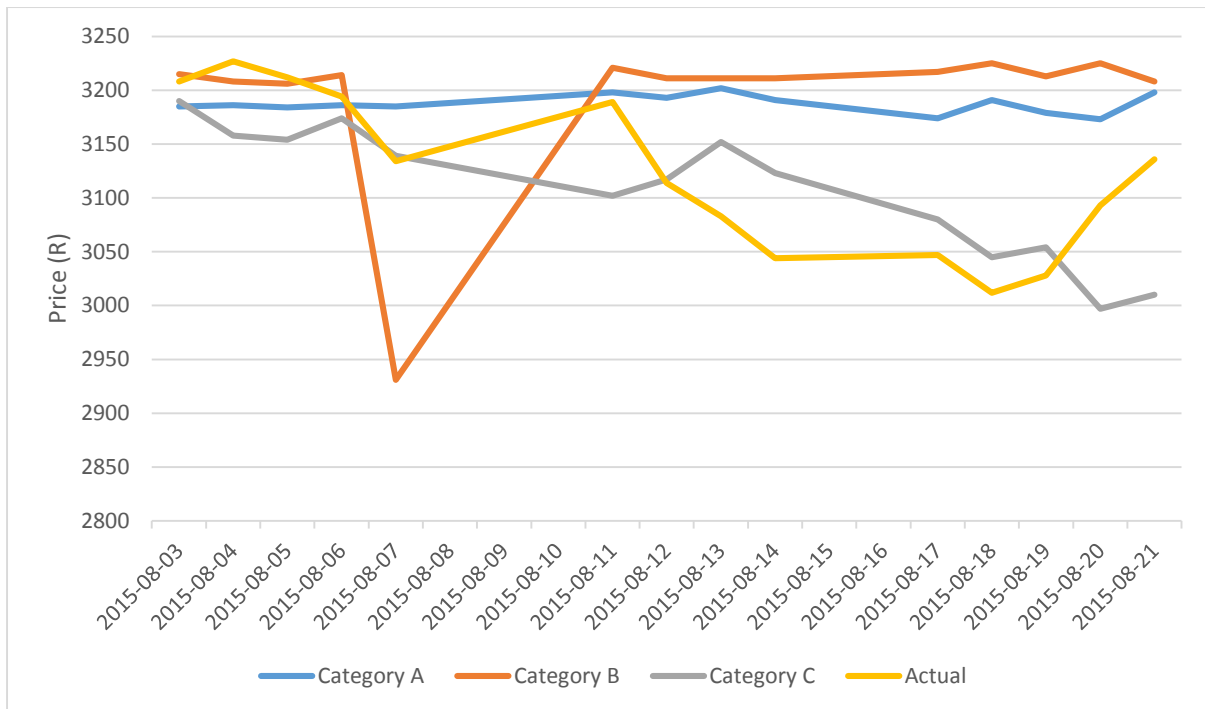


Figure 7.9: Graph showing actual and predicted December futures prices of white maize with different models

The results indicate that the category C experiments indicate that the use of a rolling input datasets for training the model and making predictions improve the accuracy of the DSS. Table 7.6 shows that the MAPE and RMSE statistics for the category C experiments which were 1.67% and 62.91 respectively were almost half those of Category A and B experiments. Moreover, the correlation coefficient of the predictions against the actual values that were recorded was 0.5803 (n=14) for category C experiments compared to 0.1364 (n=14) and -0.3268 (n=14) for category A and B experiments. The results show a significant improvement in accuracy and ability to follow the trend for category C experiments for both the spot and the December futures contract prices, when compared with the results of Iteration 1 experiments.

### 7.3.3 Comparison of Iteration 2 results and predictions by panel of experts

A further evaluation of the usefulness of the proposed framework was carried out as part of iteration by comparing the results of Category C models in Section 7.3.2 and predictions by the panel of experts. The predictions of the spot and December future prices of white maize in South Africa that resulted from Category C experiments in

Section 7.3.2 were compared with predictions made by each of the expert traders. The predictions made by the experts and those from the implemented DSS were compared to the actual end-of-day prices of the spot and December futures prices of white maize in South Africa for the first 3 weeks of August 2015. This was achieved by calculating the Root Mean Square (RMSE) and Mean Absolute Percentage Error (MAPE) statistics between each of the predictions and the actual prices. Tables 7.7 and 7.8 present the predictions for the spot and the December futures price predictions respectively.

The results in Table 7.7 indicate that the predictions from the DSS had fewer error factors than the predictions from all the experts. However, predictions by Experts A, E, F, G and H had better correlations that are greater than 0.7153 (n=14) recorded for the predictions by the DSS. The 0.7153 (n=14) correlation between the prediction of the DSS and the actual spot prices for the period in review is relatively high, but the predictions by Experts A, E, F, G and H seem to follow the trend of the actual values more closely. However, the predictions from the implemented DSS seems to be less deviated from the actual spot prices as shown in Figure 7.10. This is a relative improvement in how the model follows the trend in the actual prices of the spot prices of white maize when compared to a correlation coefficient of 0.0323 (n=14) that was recorded for the predictions of the spot price of white maize in Iteration 1. Moreover, the measurement of accuracy statistics for Iteration 2 shows a significant improvement with MAPE = 1.44% and RMSE = 49.91. This is reflected graphically in Figure 7.10 showing that the price predicted by the DSS is about the closest to the actual value recorded, although there is still room for improvements.

Table 7.7: Comparison between predictions from experts and implemented DSS for spot prices of white maize

Day	Expert A	Expert B	Expert C	Expert D	Expert E	Expert F	Expert G	Expert H	DSS	Actual
2015-08-03	3,045	2,950	3,165	3,250	3,250	3,200	3,150	3,190	3,161	3,131
2015-08-04	3,058	2,930	3,140	3,265	3,200	3,225	3,148	3,220	3,162	3,142
2015-08-05	3,035	2,900	3,120	3,280	3,150	3,195	3,155	3,260	3,094	3,138
2015-08-06	3,021	2,930	3,000	3,350	3,080	3,196	3,160	3,230	3,093	3,125
2015-08-07	2,985	2,900	3,130	3,280	3,060	3,190	3,170	3,230	3,114	3,073
2015-08-11	2,985	2,850	2,980	3,240	3,040	3,210	3,190	3,280	3,075	3,124
2015-08-12	2,912	2,820	2,982	3,190	2,980	3,200	3,200	3,330	3,075	3,074
2015-08-13	2,875	2,860	2,985	3,240	3,000	3,180	3,250	3,350	3,043	3,011
2015-08-14	2,901	2,890	2,960	3,260	2,970	3,150	3,240	3,320	3,059	2,987
2015-08-17	2,915	2,850	2,940	3,295	2,940	3,153	3,230	3,330	3,013	2,969
2015-08-18	2,874	2,800	2,950	3,330	2,910	3,180	3,200	3,350	3,008	2,941
2015-08-19	2,877	2,790	2,940	3,290	2,870	3,190	3,205	3,380	3,014	2,960
2015-08-20	2,908	2,750	2,900	3,210	2,890	3,185	3,200	3,400	3,000	3,024
2015-08-21	2,945	2,720	2,880	3,250	2,860	3,187	3,190	3,400	2,966	3,068
<b>MAPE</b>	<b>3.46%</b>	<b>7.11%</b>	<b>2.16%</b>	<b>6.46%</b>	<b>2.26%</b>	<b>4.20%</b>	<b>4.27%</b>	<b>7.50%</b>	<b>1.44%</b>	
<b>RMSE</b>	<b>106.22</b>	<b>212.67</b>	<b>85.78</b>	<b>228.51</b>	<b>87.54</b>	<b>145.40</b>	<b>167.71</b>	<b>280.49</b>	<b>49.91</b>	
<b>R-squared</b>	<b>0.9099</b>	<b>0.5241</b>	<b>0.6454</b>	<b>-0.1554</b>	<b>0.7771</b>	<b>0.7457</b>	<b>-0.7851</b>	<b>-0.7141</b>	<b>0.7153</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>

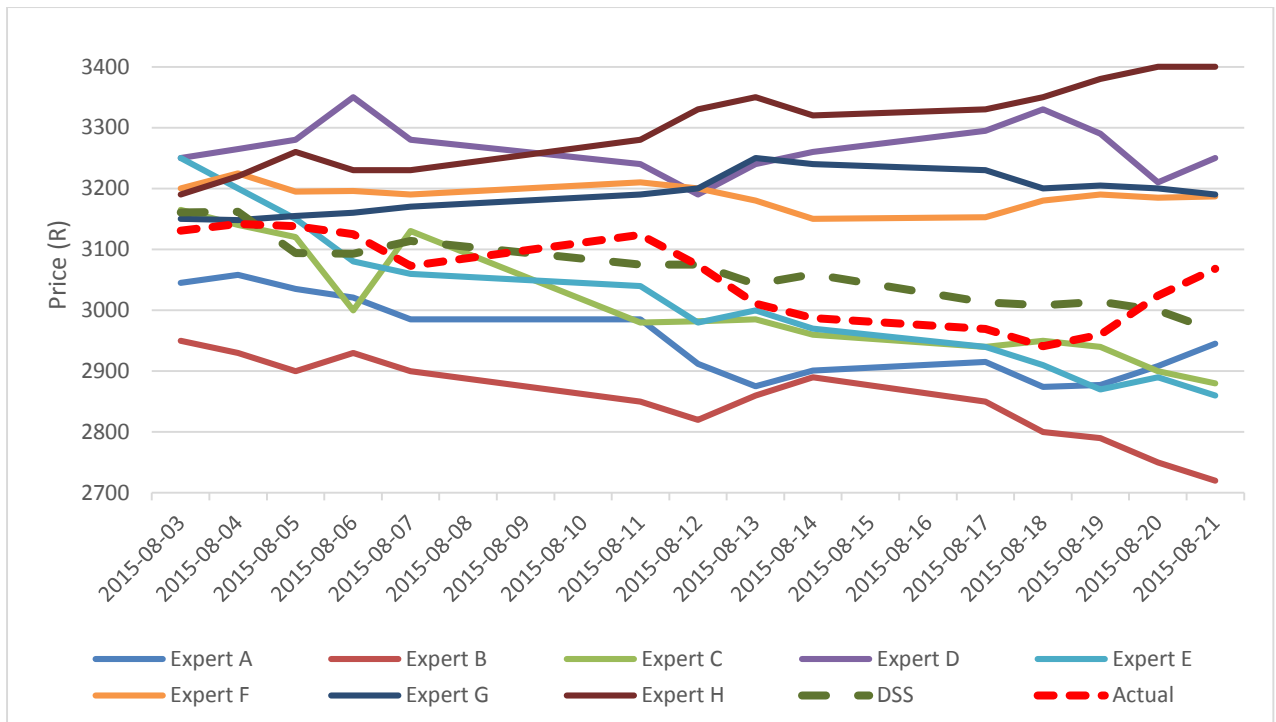


Figure 7.10: Prediction of spot prices of white maize by experts and DSS

Figures 7.11 and 7.12 present a graphical view of the error statistics and correlation coefficients respectively which compare the performance of the DSS with predictions by experts. The graphs show that the BPNN model implemented in the DSS performed relatively well with minimum deviation from the actual prices in terms of value. Results shown in Figure 7.12 indicate that the trend of the predictions from the DSS compared to the actual values, which is measured using the correlation analysis, is relatively high. However, three of the experts had predictions with trends that are much closer to those of the actual prices recorded and the prediction of Expert A had the strongest correlation at 0.9099 (n=14). The increase in performance that was recorded by using a rolling subset of the data as more data becomes available is highlighted in Figure 7.11, showing that the error statistics for the DSS is minimal when compared to the error statistics of the predictions by all the experts.

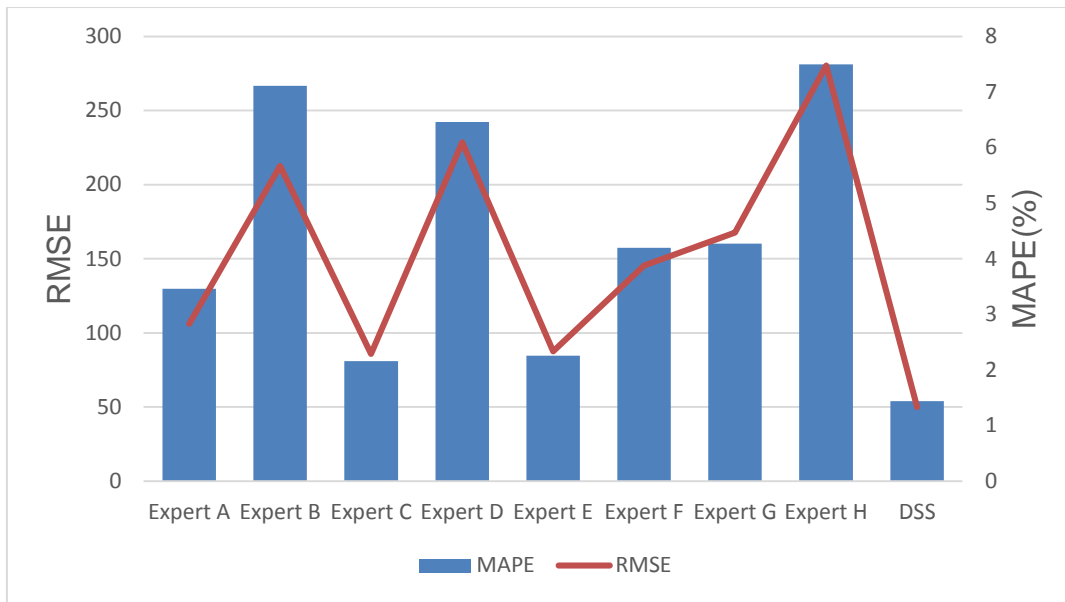


Figure 7.11: Error measurements of experts and DSS predictions for spot prices

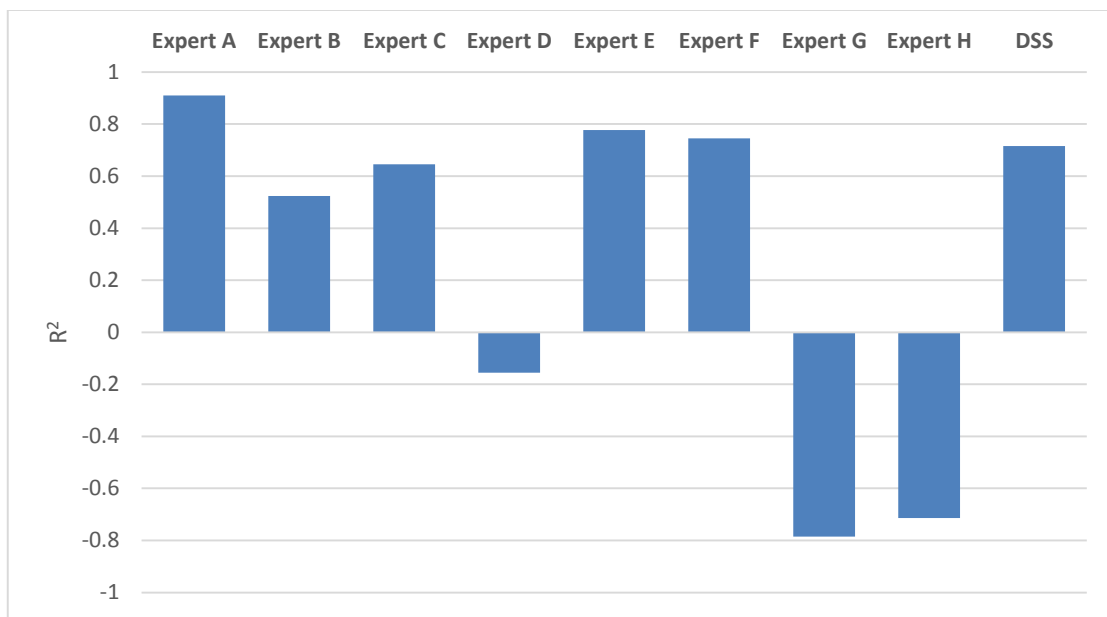


Figure 7.12: Correlation between predictions of spot prices and actual values

The same procedure used in Iteration 2 for evaluating the technical performance and overall usefulness of the predictions for the spot prices of white maize in South Africa was followed for that of the December futures contract prices. It should be noted that for the results in both Iteration 1 and 2 for the December futures contract, there is no data for Expert F because the trader opted not to make predictions for future contract

prices, making the number of expert predictions compared to the predictions using the implemented DSS to be seven.

The result of the evaluation showed that Expert C had predictions with the minimum error measurements and the highest correlation coefficient when the predictions are compared to the actual December futures contract prices. The predictions by the implemented DSS have the second minimum RMSE and MAPE after those of Expert C. This indicates that the predictions by the DSS out-perform the predictions by the other six experts when error value of deviation from the actual prices is considered. Figure 7.13 presents a line graph showing how the predictions by experts and the DSS compare with the actual prices of white maize during the same period, and Table 7.8 shows the detailed predictions by the seven experts, the DSS and the actual prices recorded.

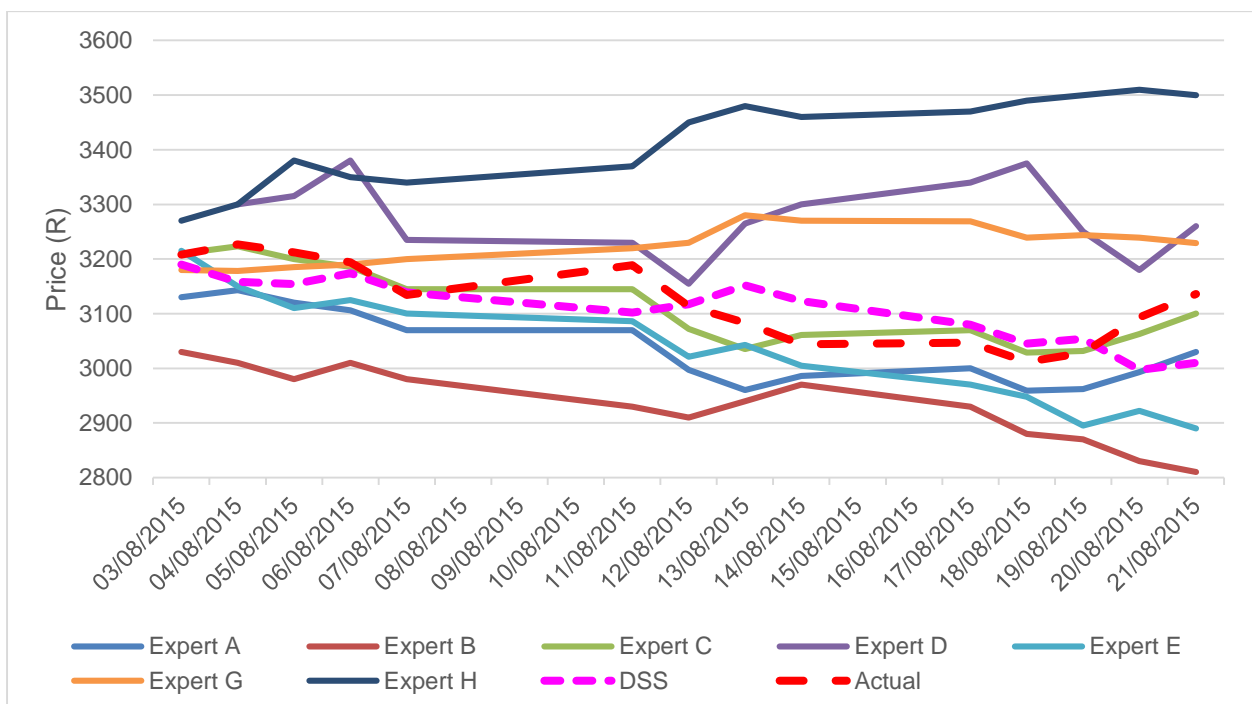


Figure 7.13: Prediction of December futures prices of white maize by experts and DSS

Table 7.8: Comparison between predictions from experts and implemented DSS for December futures contract prices

Day	Expert A	Expert B	Expert C	Expert D	Expert E	Expert G	Expert H	DSS	Actual
2015-08-03	3,130	3,030	3,210	3,275	3,215	3,180	3,270	3,190	3,208
2015-08-04	3,143	3,010	3,223	3,300	3,150	3,178	3,300	3,158	3,227
2015-08-05	3,120	2,980	3,200	3,315	3,110	3,185	3,380	3,154	3,212
2015-08-06	3,106	3,010	3,186	3,380	3,125	3,190	3,350	3,174	3,194
2015-08-07	3,070	2,980	3,145	3,235	3,100	3,200	3,340	3,139	3,134
2015-08-11	3,070	2,930	3,145	3,230	3,086	3,220	3,370	3,102	3,189
2015-08-12	2,997	2,910	3,072	3,155	3,021	3,230	3,450	3,117	3,114
2015-08-13	2,960	2,940	3,035	3,265	3,043	3,280	3,480	3,152	3,083
2015-08-14	2,986	2,970	3,061	3,300	3,005	3,270	3,460	3,123	3,044
2015-08-17	3,000	2,930	3,070	3,340	2,970	3,269	3,470	3,080	3,047
2015-08-18	2,959	2,880	3,029	3,375	2,948	3,239	3,490	3,045	3,012
2015-08-19	2,962	2,870	3,032	3,250	2,895	3,244	3,500	3,054	3,028
2015-08-20	2,993	2,830	3,063	3,180	2,922	3,239	3,510	2,997	3,093
2015-08-21	3,030	2,810	3,100	3,260	2,890	3,229	3,500	3,010	3,136
<b>MAPE</b>	<b>2.81%</b>	<b>6.45%</b>	<b>0.69%</b>	<b>4.63%</b>	<b>3.00%</b>	<b>3.63%</b>	<b>8.58%</b>	<b>1.67%</b>	
<b>RMSE</b>	<b>88.80</b>	<b>199.34</b>	<b>26.28</b>	<b>180.69</b>	<b>107.62</b>	<b>143.93</b>	<b>329.18</b>	<b>62.91</b>	
<b>R-squared</b>	<b>0.9401</b>	<b>0.5610</b>	<b>0.9421</b>	<b>-0.0206</b>	<b>0.7801</b>	<b>-0.8479</b>	<b>-0.8200</b>	<b>0.5799</b>	
	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>	<b>(n=14)</b>

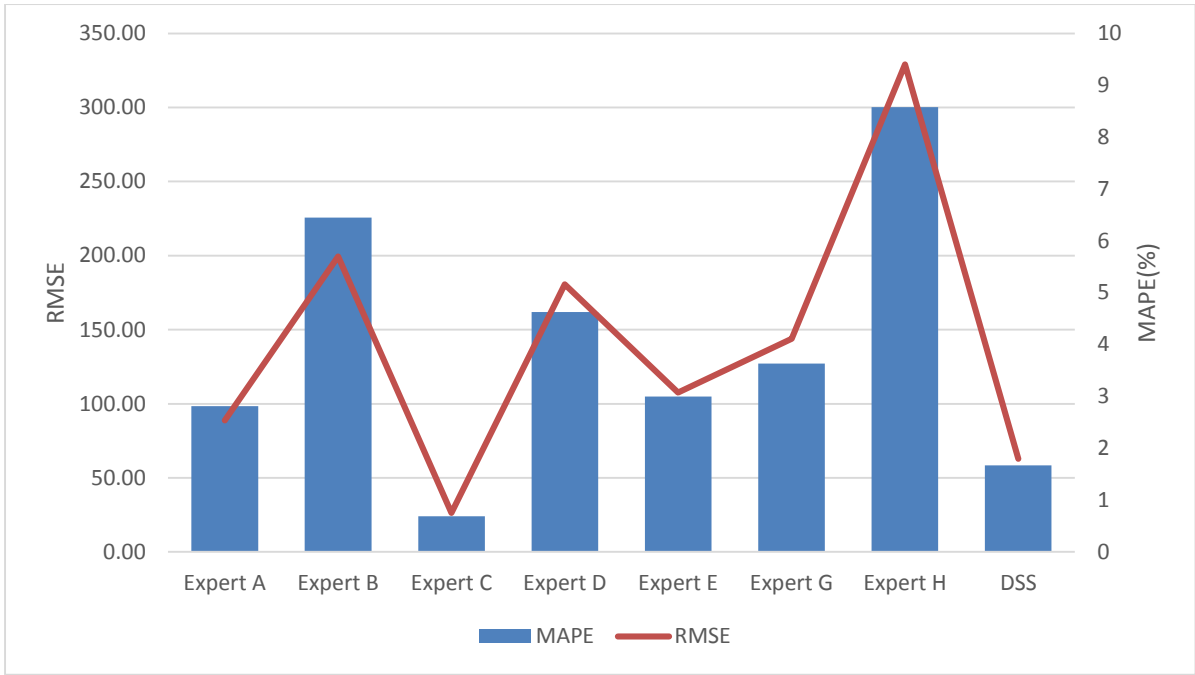


Figure 7.14: Error measurements of experts and DSS predictions for December futures prices

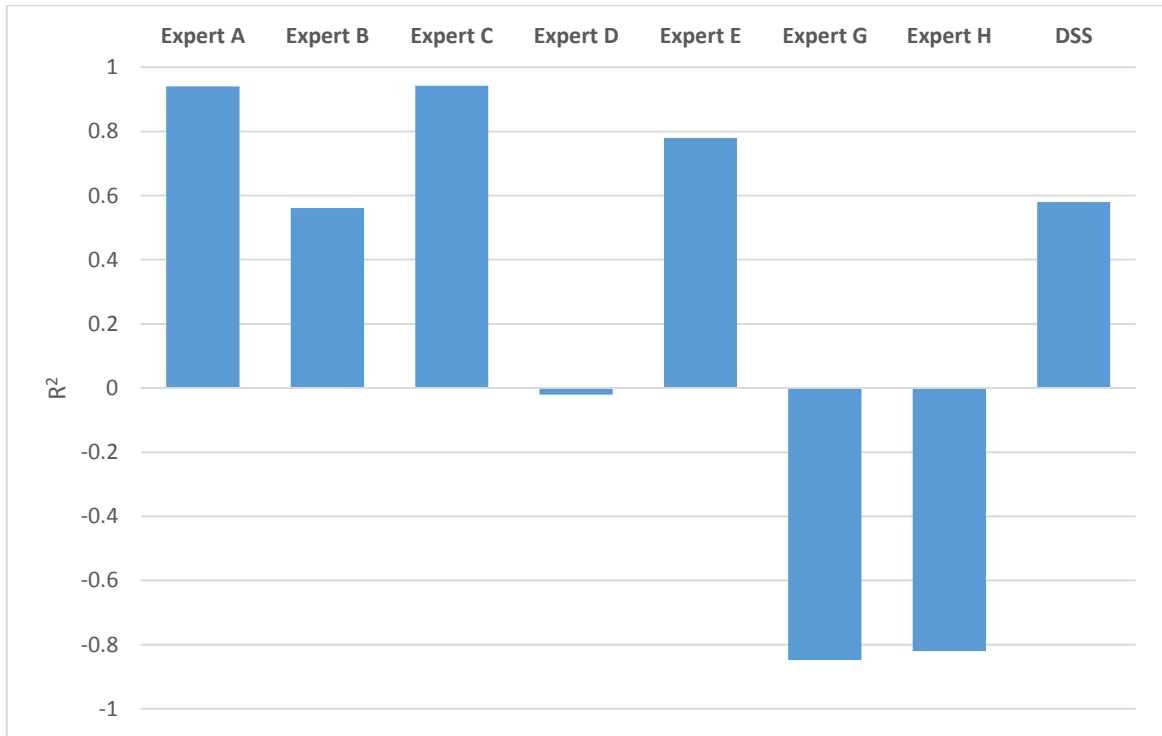


Figure 7.15: Correlation between predicted December futures contract prices and actual values



The outcome of the investigation into the usefulness of the DSS for predicting the December future contract prices as shown in Table 7.8 and Figures 7.13 and 7.14 provides an indication that the prediction made by the DSS is arguably better than the predictions of seven of the eight Experts that participated. The prediction from the DSS had a 1.67% Mean Absolute Percentage Error, while the prediction from Expert C with the minimum MAPE was 0.69%. The MAPE of the prediction from the other experts ranges from 2.81 – 8.58%. The Root Mean Square error for the prediction using the DSS is 62.91, whereas prediction with the minimum RMSE that is from Expert C is 26.28; however, the prediction with the highest RMSE was 329.18. These error measurement statistics suggest that the BPNN model implemented in the DSS performed reasonably well. However, the prediction from the DSS shows a 0.5799 (n=14) correlation with the actual values which could be interpreted as how close the model is in determining the actual trend. The correlation coefficient of the predictions by Experts A, B, C and E were still better than that of the DSS, but there is a noticeable improvement from that of the Iteration 1, which was negative.

Overall, the results from the evaluation conducted in this section have shown that the implementation of the DSS based on the proposed framework of this study is technically capable. It can be deduced that the modelling component of the proposed framework can be used to make predictions for both the spot and the futures contract prices of white maize that are reasonably accurate. The iterative experiments carried out during the evaluation process also indicate that the implementation of the proposed framework would be found useful. The evaluation process showed that the framework proposed in this study can be implemented to develop a DSS to support decision making about trading grain commodities in South Africa. However, the evaluation process also highlighted the fact that the availability of data in real-time and the ability to make use of such data in a DSS can improve the technical abilities and the overall usefulness of such DSS.

The DSS performed very well in making predictions from the modelling component and even better in some cases when compared to predictions made by experts with several years of experience in making such predictions. This indicates that the DSS will be of immense value to groups such as the farmers without such experience. The proposed framework can be used to develop a DSS that such a category of stakeholders with little experience of industry stakeholders can use to make decisions without having to completely depend on the experts.

## **7.4 Conclusion**

The purpose of this chapter was to address the fourth research objective (RO4) which is to evaluate the capabilities of a Decision Support System that is developed by following the proposed framework in predicting grain commodities prices. To achieve this purpose, Chapter 7 set out by asking the research question RQ<sub>6</sub> – how well does a DSS perform, which was developed by utilising the framework? Criteria for evaluating the DSS were identified as broadly divided into the technical capabilities and the overall usefulness of the proposed DSS. It was clearly stated that the rigour cycle of this study laid the necessary foundations which ensured that major evaluation issues under both categories have not been neglected. But the need for empirical evaluations that measure the degree of accuracy of the predictions implemented from the modelling component of the DSS was identified. It was also identified that a comparison of the predictions from the DSS to predictions by a panel of experts can give an indication of the usefulness of the implemented DSS and that of the proposed framework.

Different experiments were carried out to identify the strategy that optimises the modelling component of the proposed framework that was implemented in Chapter 6. It was identified that including near-real-time data in the input data for training the models and running iterative training experiments to identify models that will increase the accuracy of the predictions was beneficial. A prediction using this strategy was compared to the outcome of the implementation in Chapter 6, together with two other options and the former was found to make more accurate predictions which were

closer to the actual prices recorded. This signified that the proposed framework could be used to develop a DSS that is technically sound for generating actionable insights.

A panel of eight experts was also asked to make predictions for a certain period and the implemented DSS was also used to make predictions over the same period. The predictions from the experts were compared to predictions from the DSS for the spot and December futures prices of white maize in South Africa. The result of the initial iteration suggested the need for improvements of the modelling component of the implemented DSS. This led to an iteration in the research process between the evaluation phase and the development phase where the scientific rigour was updated. Thereafter, the learning from the development phase was implemented as a second iteration of the evaluation. The iteration produced an outcome that was found to perform fairly better than predictions from the experts in most cases.

It should be noted that the evaluation in this chapter was done by comparing the performance of the DSS with many years of experience and the accumulated knowledge of the experts. This indicates that the farmers without the expertise for predicting grain commodity prices can make use of an implementation based on the proposed framework when making decisions related to trading in grain commodities. This evaluation also shows that the proposed framework can be used to provide the future outlook of the different grain commodities trading strategies. This makes it possible for users such as the grain commodities farmers to choose trading options that minimise their risks and increase their profitability.

Thus, the outcome of the empirical evaluation in Chapter 7 indicates that the framework for the support of decisions about grain commodities proposed in Chapter 6 fulfils the main research objective of this study (RO<sub>m</sub>) which is to design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities. Chapter 8 will extract the main achievements of this study and provide a summary of the entire study.

# Chapter 8 : Recommendations and Conclusions

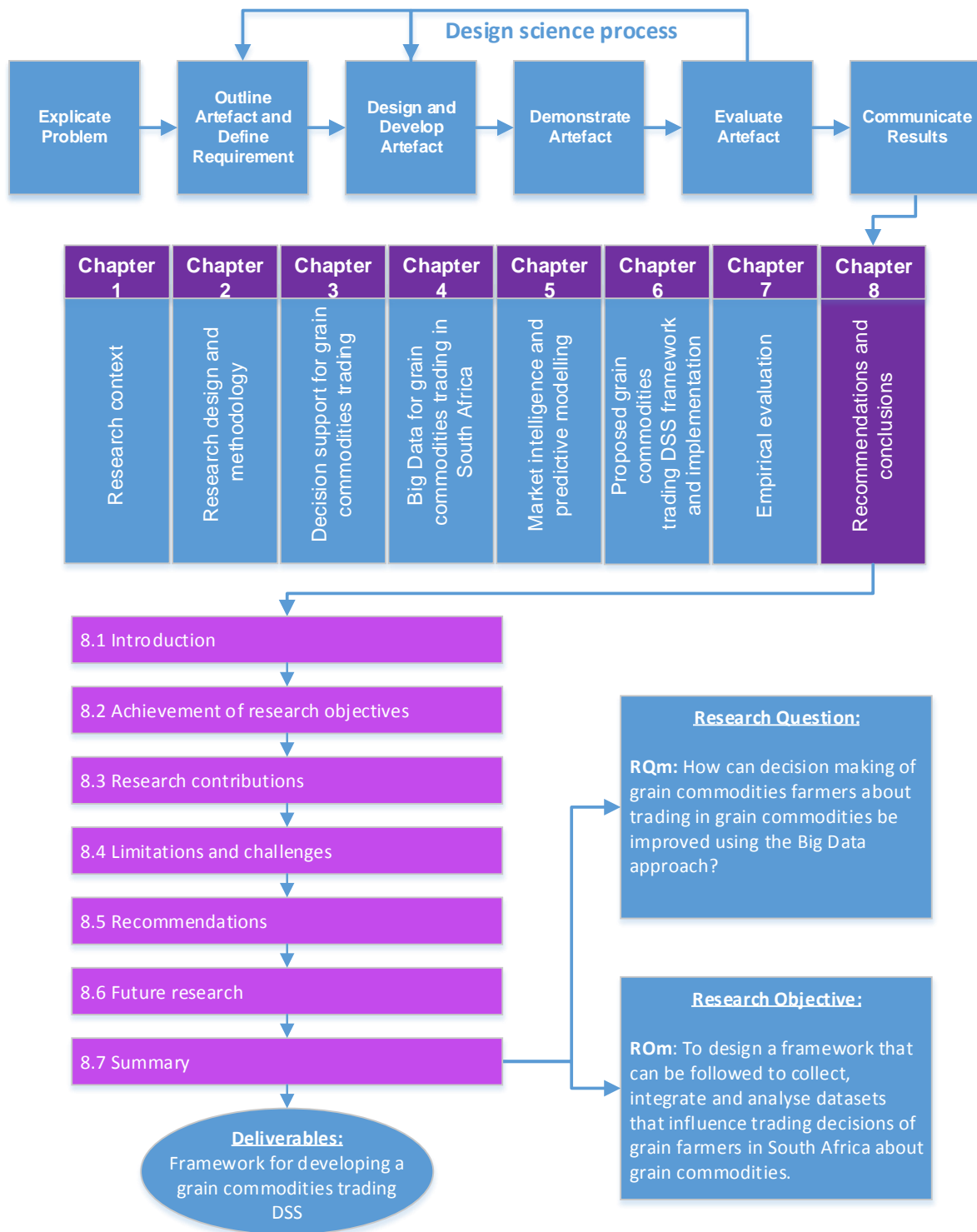


Figure 8.1: Chapter outline and deliverables

## 8.1 Introduction

Grain commodities are traded on the Johannesburg Stock Exchange in South Africa in the same way as they are traded on different stock exchanges the world over. This implies that besides selling grain commodities on a cash basis, known as spot, there are other trading strategies namely, forward contracts, futures contract and options as described in Section 3.4.1. The availability of these alternatives makes it possible for stakeholders to discover the best price for their produce and to effectively manage price-related risks (Venter, Strydom and Grové, 2013). However, the prices of grain commodities on the Johannesburg Stock Exchange locally and on other exchanges globally are described as being significantly volatile (Geysers and Cutts, 2007). This is because there are several factors that influence the prices of grain commodities as explained in Section 3.4.2. Moreover, the degree of influence of each of these factors varies with time, thereby making it necessary to keep abreast of several local and international indicators in order to understand and effectively take advantage of the available trading strategies (Venter, Strydom and Grové, 2013).

As the major producers and sellers of grain commodities in South Africa, farmers are important stakeholders in the industry. However, the farmers do not have the necessary skills and resources required to sift through, contextualise and make trading decisions using the data available on several factors that influence prices in South Africa (Section 1.2). This is because the data that provides the necessary indicators is available from disparate locations. As a result, many of the grain commodities farmers in South Africa do not take advantage of all the available trading strategies for mitigating price-related risks (Jordaan and Grové, 2010; Venter, Strydom and Grové, 2013). This could have implications of profitability and, in the long run, the sustainability of the operations of such farms due to price-related risks that they face yearly.

The purpose of this study was to propose a framework that can be followed to develop a decision support system that will enable relevant stakeholders to make decisions

regarding trading grain commodities in South Africa. In order to fulfil this purpose, the main research objective (RO<sub>m</sub>) set for this study was:

***To design a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities.***

The main objective of this study indicated the need to create an artefact that is grounded in multiple disciplines. Moreover, the main objective was broken down into four sub-objectives which are:

**RO<sub>1</sub>:** To identify data-related requirements for a system to support decisions on trading grain commodities in South Africa.

**RO<sub>2</sub>:** To identify modelling techniques for predicting the future prices of grain commodities in South Africa.

**RO<sub>3</sub>:** To develop a framework to support decisions on grain commodities trading.

**RO<sub>4</sub>:** To evaluate the capabilities of a Decision Support System that is developed by following the proposed framework in predicting grain commodities prices.

Besides the need for inter-discipline grounding that was required, the sub-objectives also indicate the possibility of an iterative research process. Therefore, the Design Science Research (DSR) methodology was adopted and used to successfully achieve the purpose and the objectives that were set at the beginning of this study. This chapter will provide an overview of how each of the set objectives and the main purpose of this study have been achieved (Section 8.2). A review of the theoretical and practical contributions of this study will be provided in Section 8.3. The limitations and constraints that were encountered during this study will be highlighted in Section 8.4. Thereafter, based on the knowledge acquired and developed during this study, recommendations will be provided in Section 8.5 and the opportunities for future

research that have been identified will be listed in Section 8.6. A summary of Chapter 8 will be provided in Section 8.7.

## **8.2 Achievement of Research Objectives**

The review of literature identified data, modelling, intelligence and visualisation as the components of a computer-based Decision Support System (DSS). These components formed the foundation of the solution that is proposed in this study. In the application of DSS for improving decision making, it was highlighted in Chapter 3 that a DSS can be used to drive the process of converting raw data to information, from which useful knowledge is created, which can then be used for improving decision making.

The ability to take advantage of the different trading strategies (spot, futures contracts, forward contracts, options and so on) was identified as opportunities for farmers to minimise their price-related risks and increase profitability (Section 3.4.1). However, for the farmers, to choose the best trading alternative, they need to predict the future prices of the grain commodities for the different trading alternatives. Hence, a study of the factors that influence the prices of grain commodities was carried out and the sources of data for each of the factors were identified (Section 3.4.2). These formed the basis modelling and subsequent prediction of prices for different trading alternatives in real-time. Thereby providing decision support for trading in grain commodities.

It was identified that the availability of the required data at the right time is important in providing decision support for trading in grain commodities as proposed by this research study. Therefore, this study adopted the Big Data approach to ensure that the data from disparate sources on the factors that influence grain commodities market in South Africa can be collected, integrated and analysed. Furthermore, the evolving Big Data tools, techniques and concepts were identified in Section 4.2, as offering opportunities for acquiring, integrating and extracting information that support decision making from datasets from multiple sources. Big Data tools and techniques also make

it possible to provide decision support in real-time and to deal with the associated challenges.

In order to achieve the main research objective of this thesis, four secondary research objective were set and each of these were also linked to research questions. Table 8.1 indicates how the research objectives were linked to research questions as well as the chapters of the thesis where the research objectives and questions were addressed. The following sub-sections summarise the achievement of the research objectives, together with the outcomes obtained in addressing the associated research question(s).

Table 8.1: Research objectives and questions addressed in study

Research Objectives		Research Questions	Chapter
<b>RO<sub>1</sub></b>	To identify data-related requirements for a system to support decisions on trading grain commodities in South Africa.	RQ <sub>1</sub>	<b>3</b>
		RQ <sub>2</sub>	
		RQ <sub>3</sub>	<b>4</b>
<b>RO<sub>2</sub></b>	To identify modelling techniques for predicting the future prices of grain commodities in South Africa.	RQ <sub>4</sub>	<b>5</b>
<b>RO<sub>3</sub></b>	To develop a framework to support decisions on grain commodities trading.	RQ <sub>5</sub>	<b>6</b>
<b>RO<sub>4</sub></b>	To evaluate the capabilities of a Decision Support System that is developed by following the proposed framework in predicting grain commodities prices	RQ <sub>6</sub>	<b>7</b>

### 8.2.1 Data-related requirements for a grain commodities trading DSS (RO<sub>1</sub>)

This objective was divided into three parts: the factors that influence the grain commodities market in South Africa, the different grain commodities trading strategies that are available and the datasets that represent the factors influencing the grain commodities market. The literature study in Section 3.4.1 identified selling grain commodities on a cash basis (spot transaction), forward contract, futures contract and



options instrument as the major strategies that are available for trading grain commodities in order to manage the price-related risks and increase profitability.

Demand and supply, storage levels, macroeconomic indicators and political factors were identified as the main grouping of the variables that influence grain commodities prices (Section 3.4.2). These factors are besides the effects of historical trade of the grain commodities themselves. It was also noted that some of the variables under each of these groups could be local, international or both in some cases. Surveys were carried out to confirm the perception about the factors that influence grain commodity prices in South Africa among local grain commodities farmers and traders. The result of the surveys validate findings from the literature (Section 3.5).

Chapter 4 examined the opportunities that exist in taking a Big Data approach for the acquisition of the necessary data required for a DSS that can be used when making decisions regarding trading grain commodities in South Africa. The possibility to stream data from different sources in real-time was identified and sources of data for the factors that were identified in Chapter 3 were discussed in Section 4.3. Therefore the first research objective (RO<sub>1</sub>) was successfully achieved.

Table 8.2: Summary of identified datasets and their sources

Factors	Datasets	Source of Data
Historical trade	Local grain commodities trade statistics	The Johannesburg Stock Exchange and SENWES
	Grain commodities trade statistics of the USA market	Chicago Board of Trade (CBOT) via EODDATA.COM
Demand and supply	Local grain commodities production, acquisition and consumption data	South African Grain Information Services (SAGIS)
	Data on the effect of international production, acquisition and consumption	Economic Research Services (ERS) of the United States

	of grain commodities (Using USA as an indicator)	Department of Agriculture (USDA)
	Data on the effect of weather conditions	South African Weather Service
Macroeconomics	Brent crude oil price data	Quandl.com
	Dollar–Rand exchange rate data	South African Reserve Bank
	South African interest rates data	South African Reserve Bank

Tables 8.2 present a summary of the identified datasets required for a grain commodities trading DSS and sources of the datasets as identified in Sections 4.3.1 and 6.4.2. The schema that shows the detail of the variables in each of the data is presented in Figures 6.3 and 6.4 for the creation of integrated data source for modelling the spot and the December futures contract of white maize as part of the demonstration of the proposed framework.

### **8.2.2 Modelling techniques for predicting the prices of grain commodities**

#### **(RO<sub>2</sub>)**

The second objective of this study was to identify the modelling techniques that will be most suitable for discovering patterns and insights from the data that was identified by achieving the first objective. The Data Science process was used for real-time acquisition, integration and investigation of the datasets to create market intelligence and predictions (Section 5.3). The literature review in Sections 5.3.1 and 5.3.2 provided an indication of initial preparation and integration of disparate data that has some of the qualities of Big Data such as was expected in the case of datasets that influence the prices of grain commodities.

The use of Neural Networks, specifically the Backpropagation Neural Network, was reviewed as a technique for developing models using time series data such as the grain commodities trading dataset (Sections 5.5 and 5.6). Although it is also possible to make use of statistical modelling techniques, it was identified from the literature

review in Section 5.6 that Neural Networks are more suitable based on the characteristics of the grain commodities trading data.

A mathematical model for using Backpropagation Neural Networks to predict the prices of grain commodities in real-time was proposed in Section 5.5.1. It was also noted that running the model periodically could be resource-intensive, but Big Data technological considerations, such as the provisions of in-memory, cloud and parallel computing provides the platform that make such intensive processing possible.

Thus, the second research objective and the fourth research question were addressed in Chapter 5 of this study. Within the DSR cycles, the first and the second research objectives were addressed as the rigour cycle in Chapters 3, 4 and 5. These chapters provided the scientific grounding, methods, important considerations and relevant examples required for this study. The outcome of this cycle formed the basis for designing a framework that can be followed to develop a system that supports decision making about trading grain commodities in subsequent parts of the study.

### **8.2.3 Developing the framework for grain commodities trading DSS (RO<sub>3</sub>)**

The third research objective of this study was to develop the framework for grain commodities trading decision support. The necessary components and considerations for developing such a framework have emerged during the rigour cycle that was carried out in Chapters 3, 4, and 5. Based on this knowledge, a framework for grain commodities trading DSS was proposed in Section 6.2.

The proposed framework that can be followed in implementing a DSS for grain commodities trading in South Africa is presented in Figure 8.2. The main components of the proposed framework are:

- **Domain knowledge:** This is for an on-going investigation of the factors that influence grain commodities market in South Africa.
- **Data acquisition and integration:** This involves the sourcing, acquisition and integration of data on identified factors that influence grain commodities prices

in South Africa. It is proposed that the data acquisition and integration should be done in real-time. A Big Data approach is proposed to deal with the envisaged volume, variety, velocity and veracity of the datasets.

- **Modelling:** The modelling component makes use of analytical techniques for an understanding of the patterns and relationships that exist in the integrated data. The proposed framework suggests the use of a Backpropagation Neural Network that is re-trained periodically, for modelling the prices of grain commodities. It was, however, noted that there are other Statistical and Artificial Intelligence techniques that might be applicable.
- **Intelligence:** This is the extraction of useful information that can enable the identification of alternatives, trade-offs and clarification of uncertainties when making decisions about trading in grain commodities. This can be in the form of price predictions, recommendations on what, when and how to trade or important discoveries about the market.
- **Visualisation:** This is the presentation layer, where the users interact with the DSS. It was identified that this layer should make it easier for all stakeholders to benefit from the system.
- **Technological consideration:** The application of Big Data tools, techniques and approach are considered critical for the success of the proposed framework.

The components of the proposed framework emphasise the tools, techniques, concepts and processes that should be followed in the implementation of a Big Data-inspired DSS for grain commodities trading in South Africa. Thus, the third research objective of the study was achieved.

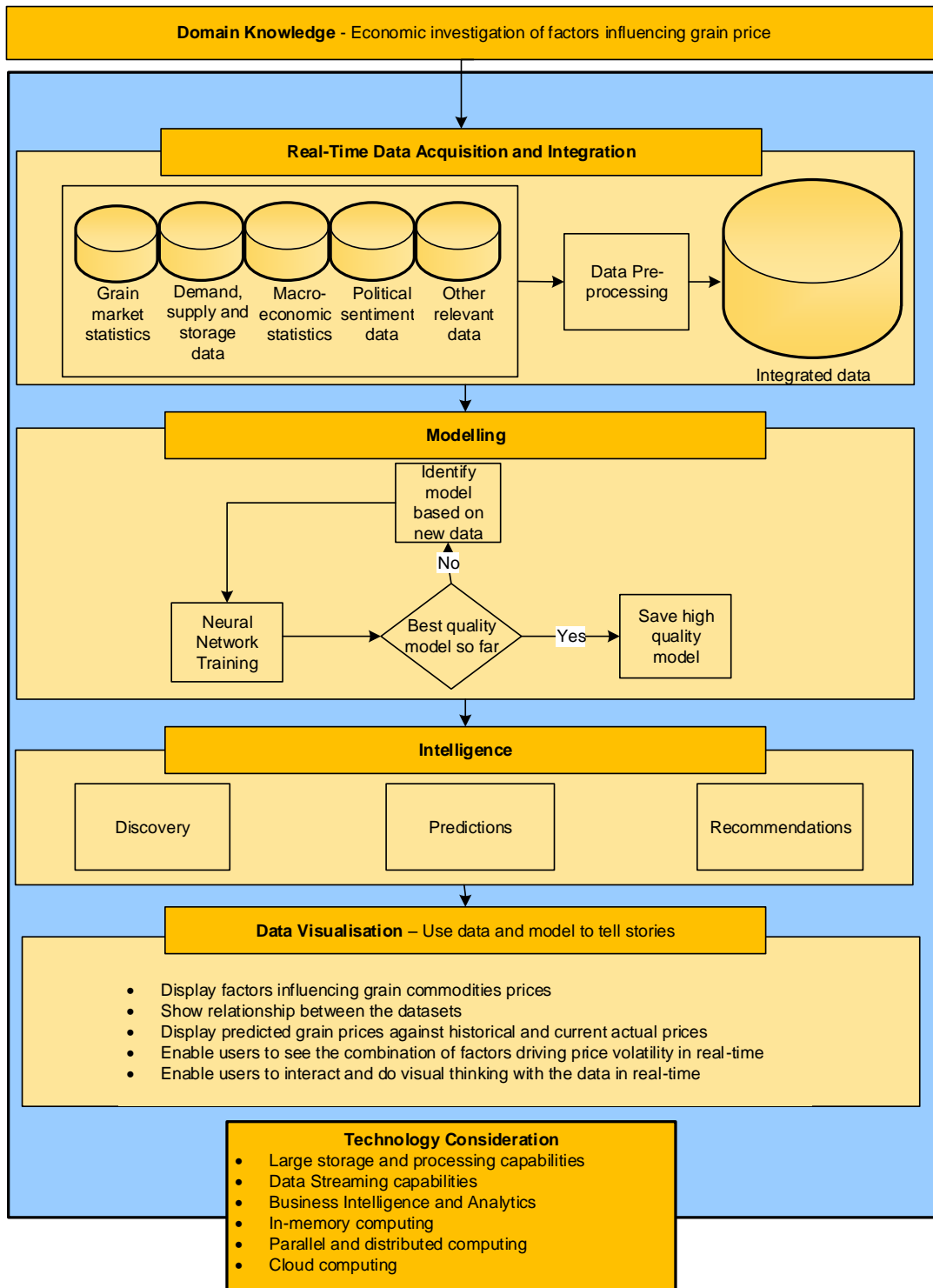


Figure 8.2: Proposed framework for grain commodities trading

### 8.2.4 Evaluation of the proposed framework (RO<sub>4</sub>)

The fourth objective that was set at the beginning of this study was to evaluate the performance of the framework providing support for making trading decisions about grain commodities. The DSR methodology that was followed in this study included a design cycle which involve the actual building, demonstration and evaluation of the artefact. The demonstration of the artefact involves showing that the developed artefact can address the identified problem, while evaluation, in this context, refers to showing how well the artefact can solve the problem (Johannesson and Perjons, 2012). An implementation of the proposed framework was carried out for the prediction of the spot and December futures contract price of white maize in South Africa (Sections 6.4 and 6.5). An exploratory analysis was carried out to determine the variables that should be included in the modelling and predictions of the spot and December futures contract prices of white maize (Sections 6.4.4). The results of exploratory analysis showing the input variables for modelling the spot and December futures contract prices of white maize are presented in Tables 8.3 and 8.4 respectively. The variables in the tables have been selected as input variable for developing the models because they exhibited relatively fair or high correlation with the spot and December futures prices of white maize respectively.

Table 8.3: Input variables for Backpropagation Neural Network model for predicting spot prices of white maize

No	Variables	Correlation with spot price of WMAZ
1	Spot price of WMAZ (lagged)	
2	Spot price of Wheat	0.6280 (n=2149)
3	USD-Rand exchange rate	0.5885 (n=2149)
4	Spot price of Brent Crude oil	0.3191 (n=2149)
5	Prime interest rate in SA	-0.3428 (n=2149)
6	Price of Corn in USA	0.2860 (n=2149)
7	Volume of Corn Trade in USA	0.2848 (n=2149)
8	Demand for WMAZ in SA	0.2474 (n=2149)
9	Demand for Wheat in SA	0.3347 (n=2149)

Table 8.4: Input variables for Backpropagation Neural Network model for predicting December future contract prices of white maize

No	Variables	Correlation with December futures of WMAZ
1	December futures prices (lagged)	
2	Spot price of WMAZ	0.9192 (n=1694)
3	Price of Wheat	0.7436 (n=1694)
4	Closing price of Corn on CBOT	0.3220 (n=1694)
5	Volume of Corn Trade in USA	0.2848 (n=1694)
6	Spot price of Brent Crude Oil	0.2492 (n=1694)
7	USD-Rand Exchange Rate	0.6314 (n=1694)
8	Prime interest rate	-0.4865 (n=1694)
9	Demand for WMAZ	0.1986 (n=1694)
10	Demand for Wheat	-0.3938 (n=1694)
11	BID price of December futures	0.7494 (n=1694)
12	OFFER prices of December futures	0.7260 (n=1694)

Furthermore, initial experiments were carried out to identify the optimum parameters for the Backpropagation Neural Network (BPNN) models for predicting the spot and December futures contracts prices of white maize (Section 6.5.1.3). The optimum BPNN models were found to be those with seven hidden layers and five lags (number of previous trading days of each variables with direct influence on the immediate future and considered as separate variables) as shown in Tables 6.8 and 6.9. The out-sample predictions made using the models suggest that the models are suitable for predicting the prices of white maize. The out-sample predictions for the spot prices of white maize are presented in Figures 6.18, 6.20 and 6.22 over a 1-month, 3-months and 6-months periods respectively. Moreover the predictions of the December futures contract prices are shown in Figures 6.24, 6.26 and 6.28 over a 1-month, 3-months and 6-months periods respectively. This proved that the framework can be followed to

develop a system that supports decision making about trading grain commodities in South Africa.

The performance, the technical abilities and the overall usefulness of the system built in Chapter 6 using the proposed framework was evaluated empirically in Chapter 7. The evaluation task resulted in an iterative process to incorporate real-time Neural Networks learning in order to improve performance of the DSS (Section 5.7). This was implemented in the evaluation process (Section 7.3.2) and performance was significantly improved. The evaluation process was carried out by comparing predictions from the system that was built based on the proposed framework to predictions by eight experienced grain commodities traders with expert knowledge in making such predictions.

The evaluation showed that a DSS that re-trains the BPNN model periodically as new data becomes available made more accurate predictions of the spot and December futures contract prices of white maize than predictions made by the experts. A comparison of the predictions by the experts and that of the BPNN model from the proposed framework for spot prices and December futures contract is presented in Figures 8.3 and 8.4 respectively.



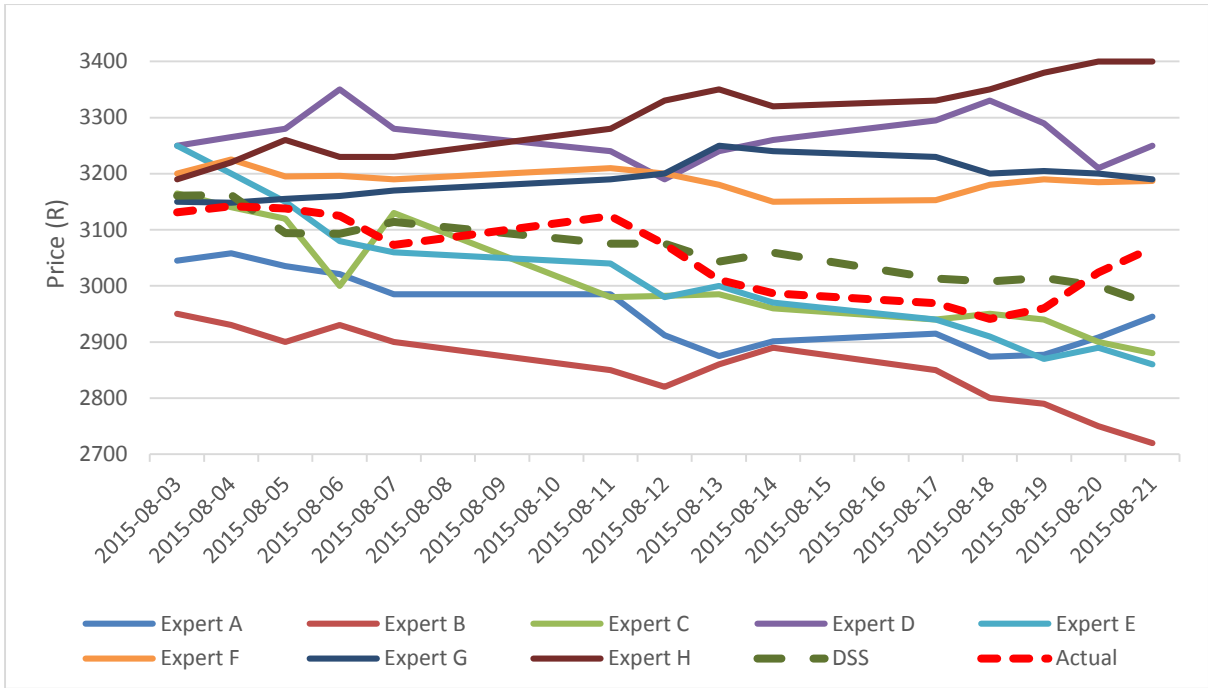


Figure 8.3: Prediction of spot prices of white maize by experts and DSS

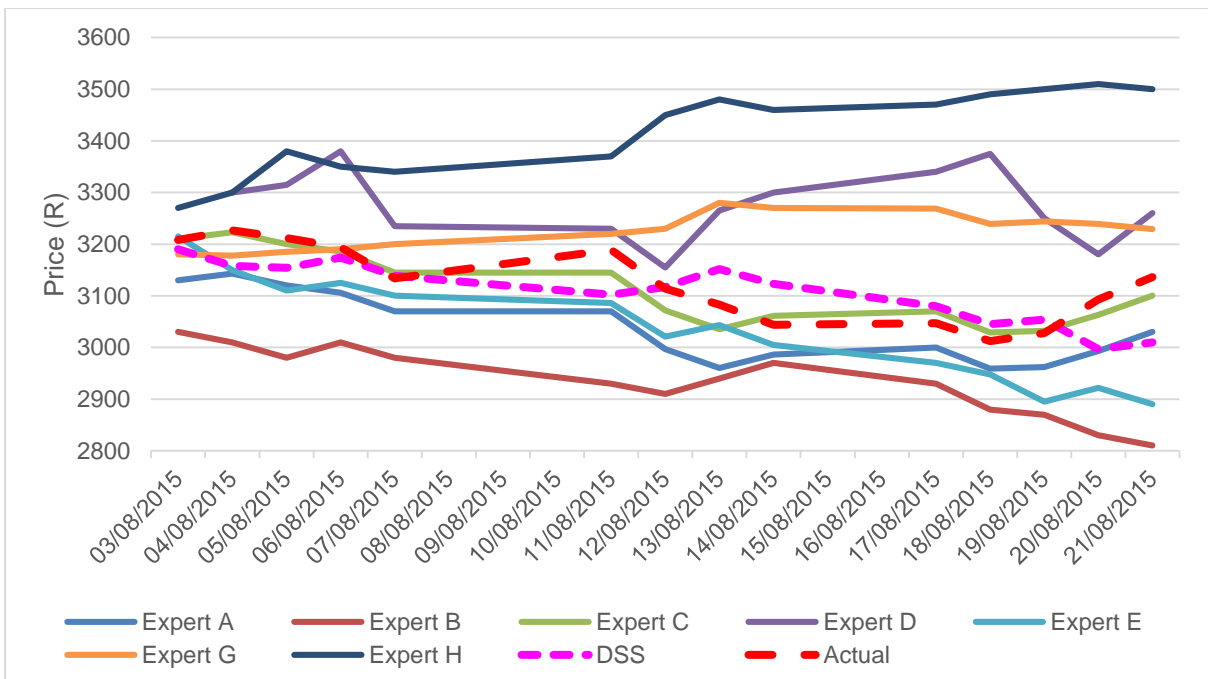


Figure 8.4: Prediction of December futures prices of white maize by experts and DSS

The prediction by the BPNN model for spot prices had the minimum error with the Mean Absolute Percentage Error (MAPE) = 1.44% and Root Mean Square Error (RMSE) = 49.91 when compared to the predictions of the experts. Out of the predictions made by eight different experts, the closest predictions for the spot prices of white maize over the same period had a MAPE = 2.16% and RMSE = 85.78. On the other hand, the BPNN model for December futures contract prices over the same period had MAPE = 1.67% and RMSE = 26.28. Predictions of the December futures contract prices of white maize by the BPNN model were compared to the predictions by the seven experts. One of the experts made predictions that out-performed those of the BPNN with MAPE = 0.69% and RMSE = 26.28. However, the BPNN out-performed the rest with the closest having MAPE = 2.81% and RMSE = 88.80.

The results showed that the system based on the proposed framework performed better than most of the traders that participated in the evaluation. Therefore, by following the proposed framework for developing a DSS, the future prices for different grain commodities trading strategies can be presented to users. Thereby, making it easier for them to make informed decisions about trading their grain commodities, without having to sift through or interpret data on several sources that influence the prices of grain commodities. Hence, the fourth sub-objective of the study was successfully achieved.

A combination of the sub-objectives proposed a framework that can be followed to collect, integrate and analyse datasets that influence trading decisions of grain farmers in South Africa about grain commodities. Thus, achieving all the sub-objectives of the study indicates that the main objective and purpose of this study was successfully achieved.

### **8.3 Research Contributions**

This study makes several significant contributions in the application of Information and Communication Technology (ICT) and other disciplines to grain commodities trading. The contributions of this study are applicable locally in South Africa and can be

adapted internationally. Contributions of this study to the body of knowledge can be divided broadly into theoretical and practical contributions.

### **8.3.1 Theoretical contributions**

This study is based on the theoretical foundations of the decision making theory and the information processing theory (Section 3.2), which formed the basis of the theoretical contributions of this study. Two theoretical contributions were made by this study. The first theoretical contribution is the list of factors that influence grain commodities trading, sources of the data and an integrated data source for developing market intelligence that support decisions regarding trading in grain commodities in South Africa. The second theoretical contribution is an adaptation of Big Data tools, techniques and concepts for the acquisition, integration and processing of relevant data for providing decision support regarding grain commodities trading.

The first theoretical contribution of this research is a list of factors that make it possible to identify the alternatives and clarify uncertainties during the decision making process. Based on the decision making theory, an integration of datasets from disparate sources on factors that influence prices can form the basis for structured decision making process when trading in grain commodities. The decision making theory suggests that the process of making decisions in an organisation should be a logical process. According to the decision making theory, the tasks and steps taken provides a model for decision making, especially in organisations as described in Section 3.2.1. Using the tasks and steps taken in decision making is presented graphically in Figure 8.5. The figure suggests that decision making involves three phase - the identification of an opportunity for decision making, identification of potential alternatives and then selecting from known alternatives (Hammond, Keeney and Raiffa, 1999; Bazerman and Chugh, 2006).

Contextually, it was identified that decision making regarding the trading of grain commodities requires the identification of options and the need for a reasonable assessment of future risks in making the right trading decision.

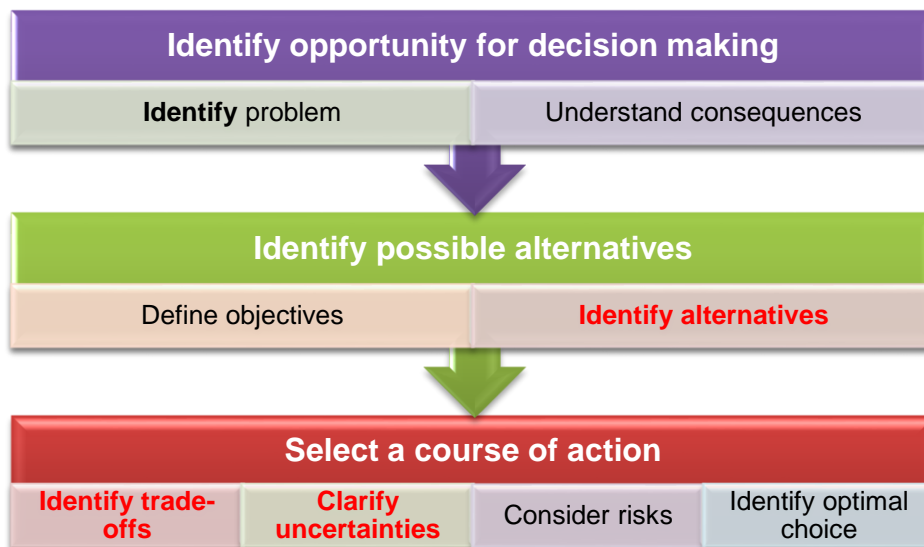


Figure 8.5: Decision making model (Adapted from Simon, 1960; Hammond, Keeney and Raiffa, 1999; Bazerman, 2006)

Table 8.2 presents, as part of the achievement of the first objective of this study, a list of datasets that can be streamed and integrated into a single data source in real-time as well as the sources of the data. The list of factors that influence the grain commodities prices were identified through a combination of literature study and surveys (Sections 3.4 and 3.5). The identified factors were categorised under demand and supply factors, macroeconomic factors and political factors. Each category of factors have local and international indicators, many of which were identified during this study. Based on the identified indicators, sources of data for the indicators were identified (Section 4.3.1) and this study demonstrated that data for the indicators can be acquired and integrated (Sections 6.4.2 and 6.4.3). Such integrated data source can then be analysed to extract patterns and intelligence in order to identify alternatives, trade-offs and clarify uncertainties (as highlighted in Figure 8.3) when making decisions regarding trading grain commodities.

The second theoretical contribution is an investigation into the possible application of the Big Data tools, techniques and concepts for the acquisition, integration and processing of relevant data, in order to provide decision support regarding trading grain commodities. It makes provision for the volume, variety, velocity and veracity of data from disparate sources required for effective decision making. The information processing theory highlights that uncertainty and ambiguity are the main impediments to optimal decision making in business (Daft and Lengel, 1986). Uncertainty is due to lack of information, while ambiguity refers to the inability to comprehend or deal with the volume of information available for decision making (Kowalczyk and Buxmann, 2014).

This study identifies uncertainty and ambiguity as the problems that are faced by stakeholders such as the farmer with limited skills and resources in managing price-related risks when trading their grain commodity. Thus, a Big Data-inspired information processing was adapted by this research study, for providing decision support in trading grain commodities. This involves the real-time acquisition, integration and analysis of data from disparate sources that influence grain commodities prices. It also includes the use of Data Science process (Section 5.3) for information processing that could enable a farmer with limited skill and resources in making the right decisions about trading his grain commodities. This will make information and insights required for decision making available at the right time and in a simplified format.

### **8.3.2 Practical contributions**

This study made practical contributions for stakeholders involved in grain commodities trading in South Africa and possibly internationally, as well as the ICT practitioners that serve the industry. The practical contributions of this study include a framework that can be followed in developing a DSS that can support grain commodities trading decision making. This study also contributed a Backpropagation Neural Network (BPNN) model for predicting the spot and December futures contract prices of white maize in South Africa. An integrated source of data on the end-of-day transactions

and external factors that influence the prices of grain commodities on the Johannesburg Stock Exchange was also an outcome this study.

Firstly, this study presented a framework that can be used for developing a DSS for trading in grain commodities as a practical contribution. This was done by identifying components, approaches and processes that can be followed in the development of a DSS that could eliminate uncertainties and ambiguities during decision making about grain commodities trading. The proposed framework is presented in Figure 8.2 as the achievement of the research objective (RO<sub>3</sub>).

The proposed framework was followed in developing a DSS for choosing between the spot and December futures contract for trading white maize on the JSE. A study was carried out to identify the factors that influence the prices of white maize in South Africa. Thereafter, datasets on the identified factors were acquired and integrated into a single data source (Section 6.4.2). Further investigation was carried out to understand the relationship between the prices of both trading alternatives (spot and December futures contract). The result of this exploratory investigation was used to decide on the input variables for modelling the spot and December futures contract prices (Section 6.4.4).

The second practical contribution of this study is a Backpropagation Neural Network (BPNN) model for the prediction of the spot and December futures contract prices of white maize in South Africa. Based on literature review, exploratory analysis and experiments (Sections 6.4 and 6.5) were carried out to determine the factors that should be included in the model. Experiments were also carried out to determine the appropriate topology and structure of the BPNN model (Section 6.5.1.3) for predicting the spot and December futures contract prices of white maize in South Africa. This was achieved as a research objective (RO<sub>4</sub>) for demonstrating the proposed framework's technical ability and usefulness in supporting decisions about which alternative to use when trading white maize in South Africa (Section 8.2.4).

Table 6.4 and 6.5 presented the input variables that were used modelling the spot and the December futures contract of white maize in South Africa respectively. Furthermore, models with seven hidden layers and five lags were found to produce optimal results for both the spot and December futures contract prices. The learning rate of 0.4 and momentum factor of 0.0001 were used for modelling both the spot and the December futures contract prices of white maize (Section 6.5.1).

The literature study (Section 5.7) and empirical results (Section 7.3.2) from this study also indicated that a real-time learning strategy can be used to improve performance of the BPNN model for predicting the prices of white maize. Predictions of this model can be used to determine the future outlook of white maize in South Africa for spot and December futures prices. Moreover, the principles followed in developing the models can also be used in developing models for other trading strategies for white maize. This will provide an avenue for measuring possible risks for whichever trading strategy alternative that is adopted.

By utilising the BPNN models that were periodically re-trained as new data, predictions of the spot and December futures contract prices made, using the models, were found to out-perform predictions by expert traders of a selected period. Predictions from the models were found to be reasonably closer to the observed actual prices and much better when compared to predictions made by experts (Section 7.3.3). Thus, it can be concluded that the proposed models are suitable for predicting the spot and December futures contract prices of white maize in South Africa, although the models can still be improved.

The third practical contribution of this research study is a database that integrates the end-of-day trade data about grain commodities and different factors that influence the prices. Based on the literature review in Section 3.4 and surveys carried out in Section 3.5, the different datasets on factors that influence the prices of grain commodities in South Africa were collected during this research study (Section 6.4.2). Table 8.2 present a list of the datasets that were collected and sources of the datasets. The

datasets were in different formats when acquired, and some of the datasets had to be restructured. Thereafter, all the datasets were integrated into a single data source that can be used for providing market intelligence and analytics on the grain commodities market in South Africa (Section 6.4.3).

The DSS that was developed by following the proposed framework from this research study, made use of the created data source for predictive modelling of the spot and December futures contract prices of white maize in South Africa. Schemas of the data tables used for developing the DSS is presented in Figure 6.3 and 6.4. The database is hosted using the SAP HANA cloud computing services provided by Amazon. Hence, the database can be updated by streaming data from the disparate sources as new data becomes available. The created database can be used by different practitioners for developing further decision support for the industry. Researchers can also make use of the same database for further research on decision support or other relevant topics for the industry.

## **8.4 Limitations and Challenges**

Several limitations and challenges were encountered during this research. The first of the challenges that was faced in the course of this study was the access to the right data. Some of the datasets that were available for the demonstration and empirical evaluation of the proposed framework were largely unstructured and incomplete. In the process of integrating datasets from different sources, a portion of the available dataset had to be ignored because complementary data for other factors could not be found. Out of all the datasets used in this study, only the data for the Chicago Board of Trade (CBOT) was structured and complete, because the data was made available by a third party agent as a paid service.

Moreover, challenges were also encountered in obtaining data in more granular format where it was deemed possible. A reputable data vendor offered access to some of the data required for this study in more structured and granular format, but the licencing costs were beyond the budgets of this research project. It is possible to have grain



commodities trading data in hourly, minutes or even seconds' basis, but this study was limited to the end-of-day data which considerably reduced the volume of data used in this study. However, it is noted that it is possible to acquire grain commodities trading data by the second, minute or hour, in which case, the framework can be used to generate market intelligence and insight at the same rate. This could not be demonstrated by this study.

It was identified in this study that sentiment analysis could be carried out on social media to collect quantitative data about the influence of political factors on grain commodities prices. However, this was not included in the model, due to the cost involved in gaining access to historical data on Twitter. Due to the constraints of time, financial resources and access to the correct data, the demonstration and the evaluation of the proposed framework was limited to a single grain commodity – white maize. Furthermore, the identification and comparison of alternatives available when making trading decisions about grain commodities were also limited to spot and the December futures contract prices of white maize alone.

## **8.5 Recommendations**

This study was built on scientific grounding, methods and concepts from multiple disciplines. Research efforts in the area of grain commodity prices and trading have previously been in the domain of agricultural economics only. This study has shown the existence of new opportunities for multi-disciplinary research. Therefore, a cross-discipline approach that involves researchers in the area of Agricultural Economics, Business Management, Mathematics, Data Science, Information Systems, and Computer Science is recommended. This will further enhance the opportunity to establish relationships among concepts across the different disciplines for improving decision making about grain commodities trading among different stakeholders.

It was identified during this study that the availability of more data could lead to the possibility of more information and knowledge that improve decision making. However, some of the mandates of the Johannesburg Stock Exchange (JSE) are risk

management and price discovery (Venter, Strydom and Grové, 2013). The JSE has done this by facilitating different trading options and making some data available. However, some of the data made freely available by the JSE was highly unstructured and difficult to understand. It is therefore recommended that the JSE improves the quality of data that is made available through its website. Furthermore, it is recommended that the JSE makes it easier for different stakeholders to access the more structured and granular data in real-time.

Part of the mandates of the Department of Agriculture, Forestry and Fisheries is to ensure food security, transformation of the sector through innovation and to promote the access to information. It is therefore suggested that this government department invests in the acquisition and development of structured open data on the factors that influence grain commodities prices in South Africa. These data sources should be updated in real-time. Such investment in Big Data for the grain commodities market in South Africa will encourage more research and innovation into reducing the uncertainty and ambiguity in grain commodities trading for all the stakeholders (McLeod, 2012; Janssen, Charalabidis and Zuiderwijk, 2012).

ICT practitioners that provide services in the agricultural sector are encouraged to consider the implementation of the proposed framework from this study into a DSS that can benefit all stakeholders in the industry which include the farmers. Although the development of the DSS involves the acquisition, integration and analysis of large datasets for information and knowledge, it is recommended that care should be ensured so that the outcome of the DSS is not just another set of data that is difficult to understand. The outcome of the DSS should be presented in a simple way that benefits all stakeholders. This can be done by further contextualising the market intelligence, such as the predictions, to tell a story that can be easily understood. It is further recommended that the use of mobile technologies such as smartphones should also be considered for the visualisation component of the proposed framework and for user experience.

The outcome of this study indicates that data can become more valuable when complementary datasets on different topics from different sources are identified and integrated. Therefore, it is recommended that Big Data centres that collect, analyse and make available datasets on different social and economic issues, be developed in South Africa. Such Big Data centres could be at national or provincial level with research entities such as Universities as custodians. These Big Data centres should drive an open data agenda by encouraging the reuse of data. This will drive innovations among researchers and practitioner which could lead to further socio-economic developments and participation of South Africa in the knowledge economy.

This study has shown that a DSS that is built by using the framework proposed in this research can help stakeholders like farmers to improve their decision making regarding the trading of their grain commodities. To maximise the benefits, it is recommended that training on how to mitigate price-related risks and increasing profitability should be done for the farmers. The training should outline the different trading options available and how the created DSS can help assist them in making better decision.

## **8.6 Future Research**

The findings of this research open up new opportunities for future research. This study found that the grain commodities farmers have been systematically disadvantaged in the grain commodities trading industry in South Africa. This is because trading assets on the stock exchange can arguably be regarded as part of the knowledge economy that is growing globally. Hence the access to the right information and knowledge is critical for success in the industry. This study has laid a foundation for the opportunities of using Big Data for improved access to information and decision making regarding grain commodities trading in South Africa. In addition, the application of Artificial Intelligence as an alternative to using statistical techniques for understanding and extracting insight from grain commodities market data was also explored. The findings of this study opens up opportunities for several future research such as:

- Political factors were found in this study as one of the factors that influence grain commodities prices. There is an opportunity to further understand how the quantitative indicators, both for the local political landscape and the global political climate, affect the grain commodities market. One of such future research projects could be into how to use social media such as Facebook and Twitter to collect data that can be analysed as quantitative political indicators. Further studies can then be carried out to include such political indicators in a model for understanding grain commodities trading.
- This study has focused on the grain commodities farmers and other stakeholders with limited skill and resources, specifically because of the identified problems. However, the output of this study can be used by all stakeholders in the industry. When the majority of the stakeholders have access to the same insight about the market, there is a tendency that they will all make the same decisions and thus react to the market in this way. Therefore the grain commodities prices for all the available trading options will also begin to reflect that all the stakeholders have the same information, thus, they will become an additional influence in the market. This introduces market efficiency where the prices reflect information that is available (Hull, 2012). How to deal with possible market efficiency without making a subset of the industry to become disadvantaged, poses opportunity for further research.
- The prices of grain commodities in the United States of America as traded on the Chicago Board of Trade (CBOT) was used in this study as an indication of the effect of the international grain commodities market. But, the grain commodities market in other countries such as the European, Australian, Canadian and so on could also affect the grain commodities market in South Africa. Therefore, future studies can be carried out to include datasets on grain commodities trade from this countries to measure their influence on the South African grain commodities market.
- Further research can also be carried out to identify other Artificial Intelligence algorithms and techniques besides Backpropagation Neural Network that can be used to extract insight from datasets on grain commodities trading in South

Africa. Moreover, the literature review in Section 5.6 suggested that combining different modelling techniques could improve the resulting insight. Therefore, further research can also be carried out to find out which combination of modelling techniques and algorithms can be used to improve the accuracy of price predictions.

- This study focused on the prediction of grain commodities prices for different trading strategies as a means of supporting decision making. However, the foundation that has been provided from this study, especially with the application of Big Data and Data Science concepts, offers opportunities for more insights. It is suggested that future research can focus on using the foundation laid by this study to extract other insights such as providing stakeholders with recommendations and prescriptive insights. Moreover, other innovative discoveries that can improve decision making about trading grain commodities in South Africa, can also be researched further by using the findings of this study as a base.
- A DSS that was developed by following the proposed framework in this study was evaluated by predicting the spot and December futures contract prices of white maize in South Africa by the DSS. The evaluation was done by comparing predictions from the DSS to predictions by experts. There are more research opportunities for the further evaluation of the proposed framework, such as the use of the proposed framework to develop a DSS for trading other grain commodities besides white maize. Furthermore, the proposed framework can be evaluated for its ability to provide other kinds of decision support other than the prediction of grain commodities prices. Moreover, a DSS that has been developed by the proposed framework can also be evaluated among farmers and other stakeholders.

## **8.7 Summary**

The main problem that was addressed by this study is that grain farmers in South Africa do not take full advantage of all the available strategies for trading their grain commodities. It was identified that this is due to the complexities associated with

monitoring the large datasets that influence the grain commodities market. In addition, the farmers do not have required skills and resources to monitor and make decisions based on such datasets. Hence, the main objective of this study (RO<sub>m</sub>) was:

***To design a framework that can be followed to collect, integrate and analyse the datasets that influence trading decisions of grain farmers in South Africa.***

Two main artefacts were developed during the course of achieving this main objective of this research (Section 8.3.2). A framework that can be followed in creating a Decision Support System (DSS) which can assist grain farmers in taking advantage of all the trading strategies for minimising their price-related risks was proposed (Figure 8.2). In the course of evaluating the proposed framework, models for predicting the spot and December futures contract of white maize in South Africa were also developed (Section 8.2.2). Moreover, this study emphasised that the Big Data and Data Science concepts can introduce new opportunities for reducing uncertainty and ambiguity in making decisions regarding trading grain commodities in South Africa, even for users with limited skills and resources.

This chapter outlines the achievement of the sub-objectives that were set at the beginning of this study that resulted in the achievement of the main objective (Section 8.2). The achievement of the first sub-objective (RO<sub>1</sub>) led to the identification of the local and international factors that influence grain commodities trading in South Africa. Furthermore, the different trading strategies that can be used to minimise price-related risks were also identified. Thereafter, datasets that can be used to represent the factors that influence the grain commodities were also identified (Section 8.2.1). The achievement of the second sub-objective (RO<sub>2</sub>) was outlined in Section 8.2.2. It was shown that Backpropagation Neural Network was adopted for modelling of price prediction, though other techniques were also identified during this study. Section 8.2.3 described the achievement of the third sub-objective (RO<sub>3</sub>) which is the proposed framework that can be followed in developing a DSS for grain commodities trading.

Lastly Section 8.2.4 outlined the successful evaluation of the capabilities of a DSS that was built by following the proposed framework as the achievement of the fourth sub-objective (RO<sub>4</sub>) of this study.

The Design Science Research (DSR) was followed in the achievement of the objectives of this study. The explication of the problem, which included the setting of the objectives of this study was achieved at the beginning of the study as the relevance cycle of this study. Pursuing the first two sub-objectives (RO<sub>1</sub> and RO<sub>2</sub>) of this study formed the rigour cycle that represent the scientific grounding for this study. The achievement of the third and fourth sub-objective (RO<sub>3</sub> and RO<sub>2</sub>) was done in the design cycle.

As part of the DSR process, it is required that results be communicated to the relevant research community and practitioners. This thesis has been written to communicate the process and the finding of this research (Johannesson and Perjons, 2012). Section 8.3 provides a summary of the theoretical and practical contributions of this study. In addition to this, knowledge developed during this research has also been presented as a peer-reviewed conference paper (Appendix K) and another article which has been submitted for publication as a peer-reviewed journal article (Appendix L).

Section 8.4 described the limitations and the challenges that were faced during this study with the access to the right data as the major issue encountered. Specific recommendations as a result of this study were made in Section 8.5. It was recommended that the government should consider making available access to data that influences the market as part of fulfilling its mandates for sustainability of the agricultural industry. Moreover, it was recommended that ICT practitioners that service the agricultural industry consider innovations that are built on Big Data such as developing Decision Support Systems that are based on the framework proposed by this study for adding value to the sector.

This study has demonstrated that the Big Data concept and approach can be used to provide support for decisions about trading grain commodities. The Big Data concept suggests that more data is being created and made accessible globally on different subjects. It was highlighted during this study that the ability and the availability of the enabling systems to integrate data from several sources can improve productivity and profitability. The results of this study have shown that it is possible to collect and integrate data on the factors that influence grain commodities trading in real-time. Furthermore, findings from this study show that the grain commodities transactions can be analysed in real-time to extract insights that improve decision making for the future for all stakeholders in the industry.

Based on the findings and outcome of this study, an opportunity for the establishment of Big Data centres that collect, analyse and make data on different socio-economic issues was recommended. Such Big Data centres should also be involved in scientific and applied Big Data research, thereby setting the Big Data agenda for South Africa, encouraging inter-disciplinary research, driving innovation. This will further encourage the participation of South Africa in the global knowledge economy.

---



## REFERENCES

- Abbott, P.C., Hurt, C., and Tyner, W.E., 2011. What's driving food prices in 2011. *Farm Foundation Issue Report*.
- Alpaydin, E., 2010. *Introduction to machine learning*. Cambridge, MA: The MIT Press.
- Apache, 2015. *Apache Hadoop*. [online] Available at: <https://hadoop.apache.org> [Accessed 9 Mar. 2015].
- Ayankoya, K., Calitz, A.P., and Greyling, J.H., 2014. Intrinsic relations between Data Science, Big Data, Business Analytics and Datafication. In: *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference*. Centurion: ACM Digital Library, pp.192–198.
- Ayankoya, K., Cullen, M., and Calitz, A.P., 2014. Social media marketing in politics. In: *International Marketing Trends Conference*. Venice.
- Bazerman, M., 2006. *Judgement in managerial decision making*. 6th ed. Hoboken, NJ: John Wiley.
- Bazerman, M., and Chugh, D., 2006. Decision-making without blinders. *Harvard Business Review*, 84(1), pp.88–97.
- Bazerman, M.H., and Moore, D.A., 2013. *Judgement in managerial decision making*. 8th ed. New York: Wiley.
- Beck, R., Weber, S., and Gregory, R.W., 2013. Theory-generating design science research. *Information Systems Frontiers*, 15(4), pp.637–651.
- Bell, J., 2015. *Machine Learning: Hands-on for developers and technical professionals*. Indianapolis: John Wiley.
- Bennett, C., Stewart, R.A., and Lu, J., 2014. Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*, 7(5), pp.2938–2960.
- Bishop, C.M., 2006. *Pattern recognition and machine learning*. New York: Springer-Verlag.
- Blumberg, B., Cooper, D.R., and Schindler, P.S., 2011. *Business research methods*. 3rd ed. Berkshire: McGraw-Hill.
- Borglund, E., and Engvall, T., 2014. Open data? *Records Management Journal*, 24(2), pp.163–180.
- Box, G., and Jenkins, G., 1970. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Brown, B., Sikes, J., and Willmott, P., 2013. *Bullish on digital: McKinsey global survey results*.
- Brynjolfsson, E., Hitt, L.M., and Kim, H.H., 2011. Strength in numbers: How does

- data-driven decision-making affect firm performance? In: *Thirty Second International Conference on Information Systems*. Shangai.
- Bukharov, O.E., and Bogolyubov, D.P., 2015. Development of a decision support system based on neural networks and a genetic algorithm. *Expert Systems with Applications*, 42, pp.6177–6183.
- Burda, M., and Wyplosz, C., 2009. *Macroeconomics*. 5th ed. New York: Oxford University Press.
- Cambria, E., Schuller, B., Xia, Y., and Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), pp.15–21.
- Carlsson, C., and Turban, E., 2002. DSS: Directions for the next decade. *Decision Support Systems*, 33(2), pp.105–110.
- Carlsson, S.A., Henningsson, S., Hrastinski, S., and Keller, C., 2011. Socio-technical IS design science research: Developing design theory for IS integration management. *Information Systems and e-Business Management*, 9(1), pp.109–131.
- Chae, B.K., and Olson, D.L., 2013. Business analytics for Supply Chain: A dynamic capabilities framework. *International Journal of Information Technology & Decision Making*, 12(1), pp.9–26.
- Chaudhuri, S., Dayal, U., and Narasayya, V., 2011. An overview of business intelligence technology. *Communications of ACM*, 54(8), pp.88–98.
- Chen, H., Chiang, R.H.L., and Storey, V.C., 2012. Business intelligence and analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), pp.1165–1188.
- Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., and Zhou, X., 2013. Big data challenge: A data management perspective. *Frontiers of Computer Science*, 7(2), pp.157–164.
- Chen, P.C.L., and Zhang, C.Y., 2014. Data-Intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, pp.314–347.
- CMEGroup, 2015. *Corn futures time & date*. [online] Available at: [http://www.cmegroup.com/trading/agricultural/grain-and-oilseed/corn\\_quotes\\_timeSales\\_globex\\_futures.html](http://www.cmegroup.com/trading/agricultural/grain-and-oilseed/corn_quotes_timeSales_globex_futures.html) [Accessed 5 May 2015].
- Co, H.C., and Boosarawongse, R., 2007. Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks. *Computers and Industrial Engineering*, 53(4), pp.610–627.
- Collis, J., and Hussey, R., 2009. *Business research: A practical guide for undergraduate and postgraduate students*. 3rd ed. New York: Palgrave Macmillan.
- Condie, T., Mineiro, P., Neoklis, P., and Weirner, M., 2013. Machine Learning for Big Data. In: *SIGMOD'13*. New York, pp.939–941.

- Courtney, H., Lovallo, D., and Clarke, C., 2013. Deciding how to decide: A tool for executives making high-risk bets. *Harvard Business Review*, pp.62–71.
- Courtney, J.F., 2001. Decision making and knowledge management in acquiring organizations: Toward a new decision. *Decision Support Systems*, 31, pp.17–38.
- Creswell, J.W., 2014. *Research design: Qualitative, quantitative and mixed method approached*. 4th ed. Los Angeles: Sage Publications.
- Crone, S.F., and Kourentzes, N., 2010. Feature selection for time series prediction: A combined filter and wrapper approach for neural networks. *Neurocomputing*, 73, pp.1923–1936.
- DAFF, 2014. Trends in the agricultural sector 2013. *Department of Agriculture, Forestry and Fisheries*.
- Daft, R.L., and Lengel, R.H., 1986. Organisational information requirements, media richness and structural design. *Management Science*, 32(5), pp.554–571.
- Davenport, T., 2010. *Analytics at work: Smarter decisions, better results*. Boston: Harvard Business Press.
- Davenport, T., 2014. *Big Data at work*. Boston: Harvard Business School Press.
- Davenport, T.H., 2009. Make better decisions. *Harvard Business Review*, 87(11), pp.117–123.
- Davenport, T.H., Barth, P., and Bean, R., 2012. How Big Data is different. *MIT Sloan Management Review*, 54(1), pp.21–24.
- Davenport, T.H., and Harris, J.G., 2007. *Competing on Analytics: The new science of winning*. Boston: Harvard Business School Press.
- Davenport, T.H., and Patil, D.J., 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, pp.70–77.
- Delen, D., and Demirkan, H., 2013. Data, information and analytics as services. *Decision Support Systems*, 55, pp.359–363.
- Demirkan, H., and Delen, D., 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and Big Data in cloud. *Decision Support Systems*, 55, pp.412–421.
- Dhar, V., 2013. Data Science and prediction. *Communications of the ACM*, 56(12), pp.64–73.
- Dix, A., 2009. Human computer interaction. In: L. Liu and T.M. Ozsu, eds., *Encyclopedia of database systems*. New York: Springer, pp.1327–1331.
- Dong, X.L., and Srivastava, D., 2013. Big Data integration. In: *IEEE International Conference on Data Engineering*. Brisbane: IEEE, pp.1245–1248.
- Doyer, O.T., D’Haese, M.F.C., Kirsten, J.F., and Van Rooyen, C.J., 2007. Strategic

- focus areas and emerging trade arrangements in the South African agricultural industry since the demise of the marketing boards. *Agrekon*, 46(4), pp.494–513.
- Elmasri, R., and Navathe, S.B., 1989. *Fundamentals of database systems*. San Francisco, CA: The Benjamin/Cummings Publishing Company.
- Enders, W., 2010. *Applied econometric time series*. Hoboken, NJ: John Wiley.
- Engelbrecht, A.P., 2007. *Computational Intelligence: An introduction*. West Sussex: John Wiley.
- Fernandez, R.C., Migliavacca, M., Kalyvianaki, E., and Pietzuch, P., 2013. Integrating scale out and fault tolerance in stream processing using operator state management. In: *ACM SIGMOD International Conference on Management of Data*. pp.725–736.
- Galbraith, J.R., 1974. Organisation design: An information processing view. *Interfaces*, 4(3), pp.28–36.
- Gartner, 2015. *Flipping to digital leadership*. [online] Available at: [http://www.gartner.com/imagesrv/cio/pdf/cio\\_agenda\\_insights2015.pdf](http://www.gartner.com/imagesrv/cio/pdf/cio_agenda_insights2015.pdf) [Accessed 12 Aug. 2015].
- Geerts, G.L., 2011. A design science research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems*, 12, pp.142–151.
- Geyser, M., and Cutts, M., 2007. SAFEX maize price volatility scrutinised. *Agrekon*, 46(3), pp.291–305.
- Ghwanmeh, S., Mohammad, A., and Al-Ibrahim, A., 2013. Innovative artificial neural networks-based decision support system for heart diseases diagnosis. *Journal of Intelligent Learning Systems and Applications*, 5, pp.176–183.
- Goes, P.B., 2014. Big Data and IS research. *MIS Quarterly*, 38(3), pp.3–8.
- Gregor, S., and Hevner, A.R., 2013. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), pp.337–355.
- Gunelius, S., 2014. *The data explosion in 2014 minute by minute*. [online] Available at: <http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic> [Accessed 28 Mar. 2015].
- Guo, P., 2014. Clarifying human-computer interaction. *Communications of the ACM*, 57(2), pp.10–11.
- Gutierrez, L., Olmeo, M.G., and Piras, F., 2015. Forecasting wheat commodity prices using a Global Vector Autoregressive model. In: *Fourth AIEAA Conference*. Ancona.
- Hammond, J.S., Keeney, R.L., and Raiffa, H., 1999. *Smart choices: A practical guide to making better decisions*. Boston: Harvard Business School Press.

- Han, J., Kamber, M., and Pei, J., 2012. *Data mining: Concepts and techniques*. 3rd ed. Waltham: Elsevier.
- Hardoon, D.R., and Shmueli, G., 2013. *Getting started with business analytics: Insightful decision-making*. Boca Raton, FL: CRC Press.
- Headey, D., and Fan, S., 2008. Anatomy of a crisis: The causes and consequences of surging food prices. *Agricultural Economics*, 39(1), pp.375–391.
- Hevner, A., and Chatterjee, S., 2010. *Design research in information systems: Theory and practice*. New York: Springer.
- Hevner, A.R., 2007. A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), pp.87–92.
- Hevner, A.R., March, S.T., Park, J., and Ram, S., 2004. Design science in Information Systems research. *Management Information Systems*, 28(1), pp.75–105.
- Hodgkinson, G.P., and Starbuck, W.H., 2008. *The Oxford handbook of organisational decision making*. New York: Oxford University Press.
- Hofstee, E., 2006. *Constructing a good dissertation: A practical guide to finishing a Masters, MBA or PhD on schedule*. Johannesburg: EPE.
- Hull, J.C., 2012. *Options, futures and other derivatives*. 8th ed. Boston: Prentice Hall.
- IBM, 2011. *What is big data?* [online] Available at: [01.ibm.com/software/data/bigdata/what-is-big-data.html](http://01.ibm.com/software/data/bigdata/what-is-big-data.html) [Accessed 22 Mar. 2014].
- Iivari, J., 2007. A paradigmatic analysis of Information Systems as a design science. *Scandinavian Journal of Information Systems*, 19(2), pp.39–64.
- Irwin, S.H., Sanders, D.R., and Merrin, R.P., 2009. Devil or Angel? The role of speculation in the recent commodity price boom (and bust). *Journal of Agricultural and Applied Economics*, 41(2), pp.377–391.
- Jabjone, S., and Wannasang, S., 2014. Decision support system using artificial neural network to predict rice production in Phimai district, Thailand. *International Journal of Computer and Electrical Engineering*, 6(2), pp.162–166.
- Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., and Shahabi, C., 2014. Big Data and its technical challenges. *Communications of the ACM*, 57(7), pp.86–94.
- Janssen, M., Charalabidis, Y., and Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), pp.258–268.
- Johannesson, P., and Perjons, E., 2012. *A design science primer*. CreateSpace: Lexington.
- Jordaan, H., and Grové, B., 2010. Factors affecting forward pricing behaviour:

- Implications of alternative regression model specifications. *South African Journal of Economic and Management Sciences*, 13(2), pp.113–122.
- Jordaan, H., Grové, B., Jooste, A., and Alemu, Z.G., 2007. Measuring the price volatility of certain field crops in South Africa using the ARCH/GARCH approach. *Agrekon*, 46(3), pp.306–322.
- JSE, 2015. *JSE*. [online] Available at: <https://www.jse.co.za/trade/derivative-market/commodity-derivatives/agricultural-derivatives> [Accessed 15 Apr. 2015].
- Kaastra, I., and Boyd, M., 1996. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10, pp.215–236.
- Kabari, L.G., and Nwachukwu, E.O., 2013. Decision support system using decision tree and neural networks. *Computer Engineering and Intelligent Systems*, 4(7), pp.8–20.
- Kadadi, A., Agrawal, R., Nyamful, C., and Atiq, R., 2014. Challenges of data integration and interoperability in Big Data. In: *IEEE International Conference on Big Data*. pp.38–40.
- Khamis, A., Nabilah, S., and Binti, S., 2014. Forecasting wheat price using Backpropagation and NARX Neural Network. *The International Journal of Engineering and Science*, 3(11), pp.19–26.
- Khashei, M., and Bijari, M., 2010. An artificial neural network (p, d, q) model for time series forecasting. *Expert Systems with Applications*, 37(1), pp.479–489.
- Khashei, M., and Bijari, M., 2011. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing Journal*, 11(2), pp.2664–2675.
- Kouloumpis, E., Wilson, T., and Moore, J., 2011. Twitter sentiment analysis: The good, the bad and the OMG! In: *Fifth International AAAI Conference on Weblogs and Social Media*. pp.538–541.
- Kowalczyk, M., and Buxmann, P., 2014. Big Data and information processing in organisational decision processes. *Business & Information Systems Engineering*, 6(5), pp.267–278.
- Kriesel, D., 2007. *A brief introduction to Neural Networks*. [online] Available at: <http://www.dkriesel.com> [Accessed 13 Jul. 2015].
- Lantz, B., 2013. *Machine Learning with R*. Birmingham: Packt Publishing.
- Larose, D.T., 2005. *Discovering knowledge in data: An introduction to data mining*. 2nd ed. Hoboken, NJ: John Wiley.
- Lim, E., Chen, H., and Chen, G., 2013. Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems*, 3(4), pp.1–10.
- Loukides, M., 2010. *What is Data Science?* [online] Available at: <https://www.oreilly.com/ideas/what-is-data-science> [Accessed 16 Dec. 2014].

- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H., 2011. Big Data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- Mayer-Schonberger, V., and Cukier, K., 2013. *Big Data*. London: John Murray.
- McAfee, A., and Brynjolfsson, E., 2012. Big Data: The management revolution. *Harvard Business Review*, 90(10), pp.61–68.
- McLeod, J., 2012. Thoughts on the opportunities for records professionals of the open access, open data agenda. *Records Management Journal*, 22(2), pp.92–97.
- Minelli, M., Chambers, M., and Dhiraj, A., 2013. *Big Data, big analytics*. Hoboken, NJ: John Wiley.
- Mofokeng, M., and Vink, N., 2013. Factors affecting the hedging decision of maize farmers in Gauteng province. In: *Fourth International Conference of the African Association of Agricultural Economists*. Hammamet.
- Năstase, P., and Stoica, D., 2010. A new business dimension: Business analytics. *Accounting and Management Information Systems*, 9(4), pp.603–618.
- Nemati, H.R., Steiger, D.M., Iyer, L.S., and Herschel, R.T., 2002. Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems*, 33(2), pp.143–161.
- Niehaves, B., 2007. On epistemological diversity in design science: New vistas for a Design-Oriented IS Research? In: *Twenty-eighth International Conference on Information Systems*. Montreal, pp.1–13.
- O’Neil, C., and Schutt, R., 2014. *Doing Data Science*. Sebastopol: O’Reilly Media.
- Osman, A., El-Refaey, M., and Ayman, E., 2013. Towards real-time analytics in the cloud. In: *IEEE Ninth World Congress on Services*. pp.428–435.
- Pang, B., and Lee, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(2), pp.91–231.
- Patil, D.J., 2012. *Data Jujitsu*. Sebastopol: O’Reilly Media.
- Peffer, K., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S., 2008. A Design Science Research methodology for Information Systems research. *Journal of Management Information Systems*, 24(3), pp.45–77.
- Phillips-Wren, G., Mora, M., Forgyionne, G.A., and Gupta, J.N., 2009. An integrative evaluation framework for intelligent decision support systems. *European Journal of Operational Research*, 195, pp.642–652.
- Piccoli, G., 2012. *Information Systems for managers: Text and cases*. 2nd ed. Hoboken: John Wiley.
- Power, D.J., 2014. Using ‘Big Data’ for analytics and decision support. *Journal of*

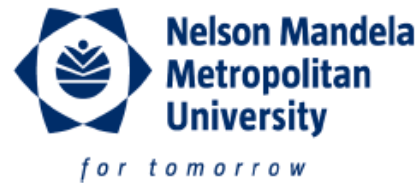
- Decision Systems*, 23(2), pp.222–228.
- Provost, F., and Fawcett, T., 2013a. Data Science and its relationship to Big Data and data-driven decision making. *Big Data*, 1(1), pp.51–59.
- Provost, F., and Fawcett, T., 2013b. *Data Science for business*. Sebastopol: O'Reilly Media.
- Qi, M., and Zhang, G.P., 2008. Trend time series modeling and forecasting with neural networks. *IEEE Transactions on Neural Networks*, 19(5), pp.808–816.
- Ravitch, S.M., and Riggan, M., 2012. *Reason and rigour: How conceptual frameworks guide research*. Thousand Oaks, CA: Sage Publications.
- Roberts, F.S., 2008. Computer science and decision theory. *Annals of Operations Research*, 163, pp.209–253.
- Rogers, P., and Blenko, M., 2006. The high-performance organization: Making good decisions and making them happen. *Handbook of Business Strategy*, 7(1), pp.133–142.
- Ruta, D., 2014. Automated trading with machine learning on Big Data. In: *2014 IEEE International Congress on Big Data*. Anchorage: IEEE.
- Sabherwal, R., and Becerra-Fernandez, I., 2011. *Business Intelligence: Practices, technologies and management*. Hoboken, NJ: John Wiley.
- SAGIS, 2015. *South African Grain Information Services*. [online] Available at: <http://www.sagis.org.za> [Accessed 2 Mar. 2015].
- SAP, 2015. *SAP HANA*. [online] Available at: [http://hana.sap.com/content/dam/website/saphana/en\\_us/abouthana/Top\\_10\\_Questions\\_for\\_Choosing\\_In-Memory\\_Databases.pdf](http://hana.sap.com/content/dam/website/saphana/en_us/abouthana/Top_10_Questions_for_Choosing_In-Memory_Databases.pdf) [Accessed 6 Apr. 2015].
- Saunders, M., Lewis, P., and Thornhill, A., 2009. *Research methods for business students*. 2nd ed. London: Pearson Education.
- Sauter, V.L., 2010. *Decision Support Systems for Business Intelligence*. 2nd ed. Hoboken, NJ: John Wiley.
- Segel, E., and Heer, J., 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), pp.1139–1148.
- Shields, P.M., and Rangarajan, N., 2013. *A playbook for research methods*. Stillwater, OK: New Forums Press.
- Shim, J.P., Warkentin, M., Courtney, J.F., and Power, D.J., 2002. Past, present, and future of decision support technology. *Decision Support Systems*, 33, pp.111–126.
- Shumway, R.H., and Stoffer, D.S., 2011. *Time series analysis and its applications*. 3rd ed. New York: Springer Science + Business Media.
- Simon, H.A., 1960. *The new science of management decision*. New York: Harper.



- Simon, H.A., 1979. Rational decision making in business organizations. *American Economic Association*, 69(4), pp.493–513.
- Simon, H.A., 1997. *Administrative behaviour: A study of decision-making processes in administrative organisations*. 4th ed. New York: The Free Press.
- Sojda, R.S., 2007. Empirical evaluation of decision support systems: Needs, definitions, potential methods, and an example pertaining to waterfowl management. *Environmental Modelling and Software*, 22(2), pp.269–277.
- Statssa, 2015. *Statistics South Africa*. [online] Available at: [http://www.statssa.gov.za/?page\\_id=750](http://www.statssa.gov.za/?page_id=750) [Accessed 16 Apr. 2015].
- Sundaram, R.K., and Das, S.R., 2011. *Derivatives: Principles and practice*. New York: McGraw-Hill/Irwin.
- Tambe, P.L., Hitt, L.M., and Brynjolfsson, E., 2012. The extroverted firm: How external information practices affect innovation and productivity. *Management Science*, 58(5), pp.843–859.
- Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operation Research Society*, 66(8), pp.1352–1362.
- Trostle, R., 2008. *Global agricultural supply and demand: Factors contributing to the recent increase in food commodity prices*. [online] Available at: <http://www.ers.usda.gov/publications/wrs0801> [Accessed 15 Mar. 2015].
- Tsadiras, A.K., Papadopoulos, C.T., and O’Kelly, M.E.J., 2013. An artificial neural network based decision support system for solving the buffer allocation problem in reliable production lines. *Computers & Industrial Engineering*, 66(4), pp.1150–1162.
- Tsay, R.S., 2010. *Analysis of financial time series*. Hoboken, NJ: John Wiley.
- USDA, 2015. *United States Department of Agriculture Economic Research Service*. [online] Available at: <http://www.ers.usda.gov/data-products/feed-grains-database/feed-grains-custom-query.aspx> [Accessed 22 May 2015].
- Vahidov, R., 2012. *Design-type research in Information Systems : Findings and practices*. Hershey: IGI Global.
- Vaishnavi, V.K., and Kuechler, W., 2015. *Design science research methods and patterns: Innovating information and communication technology*. 2nd ed. Boca Raton, FL: Taylor and Francis Group.
- Venable, J.R., 2006. The role of theory and theorising in design science research. In: *Design Science Research in Information Systems and Technology*. Claremont.
- Venter, M.M., Strydom, D.B., and Grové, B., 2013. Stochastic efficiency analysis of alternative basic grain marketing strategies. *Agrekon*, 52, pp.46–63.
- Watson, H.J., and Wixom, B.H., 2007. The current state of Business Intelligence. *Computer*, 40(9), pp.96–99.

- Wilamowski, B.M., 2009. Neural Network architectures and learning algorithms. *IEEE Industrial Electronic Magazine*, pp.56–63.
- Wiles, P.S., and Enke, D., 2014. Nonlinear modeling using Neural Networks for trading the soybean complex. *Procedia Computer Science*, 36, pp.234–239.
- Williams, G., 2011. *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. New York: Springer.
- Wright, B.D., 2011. The economics of grain price volatility. *Applied Economic Perspectives and Policy*, 33(1), pp.32–58.
- Wright, B.D., 2014. Data at our fingertips, myths in our minds: Recent grain price jumps as the perfect storm. *Australian Journal of Agricultural and Resources Economics*, 58, pp.538–553.
- Yu, S., and Ou, J., 2009. Forecasting model of agricultural products prices in wholesale markets based on combined BP neural network-time series model. In: *International Conference on Information Management, Innovation Management and Industrial Engineering*. Xian, pp.558–561.
- Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, pp.159–175.
- Zou, H.F., Xia, G.P., Yang, F.T., and Wang, H.Y., 2007. An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. *Neurocomputing*, 70, pp.2913–2923.

## Appendix A: Ethics Clearance



Faculty RTI Committee (Faculty of Science)  
Tel: +27 (0) 41 5042268  
E-mail: [lynette.roodt@nmmu.ac.za](mailto:lynette.roodt@nmmu.ac.za)

Ref: H14-SCI-CSS-12

Contact person: Mrs L Roodt

Date: 30 October 2014

Dear Mr K. Ayankoya

**TITLE OF PROJECT: A DATA SCIENCE FRAMEWORK FOR STRATEGIC DECISION SUPPORT IN SMALL AND MEDIUM SIZED ENTERPRISES**

PRP Prof Calitz K. Ayankoya  
PI: K. Ayankoya

Your above-entitled application was considered and approved by the Sub-Committee for Ethics in the Faculty of Science on 2 October 2014.

The Ethics clearance reference number is **H14-SCI-CSS-12** and is valid for three years. Please inform the Committee, via your faculty officer, if any changes (particularly in the methodology) occur during this time.

*An annual affirmation to the effect that the protocols in use are still those, for which approval was granted, will be required from you. You will be reminded timeously of this responsibility, and will receive the necessary documentation well in advance of any deadline.*

We wish you well with the project. Please inform your co-investigators of the outcome, and convey our best wishes.

Yours sincerely

A handwritten signature in black ink, appearing to read "Lynette Roodt".

Lynette Roodt  
Manager: Faculty Administrator  
Faculty of Science

## Appendix B: Invitation to participate in survey on Landbou.com

11/13/2015 Hoe verhandel jy graan? Help met navorsing | Landbou


TUIS NUUS BEDRYWE DIE WEER MARKTE KUNDIGES LEEFSTYL WINKEL WIELE

VIDEO'S OPSITKERS GEKLASSIFISEERD JOU FOTO PLASE

**Landbou**  
Boer vooruit.

Soekwoorde Soek Voermol is w Teken in Registreer

Landbou / Bedrywe / Hoe verhandel jy graan? Help met navorsing



Mielie  
Foto: www.falies.com.co.za

**Hoe verhandel jy graan? Help met navorsing**

Deur Jan Bezuidenhout  
21 Februarie 2015  
299 keer gelees

**Sal jy graag besluite met die bemerking van graan makliker wou neem? Navorsers het boere en graanhandelaars se hulp hiervoor nodig.**

'n Navorsingsprojek oor hoe Suid-Afrikaanse boere besluite vir die bemerking van graankommoditeite neem, word deur die departement rekenaarwetenskap van die Nelson Mandela Metropolitaanse Universiteit in Port Elizabeth onderneem.

Die oogmerk met die navorsing is die studie van die versameling, integrasie en ontleding van groot datastelle wat die bemerking van graankommoditeite beïnvloed, en hoe dit help om besluite deur graanboere in Suid-Afrika te verbeter.

Die navorsers het 'n webwerf geskep en vra boere en handelaars om 'n paar minute af te knyp om die vrae te beantwoord:

Boere kan [hier](#) klik om die vrae in te vul, en graanhandelaars moet [hier](#) klik om 'n soortgelyke vrae in te vul.

Hierdie studie is heeltemal anoniem en die NMMU sal die resultaat van die opname aan deelnemers stuur, wat dit versoek.

\* Rig navrae aan prof. Andre Calitz oby (041) 504 2639 of per e-pos by [andre.calitz@nmmu.ac.za](mailto:andre.calitz@nmmu.ac.za)

\* Wees die eerste van jou vriende wat hiervan hou

## Appendix C: Questionnaire for traders' survey on grain commodities trading

### 1. Section A

1.1 \* What best describes you?

1.2 \* How long have you functioned in the capacity selected in question 1.1 above?

1.3 \* Your gender.  Male  Female

1.4 \* Your age.  Less than 20 years  21 - 30 years  31 - 40 years  41 - 50 years  51 - 60 years  61+ years

1.5 Your highest education.

1.6 The province you are based in.

1.7 Please select the grain commodities that you deal with from these list.  
 White Maize  Yellow Maize  Wheat  Soybeans  Sunflower seed  Sorghum  
 Oats  Barley  Others

1.8 (If others, please specify)

1.9 What is your annual grain commodities trade volume (Tons)?

### 2. Section B - Factors influencing grain prices

Kindly select the option that best describe your opinion about each of the statements below.

2.1 Grain prices are volatile. strongly disagree      strongly agree

2.2 The US Dollar/Rand exchange rate affects grain prices in S.A. strongly disagree      strongly agree

2.3 The fluctuations in the price of crude oil affects the prices of grain commodities in S.A. strongly disagree      strongly agree

2.4 Local Interest rate affects grain prices in S.A. strongly disagree      strongly agree

2.5 The local weather conditions affects grain prices in S.A. strongly disagree      strongly agree

2.6 International weather conditions affects grain prices in S.A. strongly disagree      strongly agree

2.7 Consumer price index (CPI Inflation) affects grain prices in S.A. strongly disagree      strongly agree

2.8 Gross domestic product (GDP) affects grain prices in S.A. strongly disagree      strongly agree

2.9	The performance of Johannesburg Stock Exchange affects grain prices.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.10	Changes in the price of one grain commodity can influence the price of another grain commodity	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.11	Government policies affect grain prices.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.12	National stockpile of grains affects grain price in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.13	International stockpile of grains affects grain price in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.14	Prices of grains in other major producing countries affect grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.15	Grain production levels in SADC countries affects grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.16	The use of grains for biofuel in other countries affects grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.17	There is sufficient information for predicting grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
2.18	There is a need to collect and analyse more data on factors that influence grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree

### 3. Section C - Predicting grain prices

3.1	How do you currently determine the future price of grain commodities in S.A.	
3.2	Are there systems that currently provide you with forecast/prediction of grain prices in S.A.? If yes, please list them.	
3.3	How is the prediction of grain prices in S.A. of benefit to you?	
3.4	How does short term prediction benefit you?	
3.5	How does long term prediction benefit you?	
3.6	How do you determine short term predictions?	
3.7	How do you determine long term predictions?	
3.8	What type of information, intelligence or other predictions will help you improve your decision making for trading on grains in S.A.?	
3.9	What would you consider as a major gap in grain price prediction and futures in S.A.?	

#### 4. Section D - Decision support to farmers

4.1 Do you provide grain farmers with price predictions?  Yes  No

4.2 How often do you advise farmers?

4.3 What kind of commodity marketing advice do you provide for farmers?

4.4 What influences your pricing advice for a specific farmer?

4.5 What sort of information do you provide to farmers regarding grain prices?

#### 5. Section E - Feedback

5.1 If you want feedback of the result of this survey, please enter your email address.



## Appendix D: Questionnaire for farmers' survey on grain commodities trading

### 1. Section A

1.1 Are you a grain farmer in S.A.? If not, you need not continue with the rest of the survey.  Yes  No

1.2 How long have you been a grain farmer?

1.3 Your gender.  Male  Female

1.4 Your age.  Less than 20 years  21 - 30 years  31 - 40 years  41 - 50 years  51 - 60 years  61+ years

1.5 Your highest education.

1.6 The province you are based in.

1.7 Please select the grain commodities that you produce from these list.  White maize  Yello maize  Wheat  Soybeans  Sunflower seeds  Sorghum  Oats  Barley  Others

1.8 (If others, please specify)

1.9 What is your annual volume of grain commodities harvest (Tons)?

### 2. Section B - Factors influencing grain prices

Kindly select the option that best describe your opinion about each of the questions below.

2.1 Grain prices are volatile.  strongly disagree     strongly agree

2.2 The US Dollar/Rand exchange rate affects grain prices in S.A.  strongly disagree     strongly agree

2.3 The fluctuations in the price of crude oil affects the prices of grain commodities in S.A.  strongly disagree     strongly agree

2.4 Local Interest rate affects grain prices in S.A.  strongly disagree     strongly agree

2.5 The local weather conditions affects grain prices in S.A.  strongly disagree     strongly agree

2.6 International weather conditions affects grain prices in S.A.  strongly disagree     strongly agree

2.7 Consumer price index (CPI Inflation) affects grain prices in S.A.  strongly disagree     strongly agree

2.8 Gross domestic product (GDP) affects grain prices in S.A.  strongly disagree     strongly agree

2.9 The performance of Johannesburg Stock Exchange affects grain prices.  strongly disagree     strongly agree

2.10 Changes in the price of one grain commodity can influence the price of another grain commodity.  strongly disagree     strongly agree

2.11 Government policies affects grain prices.  strongly disagree     strongly agree

2.12 National stockpile of grains affects grain price in S.A.  strongly disagree     strongly agree

2.13 International stockpile of grains affects grain price in S.A.  strongly disagree     strongly agree

2.14 Prices of grains in other major producing countries affects grain prices in S.A.  strongly disagree     strongly agree

2.15 Grain production levels in SADC countries affects grain prices in S.A.  strongly disagree     strongly agree

2.16 The use of grains for biofuel in other countries affects grain prices in S.A.  strongly disagree     strongly agree

2.17 There is sufficient information for predicting grain prices in S.A.  strongly disagree     strongly agree

2.18 There is a need to collect and analyse more data on factors that influence grain prices in S.A.  strongly disagree     strongly agree



**3. Section C - Grain Marketing in S.A.**

3.1	I am aware that I can sell my grain commodities in the cash market after harvest (spot).	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.2	I am aware that I can sell my grain commodities using forward contracts.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.3	I am aware that I can sell my grain commodities using futures contract.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.4	I am aware that I can sell my grain commodities using put options.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.5	I understand the spot market option for selling grain commodities.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.6	I understand the forward contracts for selling grain commodities.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.7	I understand the futures contract for sell grain commodities.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.8	I understand the put options for sell grain commodities.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.9	It is better to sell grain commodities for cash after harvest than using other grain marketing strategies.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.10	It is better to use the forward contracts to sell grain commodities than using other grain marketing strategies.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.11	It is better to use futures contract to sell grain commodities than using other grain marketing strategies.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.12	It is better to use put options to sell grain commodities than using other grain marketing strategies.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.13	I market my grain commodities effectively.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.14	I would like to have access to a tool which helps me to predict grain prices in S.A.	strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> strongly agree
3.15	Which of the following grain marketing strategies do you use?	<input type="checkbox"/> Spot (at harvest) <input type="checkbox"/> Put option (after planting) <input type="checkbox"/> Put option (during pollination) <input type="checkbox"/> Futures market (multiple segment) <input type="checkbox"/> Futures market (during pollination only) <input type="checkbox"/> Forwards contracts
3.16	Which grain marketing strategy do you use the most?	<input type="text" value="(please select)"/>

**4. Section D - Forecasting grain prices**

4.1	How do you currently determine the future price of grain commodities in S.A.?	<input type="text"/>
4.2	Are there systems that currently provide you with forecast/prediction of grain prices in S.A.? If yes, please list them.	<input type="text"/>
4.3	What type of information or other predictions will help you improve your decision making for trading on grains in S.A.?	<input type="text"/>
4.4	What would you consider as a major gap in grain price prediction and futures in S.A.?	<input type="text"/>

**5. Section E - Feedback**

5.1	If you want feedback of the result of this survey, please enter your email address.	<input type="text"/>
-----	---	----------------------



## Appendix E: Permission from JSE to use data

### Ayankoya, Kayode (Mr) (s212400096)

---

**From:** Zintle Dastile  
**Sent:** 17 February 2015 12:00 PM  
**To:** Ayankoya, Kayode (Mr) (s212400096)  
**Cc:** Nomkhosi Magagula; Commodities@jse.co.za; INFO @ JSE  
**Subject:** RE: Request for grain commodities trade data - ref 798-849573  
**Attachments:** Total Tons Per Contract Per Day - Spot Deals.xls; Total Tons Per Contract Per Day.xls; Suns 1 (Mar 99 - May 05).xls; Suns 2 (Jun 05 - Jul 09).xls; Soya (Jun 08 - Jul 12).xls; Soya (May 2002 - May 2008).xls

Morning

The information you require is readily available on the JSE website but it does not date back to 1995. For both maize the data dates back to 1997, wheat was only introduced for trading in the commodities market in 1997, sunflower in 1999 and soyabean only in 2002. Please use the link below to access the spot price and the market trade data for the grain contracts. The JSE does not keep data relating to the demand and supply, export and import and stockpiling data of the grain contracts but you can access this information on the SAGIS website, I've provided the link to the website below. I have attached two reports which have data on the total number of tons delivered for each of the grain contracts, the data dates back to 2008 to date. I will attach the other outstanding files on a second mail. Please feel free to contact us, should you have further queries.

<https://www.jse.co.za/downloadable-files?RequestNode=/Safex/PriceHistory/Physically%20Settled%20Grain%20Contracts>  
<https://www.jse.co.za/downloadable-files?RequestNode=/Safex/PriceHistory/Spot%20Months%20on%20all%20Grain%20Products>  
<http://www.sagis.org.za/>

Regards

Zintle Dastile

Clearing & Settlement

Post-Trade & Information Services

T +2711 5207156

F +2711 5207558



JSE Limited · One Exchange Square · Gwen Lane · Sandown · South Africa

[www.jse.co.za](http://www.jse.co.za)



Johannesburg  
Stock Exchange

---

**From:** Krystal Reddy On Behalf Of INFO @ JSE  
**Sent:** 17 February 2015 09:46 AM  
**To:** Commodities@jse.co.za  
**Cc:** INFO @ JSE; Nomkhosi Magagula  
**Subject:** FW: Request for grain commodities trade data - ref 798-849573

Good day

Please can you assist with the below query.

## Appendix F: Sample of commodities futures transaction data



Johannesburg  
Stock Exchange

One Exchange Square,  
Gwen Lane,  
Sandown, South Africa  
Private Bag X991174  
Sandton 2146

Tel: +27 11 520 7000  
Fax: +27 11 520 8584

www.jse.co.za

Registration number: 2005/0229/  
VAT number: 4080119391

### COMMODITY DERIVATIVES MARKET

#### DOMESTIC FUTURES PRICES 07-Aug-2015

Contract	Change	Closing Bid	Closing Offer	MTM	VWAP	High	Low	Volume	Open Interest	Option Volatility
<b>WHITE MAIZE FUTURE</b>										
Aug-2015	-52.00	3065	3070	3073.00		3082.00	3050.20	507	711	0.00
Sep-2015	-63.00	3082	3085	3086.00		3118.00	3072.00	620	7792	27.00
Dec-2015	-67.00	3134	3136	3137.00	x	3171.20	3124.00	2968	16887	29.00
Mar-2016	-68.00	3090	3095	3094.00		3126.80	3082.00	675	9093	27.50
May-2016	-34.00	2685	2768	2688.00		2682.00	2680.00	5	174	28.50
Jul-2016	-5.00	2660	2668	2671.00		2669.00	2640.00	97	897	22.50
<b>YELLOW MAIZE FUTURE</b>										
Aug-2015	-25.00	2645	2650	2652.00		2650.00	2630.00	140	269	0.00
Sep-2015	-26.00	2657	2659	2662.00		2667.80	2643.00	366	6378	22.00
Dec-2015	-30.00	2684	2686	2690.00	v	2700.00	2671.00	1767	9650	20.00
Mar-2016	-23.00	2655	2660	2665.00		2679.00	2648.00	161	4408	21.00
May-2016	-35.00	2460	2499	2465.00		2460.00	2460.00	5	164	23.00
Jul-2016	-4.00	2405	2410	2415.00		2410.00	2395.00	32	787	22.00
<b>BREAD MILLING WHEAT</b>										
Aug-2015	5.00	4025	4044	4025.00		4015.00	4015.00	2	408	0.00
Sep-2015	-8.00	4040	4045	4040.00		4052.00	4040.00	217	7822	11.00
Dec-2015	-25.00	4018	4023	4022.00		4040.00	4018.00	178	7926	11.00
Mar-2016	-15.00	4079	4093	4093.00		0.00	0.00	0	431	0.00
<b>SUNFLOWER SEEDS FUTURE</b>										
Aug-2015	20.00	5585	5660	5600.00		5600.00	5560.00	32	149	0.00
Sep-2015	25.00	5605	5675	5625.00	v	5665.00	5610.00	265	2537	17.50
Dec-2015	50.00	5635	5665	5635.00		5645.00	5580.00	146	4162	17.25
Mar-2016	0.00	0	5400	5310.00		0.00	0.00	0	3	0.00
<b>SOYA FUTURE</b>										
Aug-2015	21.00	5049	5056	5056.00		5070.00	5036.00	104	169	0.00
Sep-2015	4.00	5085	5092	5092.00		5121.00	5070.00	558	4360	19.50
Dec-2015	1.00	5123	5140	5139.00		5147.00	5100.00	698	5790	21.00
Mar-2016	0.00	5113	5150	5136.00		0.00	0.00	0	180	21.00
May-2016	0.00	4925	4999	4999.00		0.00	0.00	0	25	0.00
<b>SOYBEAN CONTRACT</b>										
Sep-2015	-33.00	4494	4502	4502.00		0.00	0.00	0	252	0.00
Nov-2015	-43.00	4486	4494	4486.00		4479.00	4479.00	4	34	0.00
<b>CORN CONTRACT</b>										
Sep-2015	-29.00	1855	1859	1856.00		1856.00	1848.00	9	105	33.50
Dec-2015	-29.00	1937	1940	1938.00	v	1943.00	1928.00	579	3145	28.00
Jul-2016	-16.00	2126	2140	2138.00		2141.20	2138.20	7	263	0.00
<b>HARD RED WINTER WHEAT FUTURES</b>										
Dec-2015	-16.00	2423	2432	2423.00		2423.00	2423.00	2	0	0.00
<b>EURONEXT MILLING WHEAT CONTRACT</b>										
Sep-2015	25.00	2530	2543	2530.00		2530.20	2530.20	2	4	0.00
<b>SOYBEAN OIL CONTRACT</b>										
Dec-2015	0.00	0	0	8780.00		0.00	0.00	0	112	0.00
<b>SOFT RED WHEAT FUTURES</b>										
Sep-2015	7.00	2378	2387	2378.00		0.00	0.00	0	98	0.00
Dec-2015	8.00	2439	2449	2439.00		0.00	0.00	0	100	0.00
<b>SORGHUM FUTURES</b>										
Sep-2015	0.00	2990	3080	3050.00		0.00	0.00	0	31	0.00
Dec-2015	0.00	2920	0	3000.00		0.00	0.00	0	29	0.00
<b>WHEAT COMMODITY CANDO</b>										
Aug-2015	4.00	0	0	4014.00		0.00	0.00	0	400	0.00
<b>BRENT CRUDE OIL FUTURE</b>										
Sep-2015	0.00	0	0	650.20		0.00	0.00	0	-	0.00
Dec-2015	0.00	0	0	706.20		0.00	0.00	0	-	0.00
<b>DIESEL EUROPEAN GASOIL</b>										
Sep-2015	-0.02	0	0	5.04		0.00	0.00	0	-	0.00
Oct-2015	0.00	0	0	5.21		0.00	0.00	0	-	0.00
Nov-2015	0.00	0	0	5.28		0.00	0.00	0	-	0.00

## Appendix G: Sample of grain commodities spot transaction data

Source data

SPOT PRICE HISTORY								
DATE	WHITE	YELLOW	WEAT	\$/R	SUNS	SOYA	CORN	SORG
2007-01-02	1301	1550	1752	6.94	2410	2230		
2007-01-03	1280	1475	1720	6.87	2375	2180		
2007-01-04	1228	1478	1720	7.04	2375	2170		
2007-01-05	1268	1525	1744	7.15	2420	2170		
2007-01-08	1315	1585	1765	7.23	2448	2220		
2007-01-09	1280	1540	1763	7.18	2448	2245		
2007-01-10	1294	1546	1772	7.35	2460	2280		
2007-01-11	1295	1550	1770	7.31	2459	2265		
2007-01-12	1300	1570	1772	7.29	2490	2265		
2007-01-15	1370	1630	1792	7.21	2521	2305		
2007-01-16	1363	1665	1771	7.23	2500	2305		
2007-01-17	1366	1670	1760	7.24	2458	2290		
2007-01-18	1407	1630	1761	7.16	2458	2275		
2007-01-19	1416	1600	1762	7.16	2470	2300		
2007-01-22	1438	1605	1734	7.12	2418	2300		
2007-01-23	1415	1570	1742	7.12	2491	2300		
2007-01-24	1402	1590	1741	7.14	2450	2300		
2007-01-25	1385	1565	1734	7.15	2450	2300		
2007-01-26	1390	1575	1754	7.27	2450	2300		
2007-01-29	1380	1554	1747	7.32	2450	2300		
2007-01-30	1366	1559	1744	7.33	2450	2300		
2007-01-31	1346	1555	1752	7.31	2450	2300		
2007-02-01	1320	1520	1755	7.18	2458	2300		
2007-02-02	1305	1470	1760	7.16	2468	2300		
2007-02-05	1350	1472	1780	7.23	2502	2300		
2007-02-06	1376	1497	1782	7.20	2524	2300		
2007-02-07	1351	1469	1783	7.17	2512	2300		
2007-02-08	1380	1490	1819	7.18	2522	2300		
2007-02-09	1421	1515	1842	7.15	2540	2248		
2007-02-12	1510	1565	1886	7.24	2556	2300		
2007-02-13	1520	1595	1875	7.27	2585	2300		
2007-02-14	1520	1600	1863	7.19	2622	2300		
2007-02-15	1531	1620	1840	7.18	2638	2340		
2007-02-16	1586	1645	1823	7.15	2673	2419		
2007-02-19	1670	1700	1875	7.15	2725	2420		
2007-02-20	1676	1729	1840	7.12	2720	2430		
2007-02-21	1728	1755	1846	7.13	2700	2420		
2007-02-22	1791	1800	1891	7.11	2700	2465		
2007-02-23	1727	1759	1877	7.09	2685	2450		
2007-02-26	1772	1804	1901	7.08	2645	2430		
2007-02-27	1727	1785	1890	7.14	2643	2425		
2007-02-28	1745	1798	1916	7.23	2680	2445		
2007-03-01	1800	1820	1930	7.24	2713	2453		
2007-03-02	1848	1816	1929	7.29	2710	2441		
2007-03-05	1950	1876	1962	7.51	2755	2495		
2007-03-06	2030	1970	2005	7.51	2800	2533		
2007-03-07	1950	1975	1995	7.41	2749	2540		
2007-03-08	1937	1949	2003	7.38	2732	2494		
2007-03-09	1981	1988	2050	7.38	2732	2465		
2007-03-12	2012	1990	2090	7.31	2740	2520		
2007-03-13	1965	1970	2115	7.38	2750	2525		
2007-03-14	1940	1981	2140	7.50	2755	2553		

## Appendix H: Sample of Chicago Board of Trade transactions data

Symbol	Date	Open	High	Low	Close	Volume	OpenInt
AH	07-Jul-15	99.26	99.29	96.48	97.59	4162	20865
AH.C	07-Jul-15	98.2	98.2	96.6	97.8	4162	20865
AHH16	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHM16	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHU15	07-Jul-15	98.2	98.2	96.6	97.8	4162	20865
AHZ15	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHZ16	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHZ17	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHZ18	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AHZ19	07-Jul-15	99.1	99.1	97.8	97.8	0	0
AK	07-Jul-15	1.564	1.564	1.495	1.495	691	7317
AK.C	07-Jul-15	1.57	1.57	1.555	1.555	52	359
AKF16	07-Jul-15	1.562	1.562	1.547	1.547	0	378
AKF17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKF18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKG16	07-Jul-15	1.564	1.564	1.549	1.549	0	98
AKG17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKG18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKH16	07-Jul-15	1.57	1.57	1.555	1.555	5	359
AKH17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKH18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKJ16	07-Jul-15	1.575	1.575	1.56	1.56	5	247
AKJ17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKJ18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKK16	07-Jul-15	1.575	1.575	1.56	1.56	0	0
AKK17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKK18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKM16	07-Jul-15	1.575	1.575	1.56	1.56	0	0
AKM17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKM18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKN16	07-Jul-15	1.578	1.578	1.563	1.563	0	1
AKN17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKN18	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKQ15	07-Jul-15	1.654	1.654	1.636	1.636	289	2427
AKQ16	07-Jul-15	1.578	1.578	1.563	1.563	0	0
AKQ17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKU15	07-Jul-15	1.636	1.636	1.617	1.617	188	1057
AKU16	07-Jul-15	1.578	1.578	1.563	1.563	0	0
AKU17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKV15	07-Jul-15	1.613	1.613	1.596	1.596	85	536
AKV16	07-Jul-15	1.578	1.578	1.563	1.563	0	0
AKV17	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKX15	07-Jul-15	1.595	1.595	1.579	1.579	11	493
AKX16	07-Jul-15	1.578	1.578	1.563	1.563	0	0
AKX17	07-Jul-15	1.588	1.588	1.573	1.573	0	0
AKZ15	07-Jul-15	1.581	1.581	1.566	1.566	108	1721
AKZ16	07-Jul-15	1.58	1.58	1.565	1.565	0	0
AKZ17	07-Jul-15	1.588	1.588	1.573	1.573	0	0
BO	07-Jul-15	32.74	32.74	31.41	31.41	192096	365362



# Appendix I: Sample of grain commodities demand and supply data for South Africa

SAGIS South African Grain Information Service vnc Suid-Afrikaanse Graaninligtingsdiens vnc 1995-0000000	MAIZE / MIELES Monthly announcement of data / Maandelikse bekendmaking van data (I) 2019/16 Year (May - Apr) / 2015/16 Jaar (Me - Apr) (2) ton													SMD-072015
	May/Mei 2015			Jun 2015			Progressive/Progressief			% +/- (3)	Progressive/Progressief			2015-07-24
	White Wit	Yellow Geel	Total Totaal	White Wit	Yellow Geel	Total Totaal	White Wit	Yellow Geel	Total Totaal		White Wit	Yellow Geel	Total Totaal	
	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	1 Jun 2015	1 May/Me 2015	
(a) Opening stock	1 282 581	791 054	2 073 635	1 731 260	1 602 596	3 333 856	1 282 581	791 054	2 073 635	252,0	274 318	314 710	589 028	(a) Beginvoorraad
(b) Acquisition	836 072	1 317 500	2 153 572	1 765 592	1 803 259	3 568 851	2 621 694	3 120 559	5 742 253	-16,9	3 430 061	3 482 381	6 912 442	(b) Verkryging
Deliveries directly from farms (i)	836 072	1 284 242	2 140 314	1 765 429	1 794 486	3 559 915	2 621 591	3 040 768	5 662 359	-18,1	3 430 061	3 482 381	6 912 442	Lewerings direk vanaf plaas (i)
Imports destined for RSA	0	33 058	33 058	163	46 763	46 926	163	79 791	80 014	100,0	0	0	0	invoere bestem vir RSA
(c) Utilisation	364 016	480 411	844 427	362 073	493 929	856 002	726 089	974 940	1 700 429	4,4	861 632	767 462	1 629 094	(c) Aanwending
Processed for the local market	361 657	480 246	823 903	360 400	477 446	837 846	722 057	939 692	1 661 749	4,3	856 339	736 720	1 593 059	Verwerk vir die binnelandse mark
Human consumption (ii)	343 328	461 118	784 446	348 672	418 662	767 334	692 000	87 980	779 980	-1,7	706 541	867 752	1 574 293	Menslike verbruik (ii)
Animal feed/Industrial	16 605	415 130	431 735	9 863	434 052	443 915	26 488	846 162	876 670	10,6	143 714	643 321	792 035	Diervoer/Industrieel
Grinding	1 724	596	2 320	1 945	1 302	3 247	3 999	2 333	6 099	21,1	6 094	1 647	7 741	Malderij
BioFuel	0	0	0	0	0	0	0	0	0	0,0	0	0	0	Biobrandstof
Withdrawn by producers	1 302	4 058	5 360	914	4 243	5 157	2 216	8 301	10 517	4,1	2 191	7 911	10 102	Ontrek deur produsente
Released to end-consumers	1 057	14 107	15 164	759	12 240	12 999	1 816	26 547	28 163	8,5	3 102	22 861	25 963	Vrygestel aan eindverbruikers
(d) RSA Exports (5)	45 053	26 530	71 583	38 030	24 834	62 864	83 083	51 394	134 447	-63,1	108 751	255 763	364 514	(d) RSA Uitvoere (5)
Products (i)	7 154	9 789	16 943	6 980	9 944	16 924	14 194	18 743	32 937	-4,9	17 017	17 587	34 604	Produkke (i)
African countries	7 038	5 018	12 056	6 491	5 629	12 120	13 520	10 647	24 176	-2,0	14 421	10 246	24 667	Afrika lande
Other countries	116	4 781	4 897	489	3 315	3 804	6 005	8 096	8 701	-12,3	2 596	7 321	9 917	Andere lande
Whole maize	37 899	16 731	54 630	31 050	15 990	46 940	68 946	32 621	101 570	-69,2	91 734	238 196	329 930	Heelmale
Border posts	37 899	16 731	54 630	31 050	14 236	45 286	68 946	30 987	99 936	-16,9	91 734	26 594	120 288	Grensoorte
Warehouses	0	0	0	0	1 654	1 654	0	1 654	1 654	-99,2	0	209 942	209 942	Waarhuise
(e) Sundries	-1 676	-1 153	-2 829	1 622	-6 645	-5 023	-54	-7 769	-7 823		3 986	142	4 128	(e) Divers
Net dispatches(+)/receipts(-)	-2 030	2 581	551	3 336	-569	2 767	1 906	1 622	3 528		73	-4 068	-4 023	Netto versendinge(+)/ontvangsies(-)
Surpluses(+)/Deficits(-)	354	-3 734	-3 380	-2 314	-5 686	-8 000	-1 900	-9 420	-11 380		3 913	4 240	8 153	Surplusse(+)/Tekorte(-)
(f) Unutilised stock (arb-o-d-e)	1 731 260	1 602 596	3 333 856	3 095 127	2 893 707	5 988 834	3 095 127	2 893 707	5 988 834	8,8	2 730 010	2 774 194	5 504 204	(f) Onaanwende voorraad (arb-o-d-e)
(g) Stock stored at: (f)	1 731 260	1 602 596	3 333 856	3 095 127	2 893 707	5 988 834	3 095 127	2 893 707	5 988 834	8,8	2 730 010	2 774 194	5 504 204	(g) Voorraad geborg by: (f)
Stores and traders	1 491 830	1 394 344	2 886 174	2 662 862	2 363 585	5 026 447	2 662 862	2 363 585	5 026 447	7,3	2 366 787	2 552 235	4 918 992	Opberbers en handelaars
Processors	239 430	208 252	447 682	412 236	310 122	722 387	412 236	310 122	722 387	21,4	373 223	221 959	585 212	Verwerkers
(h) Imports destined for exports not included in the above information	0	0	0	0	0	0	0	0	0		0	0	0	(h) Invoere bestem vir uitvoer nie ingesluit in inligting hierbo nie
Opening stock	0	0	0	0	0	0	0	0	0		0	0	0	Beginvoorraad
Imported	0	0	0	0	0	0	0	0	0		0	0	0	Ingevoer
Exported - Whole maize	0	0	0	0	0	0	0	0	0		0	0	0	Uitgevoer - Heelmale
Exported - Products	0	0	0	0	0	0	0	0	0		0	0	0	Uitgevoer - Produkke
Stock surplus(+)/deficit(-)	0	0	0	0	0	0	0	0	0		0	0	0	Voorraad surplus(+)/tekort(-)
Closing stock	0	0	0	0	0	0	0	0	0		0	0	0	Eindvoorraad
Producer deliveries directly from farms (ton) White/Wit (i) Yellow/Geel (ii) Producers' deliveries direct vanaf plaas (ton) May/Me 2015 65 004 132 693 May/Me 2015 Apr 2015 109 832 234 427 Apr 2015 May - June 2015 2 621 501 3 040 708 May - June 2015														
Maize equivalent (i) Millie ekwivalent Processed for drinkable alcohol included (ii) Verwerk vir drinkbare alkohol ingesluit Also refer to general footnotes. Verwys ook na algemene voetnote.														

## Appendix J: Sample of grain commodities demand and supply data for USA

Table 4--Corn: Supply and disappearance (million bushels)

Mkt year and qtr 1/	Supply				Disappearance						Ending stocks	
	Beginning stocks	Production	Imports	Total supply 2/	Domestic use				Exports	Total disappearance 2/		
					Food, alcohol, and industrial use	Seed use	Feed and residual use	Total domestic use 2/				
2009/10	Q1 Sep-Nov	1 673	13 067	0.98	14 741	1 382		1 990	3 372	467	3 839	10 902
	Q2 Dec-Feb	10 902		1.32	10 904	1 447		1 341	2 788	422	3 210	7 694
	Q3 Mar-May	7 694		3.13	7 697	1 543	21.68	1 273	2 838	549	3 387	4 310
	Q4 Jun-Aug	4 310		2.91	4 313	1 566	0.65	496	2 063	542	2 605	1 708
	MY Sep-Aug	1 673	13 067	8.34	14 749	5 939	22.34	5 101	11 082	1 979	13 041	1 708
2010/11	Q1 Sep-Nov	1 708	12 425	5.32	14 138	1 582		2 047	3 630	452	4 082	10 057
	Q2 Dec-Feb	10 057		8.46	10 065	1 577		1 562	3 139	403	3 542	6 523
	Q3 Mar-May	6 523		10.38	6 534	1 618	20.24	715	2 353	510	2 863	3 670
	Q4 Jun-Aug	3 670		3.50	3 674	1 625	2.76	452	2 080	466	2 546	1 128
	MY Sep-Aug	1 708	12 425	27.67	14 161	6 403	23.00	4 777	11 202	1 831	13 033	1 128
2011/12	Q1 Sep-Nov	1 128	12 314	4.06	13 446	1 611		1 782	3 393	406	3 799	9 647
	Q2 Dec-Feb	9 647		3.93	9 651	1 636		1 547	3 183	444	3 627	6 023
	Q3 Mar-May	6 023		10.67	6 034	1 601	23.57	861	2 486	400	2 886	3 148
	Q4 Jun-Aug	3 148		10.71	3 159	1 548	0.96	330	1 879	291	2 170	989
	MY Sep-Aug	1 128	12 314	29.37	13 471	6 396	24.53	4 520	10 941	1 541	12 482	989
2012/13	Q1 Sep-Nov	989	10 755	34.79	11 779	1 466		2 060	3 525	221	3 746	8 033
	Q2 Dec-Feb	8 033		45.43	8 078	1 430		1 087	2 517	161	2 678	5 400
	Q3 Mar-May	5 400		40.18	5 440	1 545	22.37	921	2 488	186	2 674	2 766
	Q4 Jun-Aug	2 766		39.56	2 806	1 573	2.22	247	1 822	162	1 985	821
	MY Sep-Aug	989	10 755	159.95	11 904	6 013	24.58	4 315	10 353	730	11 083	821
2013/14	Q1 Sep-Nov	821	13 829	14.52	14 666	1 550		2 312	3 862	350	4 212	10 453
	Q2 Dec-Feb	10 453		6.58	10 459	1 607		1 451	3 058	393	3 451	7 008
	Q3 Mar-May	7 008		8.56	7 017	1 646	21.92	859	2 528	637	3 165	3 852
	Q4 Jun-Aug	3 852		6.12	3 858	1 677	1.08	411	2 089	537	2 626	1 232
	MY Sep-Aug	821	13 829	35.79	14 666	6 480	23.00	5 034	11 537	1 917	13 454	1 232
2014/15	Q1 Sep-Nov	1 232	14 216	5.01	15 452	1 610		2 223	3 833	408	4 241	11 211
	Q2 Dec-Feb	11 211		5.89	11 217	1 624		1 445	3 069	404	3 472	7 745
	MY Sep-Aug	1 232	14 216	25.00	15 472	6 498	23.22	5 250	11 772	1 825	13 597	1 876
2015/16	MY Sep-Aug	1 876	13 630	25.00	15 531	6 537	22.90	5 300	11 880	1 900	13 780	1 771

1/ September-August. Latest data may be preliminary or projected.

2/ Total may not add due to rounding.

Source: USDA, World Agricultural Outlook Board, World Agricultural Supply and Demand Estimates and supporting materials.

Date run: 6/11/2016

# Intrinsic Relations between Data Science, Big Data, Business Analytics and Datafication

Kayode Ayankoya  
Department of Computing Sciences  
NMMU  
P.O. BOX 77000  
Tel No: +27 41 504 2088  
S212400096@nmmu.ac.za

Andre Calitz  
Department of Computing Sciences  
NMMU  
P.O. BOX 77000  
Tel No: +27 41 504 2639  
Andre.calitz@nmmu.ac.za

Jean Greyling  
Department of Computing Sciences  
NMMU  
P.O. BOX 77000  
Tel No: +27 41 504 2081  
Jean.greyling.nmmu.ac.za

## ABSTRACT

Data recording and storage have evolved over the past decades from manual gathering of data by using simple writing materials to the automation of data collection. Data storage has evolved significantly in the past decades and today databases no longer suffice as the only medium for the storage and management of data. This is due to the emergence of the Big Data and Data Science concepts. Previous studies have indicated that the multiplication of processing power of computers and the availability of larger data storage at reduced cost are part of the catalysts for the volume and rate at which data is now made available and captured.

In this paper, the concepts of Big Data, Data Science and Business Analytics are reviewed. This paper discusses datafication of different aspects of life as the fundamental concept behind the growth of Big Data and Data Science. A review of the characteristics and value of Big Data and Data Science suggests that these emerging concepts will bring a paradigm change to a number of areas. Big Data was described as the basis for Data Science and Business Analytics which are tools employed in Data Science. Because these fields are still developing, there are diverse opinions, especially on the definition of Data Science. This paper provides a revised definition of Data Science, based on the review of available literature and proposes a schematic representation of the concepts.

## Categories and Subject Descriptors

H.2.5 Heterogeneous Databases

H.2.8 Database Applications

E.1 Data Structures

## General Terms

Management, Documentation, Performance, Theory.

## Keywords

Data Science, Big Data, Business Intelligence, Business

*Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.*

SAICSIT2014, September 29 - October 01 2014, Centurion, South Africa  
Copyright 2014 ACM 978-1-4503-3246-0/14/09...\$15.00  
<http://dx.doi.org/10.1145/2664591.2664619>

Analytics, Datafication.

## 1. INTRODUCTION

Data can be defined as a representation of facts that can be collected, recorded and used as a basis for decision making [15, 8]. Data is further described as a representation of the real world [16]. Although, Flores [16] took that position three decades ago, the ubiquitous nature of data today attests to this fact more than ever before. Therefore data forms the basis of collecting, organising and describing facts and information. These facts and information are entwined into the daily lives of individuals. Possession of facts is the centre of business practices and the foundation of academic research.

The study of Computer Science (CS) and Information Systems (IS) has grown significantly over the years. The application of these fields has gained importance and is a vital developmental factor for other fields of study. The application of computers and information management systems is vital in the development of critical areas such as aviation, banking, finance, business management, education, etc. Furthermore, the use of computers and information management technologies now determines how people play and relate through the use of Social Media platforms and games. Thus, the fields, CS and IS could be described as a major catalyst for accelerating growth and development in other fields of research and practice. This is because the application of CS and IS enables the collection, storage, access, analysis, interpretation and management of large numbers of facts and amounts of data and information in a fraction of a second, which is far beyond manual capacity and operation. Hence data remains one of the most important pillars of CS and IS research and practice.

The impact and influence of data in practice and research are far-reaching. In most cases, data forms the basis of decision making within organisations and is the basis of scientific inference where research is concerned. In actual fact, some organisations consider data as one of their most important assets [3, 12]. This might be irrespective of the quality or the quantity of data available to such organisations. However, in the past number of years, data has become globally available and the amount of data generated has increased significantly [24]. As a result, the expectancy of practitioners and researchers about data is intensifying and the concepts of Big Data, Data Science and Business Analytics are emerging among CS, IS and several other stakeholders. The new concepts have been described as having challenges and opportunities that will affect productivity, profitability, efficiency

and are a complete paradigm shift in the way many things are seen and done [24, 25]. There is, however, fragmented opinion about Big Data, Data Science and other concepts such as Business Analytics that are data-centric based because these concepts are evolving.

'Datafication' has been described as the foundation of the emerging concepts of Big Data, Data Science and the new potentials of Business Analytics [9, 23]. The explanation of this phenomenon is how information is being discovered in every area of life and rendered into data formats [9]. This concept aligns with and amplifies the full perspective of Flores [16] on the definition of data and explains why more data has become more available worldwide.

The aim of this paper is to provide a review of the concepts of Big Data, Data Science and Business Analytics. Furthermore, this paper will address the fragmented opinion about the relationship between the concepts and their implications for business. In order to achieve this, a critical literature review research methodology will be adopted. This will be done by examining the literature (Section 2) about each of the concepts and the relationships among them. A critical analysis and an improved schematic representation of the concepts is presented in Section 3. Opportunities for future research will be presented in the final section of this paper. Thereafter, the concluding section suggests the relationship that exist among the concepts and a definition for Data Science.

## 2. OVERVIEW OF CONCEPTS

### 2.1 Datafication

Presently, humans have the capability to collect as much data as possible from different aspects of life. Almost every human is now creating data daily [26]. Some of the areas that data is now being collected from are totally unconventional and would not be seen as sources of useful information a few years ago. This is made possible by the 'Datafication' of different aspects of life. Datafication has been described as the ability to "take all aspects of life and turn them into data" [9]. However, it is important to note that this is much more than converting existing analogue information such as books and photographs into digital formats.

An example of datafication is how social media has datafied friendships, relationships, thought and daily activities. Another example is Google's augmented-reality glasses that are described as the datafication of gaze [9]. With the increased adoption of mobile devices, it is likely that datafication of different aspects of life will only keep on increasing. There are mobile applications that are now able to datafy human movements thereby providing data and subsequent insight into weight loss and health issues. This ability to datafy different aspects of individual lives and activities within organisations is what forms the foundation of the Big Data and Data Science concepts [23, 27].

### 2.2 Big Data

The datafication of different aspects of life and more activities within the daily running of an organisation are responsible for the creation of a large volume of data [26]. The challenges and opportunities presented by this exponential increase in the amount of data started receiving attention from practitioners and academia more than a decade ago [18]. Today, the volume of data available globally has grown significantly and the rate of growth is increasing by the minute. A report from IBM in 2011 indicated

that 90% of the world's data had been created in the previous two years [19]. This deluge of data generated is now described as Big Data and has changed the way people live and how organisations operate. It has become a defining factor on how business is conducted and the impact is also being experienced in the fields of science and academic research [24, 4].

Traditionally, organisations have collected data resulting from business transactions and internal operations such as sales, marketing, finance, production and human resources management. The collection of organisational related data is as a result of the need to automate processes and systems by using information technology and the accompanying tools to simplify their internal systems and provide better service to customers. Organisations soon found the need to integrate the various sources of data into a single repository for the purpose of extracting knowledge and information [29]. This paradigm formed the basis for data warehousing, which involved the aggregation of data from different systems into a single source of intelligence. Data warehousing evolved with techniques, technologies and approaches that enabled organisations to find answers to questions on "what happened?" and "what is happening?" mostly. Over the time, the use of analytics in data warehousing also assisted organisations to find answers to "how and why did it happen?" [10, 12].

The growth in the amount of data within organisations and also external and unstructured data has amplified the value of learning from data and the use of data for supporting decision making processes. Organisations are now able to predict what will happen in the future and have evidence based scenario planning of future occurrences. These possibilities introduce a new paradigm into the use of data for decision making. Hence, the need for an understanding of the concepts and how they are related.

Over the years, theories have emerged that support and explain how organisations can derive value from information [17, 2, 30]. These continue to be relevant as practitioners and academic researchers continue to link the availability and use of data in organisations to improve decision making as well as to increase performance and profitability [3, 24].

The definition of Big Data has been based on its complexities, sources, storage and management. Big Data was described in earlier studies according to the volume of data created, the velocity of data created and the variety of data that make up Big Data [24, 26]. However, recent studies have now included veracity and value as part of what characterises big data [25].

The five main characteristics that have been used to describe Big Data bring the complexities, opportunities and risks associated with Big Data into perspective. According to Manyika, et al. [24], Big Data is best described by the fact that it is always very large in volume, the rate of creation is fast and the sources are so extremely diverse that it is beyond the ability and capacity of traditional tools, processes and management practices. Identified characteristics of big data include the following:

**Volume:** The ability to generate and collect data has increased across various spectra. A terabyte of storage space for data seemed extravagant for data storage previously, but it is estimated that Walmart generates about 2.5 petabytes of data every hour [26]. This is an equivalent of 2500 terabytes of data and that only from one of many organisations generating the same volumes. This is an indication of the volume of data that has now become available and is used to describe Big Data.

**Velocity:** The rate at which data is being created is a characteristic that redefines data. In 2012, it was estimated that 278,000 messages were sent on Twitter every minute. Within the same time frame, Facebook generates about 2.5 million posts and 1.8 million likes that amount to an average of 350 gigabytes of data and about 72 hours of video are uploaded on YouTube. This suggests that data should no longer be seen and managed from the “warehouse”. Rather, it is important to recognise the fluidity of Big Data for effective management and optimising its value [11].

**Variety:** The sources and types of Big Data that are relevant are heterogeneous; they include video, audio, images, GPS coordinates for mobile device applications, etc. The datafication of different aspects of life continues to lead to the creation of different types of data, namely structured and unstructured data that make up the Big Data phenomenon. Some organisations might be able to generate large volumes of data internally that fit the traditional description of data. In many cases, Big Data will be from different sources and will comprise different types of data, but, the ability to establish relationships among the variety of fragmented data will provide even more benefits [25].

**Veracity:** The uncleanness and inaccuracy of Big Data is as a result of the other characteristics. An IBM survey indicates that 27% of respondents were not sure of the accuracy of the data used for decision making and more than 30% use data that they do not completely trust [19]. When considering Big Data as a source of truth, it is important to consider that the data in most cases might require cleaning and there is a risk of incompleteness and inconsistency in the data.

**Value:** The purpose of Big Data in any field will be to extract value and use it as a source of truth and actionable insight [25]. However, with the characteristics of Big Data, the ability to extract value from Big Data is not automatic. This implies that a dataset that is relevant and important for one organisation might be irrelevant to another. Therefore, the mix of the right dataset, based on the purpose for which it is used will determine its value.

The characteristics discussed above define the opportunities to acquire insight, support decision making, predict the future and facilitate organisational learning from Big Data. Considering these characteristics, practitioners and researchers might be faced with a decision either to use samples drawn from a population or an entire dataset available for research and decision making. This is because large volumes of data are more trustworthy [23]. The fluidity of the (big) data available could also mean that decisions and research based on samples might become obsolete within a shorter space of time.

It is important to recognise the fluidity of Big Data and that the insight from it might be dynamic and rapidly changing because of the volume and velocity of the data [11]. Therefore, giving attention to Big Data may not be optional when seen as a tool for predicting the future, the foundation for the emerging digital economy and innovation [11]. Research has shown that organisations that are paying attention to use of Big Data are significantly more productive and profitable than others [3].

Big Data opens up new opportunities and ways of managing organisations. The evolution of Big Data and other supporting concepts have become the foundation of a new type of organisation. These organisations are replacing some of their decision making efforts by humans with algorithms that analyse

Big Data. It introduces the use of advanced analytics for system-made decisions as a replacement for intuition and experience-driven decision making [26]. This suggests a strategic and operational implication for businesses that is beyond the techniques and technologies that Big Data introduce.

Hence, the implementation of Big Data is not just an IT function, but should be considered from a strategic viewpoint and would require the involvement of management and top leadership of organisations [11]. On the other hand, the evolution of Big Data has also impacted the associated concepts of business intelligence and analytics [6]. It has been noted that the unique characteristics of Big Data are determining how business intelligence and analytics are used and the context in which they are used [6, 20]. Overall, it can be inferred that there is a need for more research focused on how Big Data can benefit different areas both in practice and in academia.

### 2.3 Business Intelligence and Analytics

Business intelligence and analytics are data-centric approaches that complement data with a set of methodologies, processes, technologies and tools for analysing and extracting information from data [12, 6, 20]. Business intelligence has been the focus of much earlier attention as sets of methodologies and processes. It was used as an enhancement of relational databases for business support and reporting. However, the introduction of business analytics has provided opportunities for the application of analytical techniques that allow for data-driven decision making and management of organisations [12, 5]. While the focus of intelligence is more on the provision of dynamic access and reporting of data, analytics offer an opportunity for extracting knowledge and insight from data [12, 31].

Business intelligence and business analytics have recently become more complementary and synchronised and are addressed together as Business Intelligence and Analytics (BI&A) [6, 20]. Overall, the two have become a prominent and almost indispensable set of tools in data management and decision making. Moreover, Business Intelligence and Analytics have evolved and are influenced by the development of different types, sources and volume of data. This evolution is described as starting with the use of BI&A for structured DBMS-based content [6]. Thereafter, BI&A is being used for unstructured content such as text and web-based content. Lately, it has evolved into being used for user-generated content on social media, data created by mobile devices, mobile application and data created by sensors. On-Line Analytical Processing (OLAP) provides analytical functions for extracting information and knowledge with the data warehousing framework which could be a part of structured Big Data [29]. But the latest developments in the nature, type and management of data have necessitated the use of more advanced analytical tools.

BI&A now has the capability to absorb and be used alongside Big Data. With the growth of Big Data, emerging research on the convergence between Big Data and BI&A shows five different categories of implementation of the concepts and areas of suggested further research [6, 20]. These include:

- Structured data analytics;
- Text analytics;
- Web analytics;
- Network analytics and
- Mobile analytics.

Each of the categories mentioned above provides opportunities for extracting actionable insight from data by using analytical processes and tools. All the categories have sources of data that form a part of the Big Data framework and definition. Previously, BI&A was used to report and understand what happened in the past [6], but the volume and fluidity of Big Data presents an opportunity to use Big Data and analytical tools to predict the future and make new discoveries [29].

## 2.4 Data Science

Emerging research in Big Data and BI&A suggests a convergence that creates a source of information and a business decision support system that cannot be ignored [6], however it has a large number of complexities. Database management and business intelligence are known to primarily belong to the field of Computer Science and Information Systems. However, computer scientists, information system researchers and practitioners have been using statistical skills to identify “what happened” from historical data. Big Data has introduced a completely new paradigm that combines the skills of a computer scientist, advanced analytical/statistical skills and knowledge of domain area for extracting value from Big Data [3, 13, 6]. This is made possible by the uniqueness of Big Data, where types of interlinked data are rapidly generated from various sources and are available for decision making.

Data Science combine the opportunities and the potentials of Big Data, BI&A, advanced analytics and the understanding of a particular field to extract value from large volumes of data [6]. But Data Science is much more than the application of statistical analysis on Big Data [21]. In an attempt to define or explain what Data Science is, a variety of perspectives have emerged. Below is a list indicating how Data Science has been viewed and defined by different authors:

- Data Science is a holistic approach to extract value from data and enables the creation of data products [21, 22].
- Data Science is the study of generalizable extraction of knowledge from data with emphasis on prediction [14].
- Data Science involves the combination of Computer Science, Data Visualisation, Machine Learning, Statistics, Mathematics, Communication and Domain Expertise skills for extracting meaning from complex data. It results in data products, data -driven decision making and the provision of answers to solve real world problems [27].
- Data Science involves principles, processes and techniques for understanding phenomena and improved decision making [29].
- Data Science involves making discoveries and providing answers to difficult questions by using Big Data [13].

Presently there is no standard or universally accepted definition of Data Science. The perspectives presented above describe Data Science based on what can be achieved with it, who is a Data Scientist, the process of Data Science as well as the tools and principles of Data Science. Although there is a fragmented approach to describing Data Science, the different perspectives indicate that Data Science offers opportunities for data-driven decision making, predictions, discoveries, recommendations and a different approach to providing solutions both in research and in practice.

Data warehousing has supported the decision making in organisations since the 1990s, but the Big Data paradigm introduces a new dimension to the use of data for improved decision making in business or in scientific research [14, 10]. Data Science provides a systematic approach that takes into consideration the characteristics of Big Data to extract and organise data. Then it makes use of tools such as predictive analysis and machine learning to provide answers to carefully asked questions and provides solutions to problems that are defined unconventionally [28, 14, 29]. The Data Scientist requires answers to questions and the ability to carefully extract these answers from data. This is why researchers and practitioners are focusing on Big Data for redefining decision support and processes.

Traditional statistical analysis focuses on building models for trial and error purposes, hypothesis testing and estimations that are based on collected samples [7]. The resulting models and outcomes of this statistical thinking have levels of uncertainty that make inferences not completely dependable [7]. However, with the paradigm shift in the nature and sources of data to Big Data, attention is moving from the traditional statistical approach. Data science takes advantage of the characteristics of Big Data and the use of advanced analytics such as predictive analysis, modelling and machine learning for prediction, recommendation and discovery [7, 10].

The foundation of machine learning is in the use of computer science together with mathematics, statistics and other forms of analytical science. It is described as a branch of artificial intelligence used to extract the unknown from known data [7, 1]. The outcome might also be an inference, as with the use of traditional statistical tools [7], but machine learning, further helps to extract patterns from different types and large volumes of data that can be used to predict future occurrences [1].

Predictive analysis and modeling are a subset of machine learning that also make use of known data to determine future occurrences [29]. The traditional use of analytical techniques in database management has focused more on the description and explanation of what happened in the past. Data Science takes advantage of the volume, variety and velocity of Big Data to identify patterns and algorithms by using predictive modeling and analysis [29]. It provides a set of tools and principles that can be used to predict future occurrences provide intelligent recommendations and actionable discoveries. This approach has been used previously with conventional data sources for tasks such as credit card fraud detection, facial recognition, trading in the stock market, voice recognition and, prediction of customer behavior, etc.

## 3. CRITICAL ANALYSIS

Data Science, Big Data and Business Analytics present opportunities for organisations to be data-driven. Such organisations will use data to predict the future and as a source of innovation. Moreover, a combination of these concepts can form the basis of an organisation’s strategy and may even be used to derive competitive advantage when implemented successfully [29]. Although considering data as an asset is not enough, it is the amount of value that an organisation is able to derive from such consideration that makes Big Data an asset. As the concepts of Data Science, Big Data and Business Analytics continue to develop, the ability to obtain value from these concepts will be determined largely on the understanding of the concepts. The

rest of this section provides an analysis that is based on reviewed literature that will provide further understanding and insight to obtaining value from the concepts.

### 3.1 Big Data: Types and Sources

Traditional datasets are structured into rows and columns and their creation is planned in most cases. However, most of the important types of Big Data are those that do not fall into the traditional dataset category. Broadly, Big Data has been categorised into structured and unstructured types. The structured data types are those generated from enterprise systems. Generally, structured data fits the storage and management principles of the relational database management system [6]. However, the datafication of several other areas of life and other business activities has resulted in the creation of other data types that cannot be stored or managed by using the conventional databases [26]. These types of data are categorised as unstructured data.

Besides the types of data that characterise Big Data, it is also important to take note of the fact that Big Data has broadened the scope of data sources [24]. Therefore, to take greatest advantage of Big Data opportunities, it will be important to explore different sources of data [3]. It is therefore important to consider Big Data based on sources. The sources of Big Data can be categorised into conventional and unconventional sources.

Conventional sources of Big Data include traditional sources where data is expected and gathered. Sources such as the enterprise systems that are implemented to manage business operations like the Customer Relationship Management Systems (CRM), Enterprise Resource Planning (ERP) systems, etc. will be included. Other conventional sources of data will be those collected directly from research, from the organisational document repository, emails and lately from Internet click stream. However, an organisation might be losing out on valuable data if its focus is on conventional sources only. According to Manyika, et al. [24] health care providers are losing out on 90% of valuable data that could be collected as video streams during surgery. This is because such data belongs to the unconventional sources of data – these are sources that generate valuable data, yet are not obvious as sources of data. Such data is drawn from social media, video, audio and images. Equipment fitted with monitoring devices, mobile devices and applications, sensors, WiFi, RFID and mobile data [6, 26] are all important sources of Big Data.

The combination of structured and unstructured, conventional and unconventional sources of data created at a torrential rate is what makes “big data” big and has become a source of learning, insight and a solution to different type of problems. The ability to source and integrate data types from different sources will be a critical success factor for organisations that want to take advantage of the Big Data concepts [11].

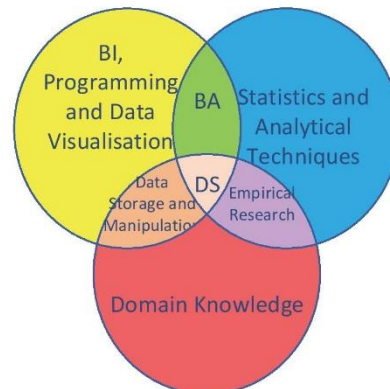
It is proposed that organisations should seek to integrate structured and unstructured data irrespective of their source into data products for data-driven decision making. Data products are data-centric products that solve real life problems for end-users without a direct emphasis on the underlying data [21, 22]. This can be achieved by making use of Data Science. These abilities of Big Data can be harnessed by using business analytics, the tools and principles of Data Science.

### 3.2 Data Science Definition

There is a divergence of opinion in the literature about the definition of Data Science [14, 21]. Many in the literature choose rather not to define it, but to provide a description of what can be achieved from Data Science [13, 27] and who is a Data Science practitioner – Data Scientist [27]. However, based on the literature review, it is apparent that:

- Data Science provides a new systematic approach and tools for dealing with and extracting actionable insight and a prediction of the future from heterogeneous and unstructured data.
- Data Science follows a holistic approach that involves Computer Science, Information System and other fields, such as Statistics and Econometrics, to actively source and identifies patterns from heterogeneous and multi-sourced data to create data products.

The use of data for decision making is not a new phenomenon [21, 29]. It is therefore important to identify how Data Science fits into other related concepts. Data Science has been described as a combination of three primary areas – Computer Science, Statistics and Domain Knowledge [27]. These can be broken down into the application of statistics and analytical techniques, computer programming techniques, data visualisation and the knowledge of a domain area together with Big Data as Data Science. Figure 1 presents a diagram that depicts how the different areas overlap. The major Computer Science aspect includes programming skills, data visualisation and business intelligence which involve reporting and presentation of information. It presents that overlap between Computer Science and the use of statistics and analytical techniques such as Business Analytics which is the combination of the two areas to draw insight and knowledge from data.



**Figure 1. Relationship between Data Science, Business Intelligence and Analytics (Adapted from [29]).**

The third aspect is Domain Knowledge that involves having a good understanding of the domain of the problem for which a solution is required. The combination of domain area and statistical techniques forms the traditional empirical research. The combination of Domain Knowledge and the Computer Science segment would result in the ability to store and manipulate data. The proposed definition of Data Science in this paper suggests that it involves the three primary areas indicated in literature. The Venn diagram in Figure 1 depicts the relationship between

Business intelligence, Programming, Data Visualisation, Business Analytics, Domain Knowledge, Statistics and Analytical Techniques.

### 3.3 Success Factors for Organisations

The following is a summary of critical success factors for implementing Data Science, Big Data and Business Analytics in order for organisations to derive maximum value.

**Data-centric objectives.** Data Science and Big Data are game changers. Several organisations might undergo a paradigm change in their problem solving, decision making process and management approach [25]. It is expected that this transformation move beyond practitioners will also influence how scientific research is carried out. Thus, to implement these concepts successfully, it will be important to view problems from a data perspective [29]. In dealing with organisational or scientific problems, the ability to set goals and objectives with the use of data in mind will be an important success factor.

**Actively collect and integrate relevant data.** It has been established that data is globally available. However, the ability to identify, collect, transform and integrate heterogeneous types of data from several sources will determine how much value is gained [12]. While it will be tempting to collect all possible data, there are risks even in the collection of useful data. An organisation might be faced with the challenges of the cost of storage and management of useless data. Furthermore, it has been suggested that a clear data policy that guides which data is collected, how data is accessed and managed is important to optimise Big Data [24].

**Talent.** Successful implementation of Big Data and Data Science requires a combination of skills – Computer Science, Statistics, Machine Learning, Data Visualisation, Communication and Domain Expertise [27]. However to find and retain individuals with competency in all these areas would be challenging. This is besides the fact that only a limited number of universities have Data Science as part of their curriculum. Research shows that for decision making, there is a shortage of skills specifically for Big Data and Data Science and there is also a shortage of managers in other functional areas that are data-driven. It would be worthwhile to consider building Big Data and Data Science teams in order to tackle the challenges posed by the shortage of skills and the need for heterogeneous skill sets [27]. Therefore, it is suggested that organisations consider the issues of skills acquisition, development and retention when adopting Big Data and Data Science. This will also include ensuring that existing functional managers are well trained to be data-driven.

### 4. FURTHER RESEARCH

Practitioners and researchers in different fields are always introducing new and different terms, concepts and phenomena. Many of these have been proven to be over-hyped and are not generally utilised. However, a careful examination of Big Data and Data Science suggests that these terms do not fall into such a category as the fundamental concepts have been integrated into general organisational terminology.

Therefore, the opportunities for further research in Big Data and Data Science will continue to grow. These include research into how Big Data and Data Science can benefit different industries and specifically in segments within each industry. There are also

opportunities to research and improve the tools, models and frameworks used in Data Science. Other areas might include research into issues of security and infrastructure. Furthermore, Big Data and Data Science have strategic, tactical and operational business implications. These present opportunities for research related to how Big Data and Data Science impact these areas for different industries and at different levels.

### 5. CONCLUSION

The review of literature in this paper suggests that Big Data comes with its challenges and opportunities, but it can be of great value. Volume, velocity, variety, veracity and value were discussed as the defining characteristics of Big Data. Moreover, it was found that Big Data could be structured or unstructured and it can be sourced from conventional and unconventional sources. It was discussed that the ability to identify and integrate different types of data from various sources will be important for deriving the best value from Big Data.

Big Data is changing the way we live, work and play. It has impact on practitioners, academic research work and might result in a complete overhaul of some industries. This paper suggests that the Big Data concept, approach, technologies, techniques, challenges and opportunities will continue to evolve. This is because of the expected increase of the factors that are driving the growth of Big Data.

It was suggested that Data Science takes advantage of Business Analytics for predictions, recommendations and discoveries. A case was made for more research into how these emerging fields can be of benefit and implemented successfully. It was identified that Computer Science support Data Science in areas such as the programming requirements, visualisation and the development of supporting technologies and techniques. Whereas Information Systems enable organisations to contextualise Data Science for different industries and organisations at different levels. Moreover, advance tools from Mathematical and Statistical fields will find more expression and more industry usage through Data Science. Lastly, Data Science needs to draw strongly from industry knowledge for optimum application in business or research.

It is suggested that the concepts of Big Data and Data Science will continue to evolve and will offer new and better ways of making decision. It is expected that the growth of the concepts will mean the introduction of new techniques, technologies and approaches that will change the way information is seen and used. The characteristics of Big Data identified in this paper will extract critical insight and alerts that are of real-time value. The application of Data Science principle and tools will also offer better opportunities for recommendations and predictions of futuristic events and solution to difficult problems.

Based on the critical review of literature, this study identified a relationship between Data Science, Big Data, Business Analytics and Datafication. This study identifies Datafication as a key driving force of the volume, velocity and variety of Big Data and suggests that more aspects of human lives and business activities will be datafied. Also, this study identifies the evolution of Big Data as a basis for the use of advanced analytics to create



algorithms that make and support improved decision making. Finally, this study proposes the following Data Science definition:

*Data Science is a collection of tools and a set of principles that can be used for extracting insight from Big Data and making data-driven decisions with emphasis on predictions, recommendation and discoveries.*

Data Science may be applicable to data that cannot be classified as Big Data, but introduces a new paradigm in decision making for business and research.

#### REFERENCES

- [1] Alpaydin, E. 2010. Introduction to machine learning. MIT Press, Massachusetts.
- [2] Atwell, P. and Rule, J. 1984. Computing and organisations: What we know and what we don't know. *Communications of the ACM*. Volume 27, 12, pp. 1184–1192.
- [3] Brynjolfsson, E., Hitt, L. M. and Kim, H. H. 2011. Strength in numbers: How does data-driven decision making affect firm performance? Available at SSRN: <http://ssrn.com/abstract=1819486> [Accessed 24 March 2014].
- [4] Chae, B. and Olson, D. L. 2013. Business analytics for supply chain: A dynamic-capabilities framework. *International Journal of Information Technology & Decision Making*. Volume 12, 1, pp. 9-26.
- [5] Chaudhuri, S., Dayal, U. and Narasayya, V. 2011. An overview of business intelligence technology. *Communication of ACM*. Volume 54, 8, pp. 88-98.
- [6] Chen, H., Chiang, R. H. L. and Storey, V. C. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*. Volume 36, 4, pp. 1165-1188.
- [7] Clarke, B., Fokoue, E. and Zhang, H. H. 2009. Principle and theory of data mining and machine learning. Springer Science + Business Media, New York.
- [8] Collis, J. and Hussey, R. 2009. Business research: A practical guide for undergraduate and postgraduate students. 3rd edition. Hampshire: Palgrave Macmillan.
- [9] Cukier, K. N. and Mayer-Schoenberger, V. 2013. The Rise of Big Data. *Foreign Affairs* (June 2013). Available at: <http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data> [Accessed 05 May 2014].
- [10] Davenport, T. H. 2014. Big data at work: Dispelling the myths, uncovering the opportunities. Harvard Business School Publishing, Boston.
- [11] Davenport, T. H., Barth, P. and Bean, R. 2012. How big data is different. *MIT Sloan Management Review*. Volume 54, 1, pp. 21-24.
- [12] Davenport, T. H. and Harris, J. G. 2007. Competing on analytics: The new science of winning. Harvard Business School Press, Boston.
- [13] Davenport, T. H. and Patil, D. J. 2012. Data scientist: The sexiest job of the 21<sup>st</sup> century. *Harvard Business Review*, October 2012, pp. 70-76.
- [14] Dhar, V., 2013. Data science and prediction. *Communication of the ACM*. Volume 56, 12, pp. 64-73.
- [15] Elmasri, R. and Navathe, S. B. 1989. *Fundamentals of database systems*. The Benjamin/Cummings Publishing Company, California.
- [16] Flores, I. 1981. *Data base architecture*. Van Nostrand Reinhold Company, New York.
- [17] Galbraith, J. R. 1974. Organisation design: An information processing view. *Interfaces*. Volume 4, 3, pp. 28–36.
- [18] Hey, T. and Trefethen, A. 2003. The data deluge: An e-science perspective. Chap. 38 In *Making the global infrastructure a reality*, edited by Berman, F., Geoffrey C. F. and Hey, A. J. G. pp. 809-858. West Sussex, England: John Wiley and Sons.
- [19] IBM 2011. What is big data? Available at: <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html> [Accessed 22 March 2014].
- [20] Lim, E. P., Chen, H. and Chen, G. 2013. Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems*. Volume 3, 4, pp. 17.1–17.10.
- [21] Loukides, M. 2011a. What is data science? The future belongs to the companies and people that turn data into products. O'Reilly, Cambridge.
- [22] Loukides, M. 2011b. The evolution of data products. O'Reilly, Cambridge.
- [23] Lycett, M. 2013. Datafication: Making sense of Big Data in a complex world. *European Journal of Information Systems*. Volume 22, pp. 381-386.
- [24] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. Big data: The next frontier for innovation, competition and productivity. Available at: [http://www.mckinsey.com/insights/mgi/research/technology\\_and\\_innovation/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation) [Accessed 19 March 2014].
- [25] Mayer-Schroengerger, V. and Cukier, K. 2013. Big data: A revolution that will transform how we live, work and think. New York, Houghton Mifflin Harcourt.
- [26] McAfee, A. and Brynjolfsson, E. 2012. Big data: The management revolution. *Harvard Business Review*. October 2012, pp. 60-69.
- [27] O'Neil, C. and Schutt, R. 2014. *Doing Data Science*. O'Reilly Media, CA.
- [28] Patil, D. J. 2012. Data Jujitsu: The art of turning data into products. Available at: <http://radar.oreilly.com/2012/07/data-jujitsu.html> [Assessed 02 April 2014].
- [29] Provost, F. and Fawcett, T. 2013. Data science for business. O'Reilly Media, Sebastopol.
- [30] Radner, R. 1993. The organisation of decentralised information processing. *Econometrica*. Volume 61, 5, pp. 1109-1146.
- [31] Watson, H. J. and Wixom, B. H. 2007. The current state of business intelligence. *Computer*. Volume 40, 9, pp. 96-99.

## Appendix L: Submitted journal article

Agrekon



### Real-time grain commodities price predictions in South Africa: A Big Data and Neural Networks approach

Journal:	<i>Agrekon</i>
Manuscript ID	Draft
Manuscript Type:	Research Paper
Keywords:	Big Data, Neural Networks, Grain Commodities Prices, Predictions, White Maize

SCHOLARONE™  
Manuscripts

URL: <http://mc.manuscriptcentral.com/ragr>

## **REAL-TIME GRAIN COMMODITIES PRICE PREDICTIONS IN SOUTH AFRICA: A BIG DATA AND NEURAL NETWORKS APPROACH**

### **Kayode Ayankoya**

Department of Computing Sciences  
Nelson Mandela Metropolitan University  
Email: s212400096@nmmu.ac.za

### **Andre P. Calitz**

Department of Computing Sciences  
Nelson Mandela Metropolitan University  
Email: andre.calitz@nmmu.ac.za

### **Jean H. Greyling**

Department of Computing Sciences  
Nelson Mandela Metropolitan University  
Email: jean.greyling@nmmu.ac.za

### **ABSTRACT**

The prices of agricultural grain commodities are known to be volatile due to several factors that influence these prices. Moreover, different combinations of these factors, such as demand, supply and macroeconomic indicators, are responsible for the price volatility at different times. Big Data presents opportunities to collect and integrate datasets from several sources for the purpose of discovering useful patterns and extracting actionable insights that can be used to gain competitive advantage or improve decision making. Neural Networks presents research opportunities for training computer algorithms to model linear and non-linear patterns that might exist in datasets for the purpose of extracting actionable insights such as making predictions.

This paper proposes a Big Data and Neural Networks approach for predicting prices of grain commodities in South Africa. It was identified that disparate data that influence the grain commodities market can be acquired, integrated and analysed in real-time to predict future prices of grain commodities. By utilising SAP HANA as the enabling Big Data technology, data acquired from several sources was used to create an integrated dataset, and a predictive model was developed using Backpropagation Neural Network algorithms. This model was used to predict the daily spot prices of white maize on the Johannesburg Stock Exchange (JSE) at the end of each trading day. The initial results indicate that the approach can be used to predict future prices of grain commodities as the market evolves.

**Keywords:** Big Data, Neural Networks, Grain Commodities Prices, Predictions, White Maize.

**JEL Codes:** Q11, Q13, C63, C89

## 1. INTRODUCTION

The trading of grain commodities is coordinated in South Africa by the Johannesburg Stock Exchange (JSE). The implication of trading grain commodities on the stock exchange is that the grain commodities market in South Africa is *Laissez Faire* in nature. In essence, this indicates that the market, and effectively the prices of grain commodities are controlled by several local and international economic, political and social factors that are rapidly changing. Therefore, stakeholders in the grain commodities market are constantly exposed to price-related risks due to the volatility of prices of grain commodities (Venter, Strydom and Grové, 2013). The volatility of prices of grain commodities and the associated price-related risks suggest that stakeholders will be confronted with important decisions when marketing their products.

The volatility in the prices of grain commodities and other agricultural products has been a source of concern for academic researchers, governmental and non-governmental organisations for many decades (Trostle, 2008; Wright, 2011). Previous studies have shown that many South African grain commodities farmers might be disadvantaged in the market because they do not have the required skills, knowledge and time to monitor and interpret several market indicators (Jordaan and Grove, 2010; Venter, Strydom and Grové, 2013). This has been attributed to the complexities associated with determining the grain commodities market intelligence and future outlook (Jordaan, Grové, Jooste and Alemu, 2007; Venter, Strydom and Grové, 2013).

In order to optimise income and reduce price risks, it is required that stakeholders in the industry sift through volumes of economic, political and social data (Wright, 2011; Trostle, 2008) that has to be sourced from various places. Moreover, they are required to make sense out of the changes in this data as it relates to the grain commodities price on a regular basis (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013). This is essential for them to devise strategies for selling their produce in order to manage price-related risks and increase profitability (Mofokeng and Vink, 2013; Venter, Strydom and Grové, 2013).

Contextually, this could be described as the dilemma of the average grain commodities farmer that enjoys farming activities but is unable to get the best price for his/her produce. Within the value chain of the grain commodities production and trade in South Africa, the grain commodities farmer that is unable to get the best value for his/her produce seems to be absolute price-takers. In the long run, this can be seen as a threat to the sustainability of the operation of such farms due to price-related risks that are faced yearly. Therefore, a system that helps stakeholders, with limited skill and experience, in forecasting grain commodities prices so that they can make better decisions in managing their price risks and increase profitability will be beneficial.

The factors that influence the grain commodities industry include several variables that affect grain prices (Trostle, 2008; Abbott, Hurt and Tyner, 2011; Wright, 2011; Venter, Strydom and Grové, 2013; Khamis, Nabilah and Binti, 2014). These factors can be categorised as:

- Historical and recent market data;
- International demand and supply;
- Macroeconomics; and
- Political factors.

Recent developments regarding the concept of Big Data make it easier to gain access to data on several subjects. Big Data has been described as a concept with the potential to influence all aspects of life including work and play (Manyika et al., 2011; McAfee and Brynjolfsson, 2012).

Big Data is based on the ability that now exists to collect a large volume of datasets compared to what was possible previously. Other characteristics that define Big Data include the wide variety of datasets that complement each other, the velocity at which data is created and the associated veracity (complexity, uncleanness and inaccuracy) of Big Data as a result of the other characteristics (Manyika et al., 2011; Davenport, Barth and Bean, 2012; Mayer-Schonberger and Cukier, 2013). However, it might be erroneous to consider large datasets as Big Data just because of their volume (Goes, 2014). It is the combination of some of these characteristics that makes Big Data different, hence, requiring new thinking and approach for storing and processing data (Chen and Zhang, 2014). This uniqueness, compared to traditional data, is also what defines the opportunities to generate dynamic and real-time insights, support decision making, predict the future and facilitate organisational learning from Big Data.

The ability to collect and integrate datasets from several sources open new opportunities in different fields of interest for the purpose of discovering useful patterns and extracting actionable insights. Big Data also enables new research opportunities to investigate relevant concepts and provide solutions to difficult challenges in different fields (Ayankoya, Calitz and Greyling, 2014). This has led to the evolution of several tools, techniques and technologies that make it possible to leverage large datasets for innovations both in research and practice (Chen and Zhang, 2014).

Predicting the prices of grain commodities will require the collection of the market data and data on the external factors that influence grain commodities prices. It will be necessary to collect historical data to understand patterns, however, there is also the need to collect data as events take place in order to be able to provide real-time intelligence and insight. This can be achieved by taking advantage of the availability of large datasets, together with new technologies, tools and the ability to incorporate all these into a real-time solution that provides a platform for better support for decision makers (Power, 2014). Having more data available in real-time or near real-time together with sufficient tools, techniques and technologies that can be used to extract insight from such data could help decision makers to make decisions quicker by using more relevant information. The financial markets have been a generator of large datasets for many years through millions of transactions processed daily. But the availability of relevant datasets in real-time creates new opportunities to analyse data from such transactions as they take place, which offers improved decision making (Ruta, 2014). The implementation of Big Data concepts, tools and technologies makes it possible to capture, store and use such torrents of data in a stream as they are created (Chen and Zhang, 2014).

Neural Networks can be implemented together with Big Data and enabling environments for extracting actionable insights from large datasets. It is a branch of Artificial Intelligence that is able to learn complex patterns from data for the purpose of solving difficult problems and making decisions. The implementation of Neural Networks is founded on the biological research into the ability of the neural system of the human/animal brain to learn, recognise, store information, generalise and make decisions based on prior knowledge. Research on the application of Neural Networks for understanding complex time series data indicates that it is suitable for making predictions from patterns that can be found in historical time series data (Qi and Zhang, 2008; Crone and Kourentzes, 2010).

It has been found that using Neural Networks for modelling and forecasting future time series observations is not limited by the constraints of statistical approaches such as seasonal trends and stationarity (Qi and Zhang, 2008). Moreover, Neural Networks are able to deal with complex patterns and significant changes in patterns that might occur in the time series because of the ability to use non-linear learning to detect changes and relationships that might exist in the data (Zhang, 2003; Qi and Zhang, 2008; Bukharov and Bogolyubov, 2015). Neural Networks are also considered to be better than statistical techniques in time series analysis because they are able to analyse and forecast qualitative and discrete data types (Bukharov and

Bogolyubov, 2015). Therefore, comparative studies from different areas of application have found Neural Networks to be more efficient than time series analysis that is based on statistical techniques (Co and Boosarawongse, 2007; Zou, Xia, Yang and Wang, 2007; Bennett, Stewart and Lu, 2014).

This study explores the use of Neural Network for predicting prices of grain commodities in South Africa. The spot prices of white maize will be used as an experimental case study. The paper is structured as follows; Section 2 will discuss the factors that influence prices of grain commodities in South Africa. Section 3 will provide an overview of scientific grounding, tools, techniques and methodology used in this study. This is followed by the experimental results of an implementation of the suggestions in this paper in Section 4 and concluding remarks in Section 5.

## **2. FACTORS AFFECTING GRAIN COMMODITIES PRICES IN SOUTH AFRICA**

Past trading activities on the grain commodities market are known to influence future trading and the prices of grain commodities (Jordaan et al., 2007; Wright, 2011). It is, therefore, important to consider local market transactions of the grain commodity of interest, as well as international trade, for the same commodity in countries where their commodities trading markets affect those of South Africa. Variables to consider should include trade data such as price, volume traded, bidding prices and so on, as provided by the stock exchanges. Several of the grain commodities can be used for the same purpose, therefore they are considered as substitutes and their economics are considered to be interdependent (Wright, 2011). Hence it will be important to include the effect of substitutes in studying the factors that affect grain commodities prices in South Africa. The following discussion provides an overview about other factors that affect grain commodities prices in South Africa.

### **2.1 Demand, supply and storage**

Economic theories suggest that prices will go up when there is an increase in demand for any commodity, especially when the supply of such a commodity does not increase with demand (Burda and Wyplosz, 2009). In reverse, the price of commodities is forced downward when there is over-production, reduced demand or a huge stockpile of commodities. This summarises the impact that the local and international utilisation of grain commodities for domestic and industrial purposes has on the grain commodities price.

Variables under this theme include factors that influence the ability of farmers to supply or those factors that cause over-supply and the calming or panic effect that the level of the grain stockpile has on the volatility of grain prices (Wright, 2011; Abbott, Hurt and Tyner, 2011; DAFF, 2014). The demand for grain commodities as an important source of calories for human consumption and industrial demand for animal feeds and biofuel also play a role (Wright, 2011; Trostle, 2008). The prominent variables under the demand, supply and storage theme that influence grain prices are as follows:

- Domestic utilisation;
- Industrial utilisation;
- Utilisation by major importing countries;
- Production level in major exporting countries;
- Influence of weather on production;
- Input costs;
- Local stockpile;

- International stockpile;
- Price, demand, supply and storage of substitutes; and
- Level of utilisation compared to stockpile (stock-to-use-ratio).

## **2.2 Macroeconomics**

Macroeconomic factors have been identified as influencing the changes that occur in the price of grain commodities. Similar to the previous theme, there are several variables which influence grain prices that fall under this theme. Studies show that some of the macroeconomic variables influence the prices because they are linked directly to the factors of production (Trostle, 2008). On the other hand, the influence of the other macroeconomic factors is simply a reflection of the state of the local or global economy (Abbott, Hurt and Tyner, 2011). Although, there are suggestions that the use of macroeconomic variables for understanding grain commodities prices requires further research (Wright, 2011), it remains an important part of the discourse on the price of grain commodities (Abbott, Hurt and Tyner, 2011; DAFF, 2014; Wright, 2014; Trostle, 2008). The macroeconomic variables that influence the price of grain commodities, identified from the literature that has been cited above, include:

- Currency exchange rates (especially US Dollars to other currencies);
- Price of crude oil;
- Local interest rates; and
- Consumer price index.

## **2.3 Political factors**

The influence of government policies, political interactions and international trade which is largely driven by politics, cannot be separated from the fluctuations in the prices of grain commodities (Abbott, Hurt and Tyner, 2011). Although the grain commodities market is deregulated in South Africa, political influence on economic, social and trade-related issues is a reality. It is generally understood that issues relating to politics impact the prices of grain commodities (Abbott, Hurt and Tyner, 2011; Wright, 2014; Trostle, 2008). However, unlike the variables under the other themes which have quantitative indicators for which data is collected and analysed, the impact of politics on the prices of grain commodities can be said to be largely subjective and opinion driven. The influence of political factors on the prices of grain commodities will not be included in this study. There is, however, an opportunity for future studies to determine quantitative indicators for political factors that influence the grain commodities market.

The factors explored in this segment provide insights into the factors that influence the grain commodities prices. These include variables for which data in monthly, daily, hourly and in some cases minute by minute is generated and stored.

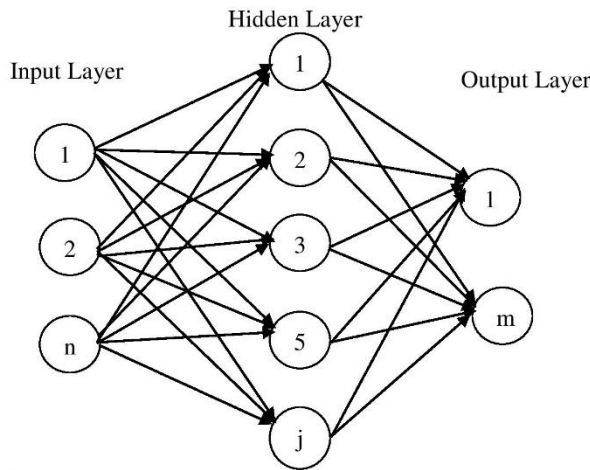
## **3. METHODOLOGY**

### **3.1 Backpropagation Neural Network for forecasting**

Several modelling algorithms exist for the different Neural Network architectures for making approximations such as making predictions or classifications. Wilamowski (2009) alluded that the choice of algorithm should be based on the type and complexity of the problem for which a model is being trained. The Backpropagation Neural Network (BPNN) which is based on the feed forward Neural Network has been found to be widely suitable for problems requiring prediction from data such as a time series (Khashei and Bijari, 2010; Evans, Pappas and Xhafa,

2013; Khamis, Nabilah and Binti, 2014). BPNN is made up of layers of intertwined neurons consisting an input layer, hidden layers and the output layer as shown in Figure 1.

BPNN follows the multi-layer learning networks system as shown in Figure 1, where there is an input layer that is composed of neurons (1 to n) representing the independent variable. A typical BPNN comprise one or more hidden layers with neurons (1 to j) which are weighted and determine the degree of influence during the learning process; these hidden layers enable the network to use a non-linear function to model complex patterns (Alpaydin, 2010). Finally, Back Propagation Neural Networks also contain an output layer with neurons (1 to m) representing the estimated variables (dependent variables) as shown in Figure 1. During the learning process, BPNN sends a signal about errors from the output layer back to the hidden layer. The connections that exist between the neurons from each of the layers facilitate the learning process through the use of mathematical functions depending on the type of Neural Network and its setup (Engelbrecht, 2007). This ensures that subsequent learning produces an output with lesser error value until an optimal output is discovered (Alpaydin, 2010).



**Figure 1: Simple Backpropagation Neural Networks**

Previous studies have shown that BPNN is suitable for making predictions that are based on historical data even when they involve complex patterns (Ghwanmeh, Mohammad and Al-ibrahim, 2013; Tsadiras, Papadopoulos and O’Kelly, 2013). Hence, many time series-related problems in areas such as financial forecasting, engineering and medical research have successfully implemented BPNN. Thus, this study adopts BPNN for the implementation of the Neural Network modelling and predicting prices of grain commodities.

Zhang (2003) and Qi and Zhang (2008) suggested that the relationship that exists between the input variables and the output variables in a feed-forward Neural Network can be represented mathematically as;

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g \left( \beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t \quad (1)$$

where  $y_t$  is the expected output,  $\alpha_j$  ( $j = 0, 1, 2, \dots, q$ ) and  $\beta_i$  ( $i = 0, 1, 2, \dots, p$ ) represents the weights for the connections between the neurons in the hidden layer and the output nodes;  $p$  represents the number of input nodes and the number of hidden nodes is represented by  $q$  in



the equation. The transfer function between the hidden layer and the output node is denoted by  $g$  which could be a sigmoid function, expressed as:

$$g(x) = \frac{1}{1 + \exp(-x)} \quad (1a)$$

The mathematical representation of the Neural Network denotes a non-linear autoregressive relationship that exists between the future value  $y_t$  and past observation  $(y_{t-1}, y_{t-2}, \dots, y_{t-p})$  (Khashei and Bijari, 2011). Hence, the Neural Network model in equation (1) can be presented mathematically as:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, \emptyset) + \varepsilon_t \quad (2)$$

Where  $f(\cdot)$  denotes the Neural Network model and  $\emptyset$  is a vector of the parameter in equation (1). However, for a time series model where external variables are considered besides the internal autocorrelation function, the past observations of the external variables can be included in the model as:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, x_{1t-1}, x_{1t-2}, \dots, x_{1t-p}, \dots, x_{rt-1}, x_{rt-2}, \dots, x_{rt-p}, \emptyset) + \varepsilon_t \quad (3)$$

where  $x_{rt-1}$  denotes the observation for the external variable  $r$  collected during period  $t - 1$ . This has been included because the predicted outcome,  $y_t$ , is influenced by past observations of the same series, as well as by the past observations of the external variables. Hence, this implementation will consider the prices of previous trading days as input, as well as observations of the past trading days for the factors that influence the price of white maize.

By using the model represented in equation (3) to make predictions for a period  $t + n$  in the time series, where  $t$  is the current time and  $n$  is a positive integer, there is a need to make provision for the fact that data from the period between time  $t$  and  $t + n$  will not exist. Hence, the Neural Networks model can be built to find the pattern between the independent variables at the current time  $t$  and the associated past observations, for predicting the future time for time  $t + n$ . Therefore the equation (3) can be written as:

$$y_{t+n} = f(y_t, y_{t-1}, \dots, y_{t-p}, x_{1t}, x_{1t-1}, \dots, x_{1t-p}, \dots, x_{1t}, x_{1t-1}, \dots, x_{rt-p}, \emptyset) + \varepsilon_t \quad (4)$$

This model can be used as the foundation for predicting future prices while taking into consideration the changes in the market dynamics. It also provides a basis for the retraining of the model to ensure that changes in the market dynamics are captured continuously. Thus, technological advancements that come with a Big Data environment such as in-memory, cloud and parallel computing can be leveraged by developing and retraining different models on how the combination of historical and real-time data influences different periods in the future. Hence, at the close of a business day, different models can be retrained based on the historical patterns that include the day's transactions to determine what will happen in the next 1, 2, 3 days and so forth.

### 3.1.1 Features selection for model

Selection of the right input variable that optimally captures and explains the patterns in a time series model is considered very crucial for the degree of accuracy of the model resulting from a Neural Network (Co and Boosarawongse, 2007; Crone and Kourentzes, 2010; Qi and Zhang, 2008). The extant literature shows that deciding on the input variable for time series modelling, using Neural Networks, might be an art as much as a scientific expedition. Several authors conclude that there is generally no accepted theoretical background to follow in deciding the

input variables in a Neural Network based time series modelling (Zou et al., 2007; Khashei and Bijari, 2011; Jabjone and Wannasang, 2014).

In developing a multivariate model, there is a need to consider not only the past observations of the variable being modelled but also to examine the influence of external variables as denoted in equation (3). Thus, in a multivariate analysis, the choice of external variables that will lead to an optimized model is crucial. Several studies on multivariate time series modelling used the analysis of correlation to support the choice of external variable (Yu and Ou, 2009; Khamis, Nabilah and Binti, 2014; Jabjone and Wannasang, 2014). However, it is important to highlight that correlation analysis does not imply that these variables are, of a certainty, responsible for the patterns that exist in the price data (Irwin, Sanders and Merrin, 2009; Bukharov and Bogolyubov, 2015). Hence, the choice of external variables can also be supported by previous knowledge in the field of interest (Wiles and Enke, 2014).

Also, for univariate analysis, as well as multivariate time series analysis after the external variables have been selected, there is still a need to decide how far back to go in including the effect of past observations in predicting future values. One of the major reasons for using the Neural Networks for time series analysis is to identify and capture non-linear relationships that might be in the dataset (Qi and Zhang, 2008; Bukharov and Bogolyubov, 2015). However, there is empirical and theoretical evidence that complex time series data with non-linearity patterns can also possess some linear characteristics (Khashei and Bijari, 2011). Qi and Zhang (2008) supported the use of techniques such as lagging to include the linear effect of past observations in the Neural Network-based time series data. There are no generally acceptable foundations for selecting the lag length for a Neural Network based time series modelling. However, similar work has made use of random experiments to determine the lag length that produced the best model (Zou et al., 2007; Khashei and Bijari, 2011).

The total number of input variables for the Neural Networks will consist of the dependent and independent variable(s) that have been selected. These will also include the lagged variables for each of the selected variables.

### **3.1.2 Other Neural Network parameters**

Other parameters that are required in setting up a BPNN include the number of hidden layers, learning rate, momentum factor and the transfer function. The learning rate,  $\eta$ , determines the number of steps that is taken in the search for the output. If the chosen learning rate is too large, the optimum can be missed, and when it is too small, the network can take too long to train (Engelbrecht, 2007). The momentum factor,  $\alpha$ , determines the degree of influence that weights of previous learning will have on the current learning. It allows the training process to use the identified weights of the previous learning iteration so that the weights of the past iteration are introduced as inertia into the current learning iteration (Larose, 2005). The momentum factor ranges from 0 to 1, meaning that when the momentum term is close to or equal to one, the weight of the current iteration will be essentially the same as the previous one.

The learning process in Neural Networks is iterative; therefore the number of iterations for the learning process should be set from the beginning as an exit criterion for the network. However, depending on the selected learning rate or the momentum rate, it could take much longer to achieve the set number of iterations. In such cases, a target error level can also be set for which the learning process will be terminated when it is achieved (Larose, 2005). The selection of the optimum learning rate, momentum term and exit criterion is a balancing act considering the implication that each of the parameters has on performance of the network. The use of experiments is also suggested as the approach for choosing the other network topology parameters such as the hidden layer, learning rates and momentum factor (Qi and Zhang, 2008;

Tsadiras, Papadopoulos and O’Kelly, 2013; Ghwanmeh, Mohammad and Al-ibrahim, 2013; Khamis, Nabilah and Binti, 2014).

Besides the mentioned network parameters, it is suggested that the input data for Neural Network modelling be pre-processed into a normalised format ranging between -1 and 1 or 0 and 1 (Engelbrecht, 2007; Co and Boosarawongse, 2007; Khamis, Nabilah and Binti, 2014). Transforming the input data into a range of 0 to 1 can be achieved by using the equation:

$$y_t = \frac{y_t - y_{min}}{y_{max} - y_{min}} \quad (5)$$

Where  $y_t$  is an observation for time  $t$ ,  $y_{min}$  and  $y_{max}$  are the minimum and the maximum observed values of all the observations of a given variable.

### 3.2 Model evaluation

The purpose of identifying the correct parameters and topology for training an optimal Neural Network for different problems is so that the resulting model can adequately be used to estimate future occurrences based on historical data. However, care needs to be taken to ensure that the output from a model is not a result of just memorising the historical data (Provost and Fawcett, 2013a). In such cases, the model will have high accuracy when used to forecast a subset of the data used in training the model, but will not produce a reasonable result when used to predict a dataset not seen by the model during training. Hence the model has not learned the patterns in the data, but it has only memorised the observations in the data. This problem is regarded as overfitting (O’Neil and Schutt, 2014).

Contrary to just memorising the observations in a training dataset, the desired model from a modelling exercise is one that is able to accurately estimate future outcomes based on input data that has not been seen by the training model at all. This is regarded as generalisation (Provost and Fawcett, 2013b). The level of accuracy of a model can be measured by its ability to generalise even when there is a significant change in the input data. There is a risk of overfitting a model to the training data when the model becomes too complex, such as having too many hidden nodes or too few observations compared to the number of input nodes (Alpaydin, 2010). However, the ability of the model is reduced greatly with an overly simple network (Co and Boosarawongse, 2007). Hence, there is a need to strike a balance between generalisation and overfitting.

A common practice for avoiding overfitting is to split the available dataset into a training and a test set (O’Neil and Schutt, 2014; Provost and Fawcett, 2013b; Alpaydin, 2010), where the test set is kept completely separate and not used in the training process. The performance of the model is then checked by using the model to forecast the series in the test set and to compare the results with the actual data. Statistical measures such as the Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), provide quantitative measures for comparing the results of predictions from the training set and the test set. MSE is a modelling evaluation statistic that gives an indication of how much a set of values that has been predicted by using a model varies from the actual observations (O’Neil and Schutt, 2014). It represents the loss function between the result of a trained model when compared with the actual value, hence it is regarded as the training error for Neural Networks (Wilamowski, 2009; O’Neil and Schutt, 2014). The MSE is defined as:

$$MSE = \frac{1}{n} \sum_{t=1}^n (Y_t - F_t)^2$$

Where  $F_t$  represents the predicted values,  $Y_t$  the observed actual values and  $n$  the total number of values. However, a more popular measure of the accuracy of a model is the Root Mean Squared Error that is obtained by taking the square root of MSE (Khashei and Bijari, 2011; Bennett, Stewart and Lu, 2014). RMSE is represented as:

$$RMSE = \sqrt{MSE}$$

The Mean Absolute Percentage Error (MAPE) is another measurement of accuracy of a predictive model which presents the predictions error as a percentage of the actual observed values. It calculates the absolute value of the ratio of the error to actual values (Tofallis, 2015), and calculates it as a percentage by multiplying it by 100. MAPE is obtained as:

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{Y_t - F_t}{Y_t} \right|$$

These statistics are generally used for measuring model accuracy in time series forecasting (Enders, 2010; Tsay, 2010) and have also been adopted in measuring the accuracy of a Neural Network-based time series model as well (Zou et al., 2007; Crone and Kourentzes, 2010; Khashei and Bijari, 2011; Khamis, Nabilah and Binti, 2014).

#### 4. EXPERIMENTAL RESULTS

The proposed approach was implemented by setting up experiments for predicting the spot price of white maize traded on the Johannesburg Stock Exchange. SAP HANA was adopted as the technology of choice to demonstrate the proposed approach because of its ability to support Big Data and advance analytics solutions (Chen and Zhang, 2014). It provides the required technology to handle real-time acquisition, pre-processing and predictive analytics of large datasets.

##### 4.1 Implementation

For the purpose of this implementation, historical data on spot transactions on grain commodities was obtained from the website of JSE with permission to use the data for research purposes. End-of-day data for spot prices was captured directly from the newsfeed provided on the website of the JSE while end-of-day data was captured from the website of a major grain commodities storage company ([www.senwes.co.za](http://www.senwes.co.za)). To include the influence of other markets outside South Africa, this implementation included the effect of the grain commodities market in the USA as a major producer of white maize. Data on the corn trade in the USA was collected from the Chicago Board of Trade (CBOT) through third party data subscription. Moreover, the demand and supply data was collected from a service made available by the South African Grain Information Services (SAGIS) website. Data on the production and consumption in the USA was also collected through free services offered by the Economic Research Services (ERS) of the United States Department of Agriculture (USDA) on its website. Other sources of data include the websites of the Reserve Bank, where historical and the current interest rate in South Africa, as well as the daily South African Rand – US Dollar currency exchange rates data was collected. Finally, historical and current data on the prices of Brent Crude Oil was accessed through open data services on [www.quandl.com](http://www.quandl.com).

The data collected was largely unstructured and in various formats. The Big Data tools and techniques, however, made it possible to clean, structure and integrate the datasets into a single time series data, by using the date as the integrating factor. Data for this experimental

analysis was then extracted from the integrated repository of data. Historical data of the end-of-day spot price of white maize was taken from 02 January 2007 till 31 July 2015 resulting in a total of 2149 observations together with the data on factors influencing the prices as independent variables.

#### 4.2 Model training

Correlation analysis was carried between each of the dependent variables (spot prices of white maize) and the independent variable in order to decide the input variables for the training of the networks. This was carried out with the support of extracts from literature on the factors that influence grain commodities prices in South Africa. Table 1 presents a list of the variables that were selected for modelling the spot prices of white maize.

**Table 1: Input variables for Neural Network model for WMAZ spot price**

No	Variables	Correlation with spot price of WMAZ
1	Spot price of WMAZ (lagged)	
2	Spot price of Wheat	0.6280 (n=2149)
3	USD-Rand exchange rate	0.5885 (n=2149)
4	Spot price of Brent Crude oil	0.3191 (n=2149)
5	Prime interest rate in SA	-0.3428 (n=2149)
6	Price of Corn in USA	0.2860 (n=2149)
7	Volume of Corn Trade in USA	0.2848 (n=2149)
8	Demand for WMAZ in SA	0.2474 (n=2149)
9	Demand for Wheat in SA	0.3347 (n=2149)

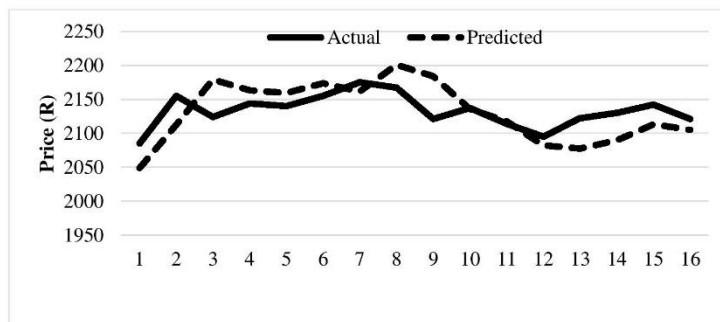
Initial experiments were also carried out to determine the other network parameters that would produce the best model. The optimal model was identified to have a lag length of 5, signifying the inclusion of the effect of the previous 5 trading days in the model. Considering the number of variables and lag length, the model with 7 hidden layers were also found to perform better than the rest. The model was also found to perform optimally with learning rate set to 0.4 and the momentum factor set to 0.001.

#### 4.3 Cross-validation of model

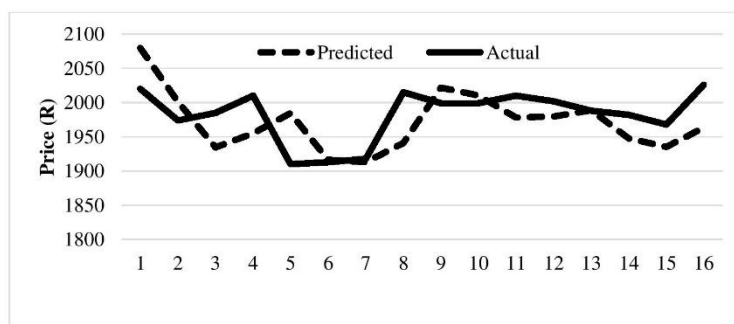
BPNN model for spot prices of white maize in South Africa were created. Subsequently, a validation process was also carried out to ensure that the model was able to generalise and not just memorising the input data. The model was used to make predictions by using subsets of the training dataset and the testing dataset. Using a subset of the training data to make predictions is known as the in-sample evaluation while predicting with a dataset that is totally separate from the one used in training the model is known as the out-sample evaluation (O'Neil and Schutt, 2014). The BPNN model for the spot price of white maize in South Africa was trained by using historical data of transactions that happened between 01 January 2010 and 31 December 2014. In-sample evaluations were carried out with subsets of the training data while out-of-sample evaluations were carried out with the testing data. For both categories, the created model was used to make predictions for 1 and 3 month periods.

The dataset for the trading days in the last month of the training data from 01 December 2014 to 31 December 2014 was predicted and compared against the actual prices. In-sample predictions were also made and compared with actual prices over a period of 3 months using data from 01 October 2014 till 31 December 2014. A comparison of the predicted and actual spot prices of white maize for in-sample as well as out-sample predictions within a 1-month

period are presented in Figures 2 and 3 respectively. The graph in Figure 2 show that the in-sample predictions are very close to the actual values and it also indicate that the in-sample predictions followed the trend of the actual prices quite closely. On the other hand, Figure 3 present the result if the out-sample predictions. The predicted prices were also close to the actual prices, however, not as much as it is with the in-sample predictions.

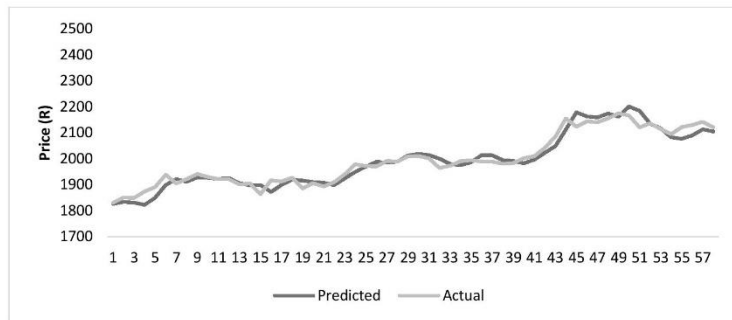


**Figure 2: Comparison of actual vs predicted spot prices of white maize (1 month in-sample).**

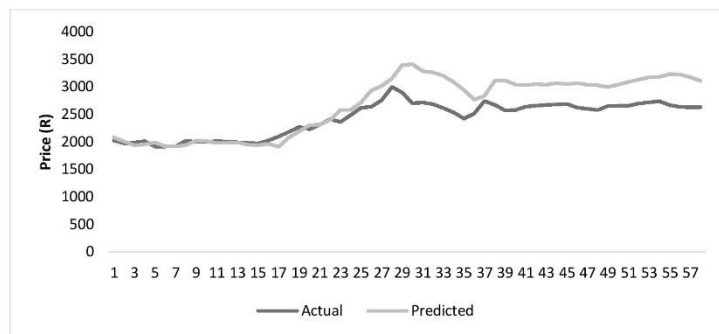


**Figure 3: Comparison of actual vs predicted spot prices of white maize (1 month out-sample)**

By using the same model, the spot price of white maize over a period of 3 months were also predicted. Figures 4 and 5 shows a comparison of the actual spot prices of white maize and the predicted prices over a 3 months period for the in-sample and out-sample predictions respectively. The results show that the in-sample predictions over the 3 month period were very close to the actual prices of white maize. However, the accuracy of the out-sample predictions depreciated significantly from about the prediction for the 25th trading day as shown in Figure 5.



**Figure 4: Comparison of actual vs predicted spot prices of white maize (3 months in-sample).**



**Figure 5: Comparison of actual vs predicted spot prices of white maize (3 months out-sample).**

To measure the prediction accuracy of the model, the Mean Absolute Percentage Error (MAPE) statistic for the in-sample and out-sample predictions for the two different periods were compared. The results, as shown in Table 2, indicate that the MAPE of the in-sample predictions (1.31%) and that of the out-sample predictions (2.26%) over a period of a single month were relatively close. However, the predictions over a 3-month period the MAPE for the in-sample predictions was 0.97%, while that of the out-sample predictions was 9.20%. This signify a noticeable difference when compared to the result obtained for predictions over a single month. However, the correlation between the predicted prices and the actual prices for the 3-months was 0.9709 and 0.9598 for the in-sample and out-sample predictions respectively. This suggests that both the in-sample and the out-sample predictions over the 3-month period followed the trend of the actual spot price than the predictions over 1-month period with 0.6568 and 0.1412 correlation for the in-sample and out-sample predictions respectively. These results suggest that the model is able to generalise and make predictions for unseen data, although there is room for further research into improving the model.

**Table 2: Summary of verification of BPNN model for spot prices**

Period	In-sample			Out-sample		
	MAPE(%)	RMSE	R <sup>2</sup>	MAPE(%)	RMSE	R <sup>2</sup>
<b>1 month</b>	1.31	32.97	0.6568	2.26	61.02	0.1412
<b>3 month</b>	0.97	24.61	0.9709	9.20	348.64	0.9598

The predictions depicted by the graphs in Figures 2 - 5 show that the model is more accurate with in-sample predictions as expected, especially for predictions over 3 months. When the same model is applied for making out-sample predictions using the input dataset that was not used for the training process, the model was less accurate. However, the results of the out-sample predictions suggest that the model was intelligent enough to recognise the market trend, although, the deviation between the actual and the predicted price increased significantly with time. This result suggests that the identified BPNN topology and architecture could be used for predicting spot prices of white maize in South Africa. However, there is a need to implement strategies that will improve the accuracy of the predictions.

#### 4.4 Real-time predictions

The cross-validation in Section 4.3 is based on the assumptions that the external data is available for the period for which the price of white maize is being predicted. However, as proposed in Section 2.2 with the model denoted as equation (4), a model can be built based on all the available data for predicting the spot and futures contract prices for different days into the future. This model can then be retrained periodically as new data becomes available to ensure that new market dynamics are captured in the Neural Network.

Based on the suggestions of Ruta (2014) on the use of Big Data for real-time learning for financial assets trading, new BPNN algorithms for building models for 14 trading days ahead were written. Each of the models was run continuously until 10 different predictions were recorded for each day. Thereafter, the mean value of the 10 predictions captured for each day was taken as the final prediction.

**Table 3: Tables showing the input datasets used for modelling**

Training		Prediction		
Start	End	Start	End	Results for
2010-01-01	2015-07-15	2015-07-16	2015-07-31	2015-08-03
2010-01-01	2015-07-18	2015-07-19	2015-08-03	2015-08-04
2010-01-01	2015-07-19	2015-07-20	2015-08-04	2015-08-05
2010-01-01	2015-07-20	2015-07-21	2015-08-05	2015-08-06
2010-01-01	2015-07-21	2015-07-22	2015-08-06	2015-08-07
2010-01-01	2015-07-22	2015-07-23	2015-08-07	2015-08-10
2010-01-01	2015-07-25	2015-07-26	2015-08-10	2015-08-11
2010-01-01	2015-07-26	2015-07-27	2015-08-11	2015-08-12
2010-01-01	2015-07-27	2015-07-28	2015-08-12	2015-08-13
2010-01-01	2015-07-28	2015-07-29	2015-08-13	2015-08-14
2010-01-01	2015-07-29	2015-07-30	2015-08-14	2015-08-17
2010-01-01	2015-08-01	2015-08-02	2015-08-17	2015-08-18
2010-01-01	2015-08-02	2015-08-03	2015-08-18	2015-08-19
2010-01-01	2015-08-03	2015-08-04	2015-08-19	2015-08-20

The experiments were set up to use a rolling subset of data as the input for the training and the predictions as shown in Table 3. The experiments made use of datasets between 01 January 2010 and 15 July 2015 as the training set for building the model for the first trading day in the month of August. New daily data was included in the input data for retraining the



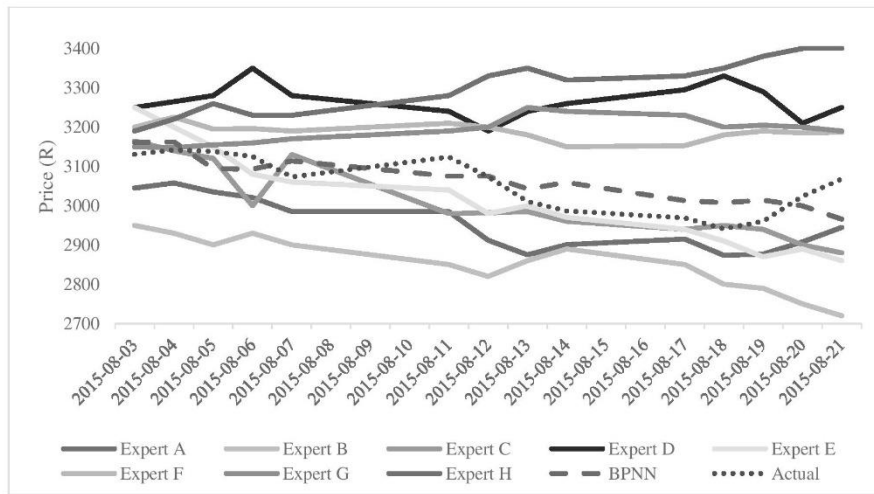
model at the end of each day. This was also applied to the input data for the predictions, by adding data from the previous trading day as shown in Table 3.

Besides the use of the measurement of accuracy statistics to measure the technical abilities of the models, 8 expert grain commodities traders (referred to experts A – H) agreed to voluntarily participate in the evaluation exercise. The panel of experts that agreed to participate are from 3 different companies that are listed on the Johannesburg Stock Exchange’s website as registered to trade grain commodities in South Africa. Moreover, some of these trading companies also buy and sell grain commodities as financial assets on the Johannesburg Stock Exchange. The experts were asked to predict the future spot prices of white maize on the Johannesburg Stock Exchange for the month of August 2015 before the beginning of the month of August 2015.

The results in Table 4 indicate that the predictions from the BPNN model had lesser deviation from the actual spot prices than the predictions from all the experts. The measurement of accuracy statistics shows that the predictions by the BPNN model had the minimum error with the Mean Absolute Percentage Error (MAPE) = 1.44% and Root Mean Square Error (RMSE) = 49.91 when compared to the predictions of the experts. This is only followed by the predictions of Expert C with MAPE = 2.16% and RMSE = 85.78. Figure 6 provides a graphical representation of the results, showing that the price predicted by the BPNN model is about the closest to the actual prices recorded, although there is more room for improvements.

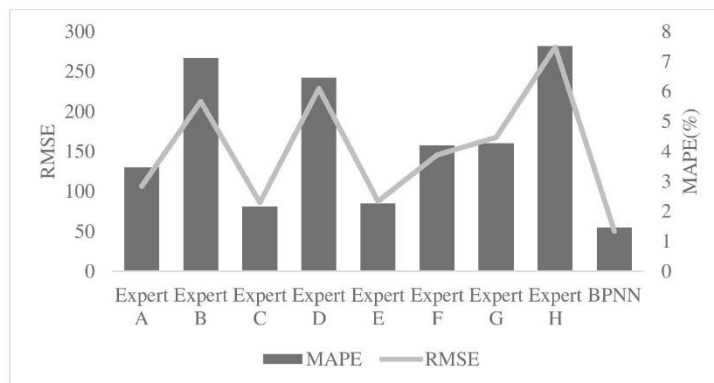
**Table 4: Comparison between predictions from experts and implemented DSS for spot prices of white maize.**

Day	Expert A	Expert B	Expert C	Expert D	Expert E	Expert F	Expert G	Expert H	BPNN	Actual
1	3045	2950	3165	3250	3250	3200	3150	3190	3161	3131
2	3058	2930	3140	3265	3200	3225	3148	3220	3162	3142
3	3035	2900	3120	3280	3150	3195	3155	3260	3094	3138
4	3021	2930	3000	3350	3080	3196	3160	3230	3093	3125
5	2985	2900	3130	3280	3060	3190	3170	3230	3114	3073
6	2985	2850	2980	3240	3040	3210	3190	3280	3075	3124
7	2912	2820	2982	3190	2980	3200	3200	3330	3075	3074
8	2875	2860	2985	3240	3000	3180	3250	3350	3043	3011
9	2901	2890	2960	3260	2970	3150	3240	3320	3059	2987
10	2915	2850	2940	3295	2940	3153	3230	3330	3013	2969
11	2874	2800	2950	3330	2910	3180	3200	3350	3008	2941
12	2877	2790	2940	3290	2870	3190	3205	3380	3014	2960
13	2908	2750	2900	3210	2890	3185	3200	3400	3000	3024
14	2945	2720	2880	3250	2860	3187	3190	3400	2966	3068
MAPE	3.46%	7.11%	2.16%	6.46%	2.26%	4.20%	4.27%	7.50%	1.44%	
RMSE	106.22	212.67	85.78	228.51	87.54	145.40	167.71	280.49	49.91	
R-squared	0.9099	0.5241	0.6454	-0.1554	0.7771	0.7457	-0.7851	-0.7141	0.7153	
	(n=14)	(n=14)	(n=14)	(n=14)	(n=14)	(n=14)	(n=14)	(n=14)	(n=14)	



**Figure 6: Prediction of spot prices of white maize by experts and BPNN model.**

The calculation of MAPE on Table 4 indicate the difference between each predicted value and the actual value that was recorded as the percentage, showing the size of error between the predicted value and the actual value. On the other hand, the RMSE shows a measurement of how much the predicted prices deviates from the actual prices. Figures 7 present a graphical view of the error statistics which compare the performance of the BPNN model with predictions by experts. The graph shows that the BPNN model implemented performed relatively better with minimum deviation from the actual prices in terms of value.



**Figure 7: Error measurements of experts and BPNN model predictions for spot prices.**

Both measurements of accuracy suggest that the predictions by the BPNN model performed better than the predictions made by the experts. The practical implication of this result is that the acquisition and analysis of data on factors that influence the grain commodities market in real-time present opportunities to create Decision Support Systems (DSS) for trading in grain commodities in South Africa. Such DSS can be used to assist stakeholders, such as the farmers, with limited skills and resources, with making decisions about trading their grain commodities.

## 5. CONCLUSION

This paper set out to demonstrate that grain commodities prices in South Africa can be predicted in real-time or near real-time by using Neural Networks and by taking advantage of the evolution in the concept of Big Data. It was identified that the grain commodities market data and data on the factors that influence the markets are available from different sources. Although the data is scattered in different locations and is often available in different formats, the tools and techniques of Big Data make it possible to source, acquire and integrate this data, even in real-time. Local demand and supply of grain commodities, international grain commodities markets, and macroeconomic indicators were identified as some of the factors that influence the grain commodities market in South Africa, this is besides the influence of past grain commodities market transactions. However, it should be acknowledged that there might be other variables that could influence the price of grain commodities not identified by this study, such variables can be added in the future to improve the outcome of the propositions in this study.

Backpropagation Neural Network makes it possible to explore patterns in datasets regardless of the fact that these patterns might be linear or non-linear. It has also found its application in modelling time series problems in fields such as medical, econometrics and engineering. It was demonstrated in this paper that the BPNN can be used to model and predict grain commodities prices. Furthermore, by using SAP HANA as a Big Data platform, it was demonstrated that with the acquisition of data in real-time or near real-time, a BPNN model can be retrained periodically as new data becomes available. This will ensure that changes in the market data are captured early enough and used in making predictions about future grain commodities prices. Empirical results in this study revealed that this approach could provide better results than experienced grain commodities traders.

This study is part of a bigger research into how a decision support can be provided for grain commodities trading in South Africa, especially for farmers with limited skills and resources for predicting grain commodities prices. Further studies will explore creating decision support, such as predictions, recommendations and discoveries that can be extracted from relevant datasets in real-time for trading grain commodities. Studies will also be carried out to explore how such market intelligence can be made available by using mobile technologies for easy access.

## REFERENCES

- Abbott, P.C., Hurt, C., and Tyner, W.E., 2011. What's driving food prices in 2011. *Farm Foundation Issue Report*.
- Alpaydin, E., 2010. *Introduction to Machine Learning*. Massachusetts: The MIT Press.
- Ayankoya, K., Calitz, A.P., and Greyling, J.H., 2014. Intrinsic relations between Data Science, Big Data, Business Analytics and Datafication. In: *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference*. Centurion: ACM Digital Library:192–198.
- Bennett, C., Stewart, R.A., and Lu, J., 2014. Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*, 7(5):2938–2960.
- Bukharov, O.E., and Bogolyubov, D.P., 2015. Development of a decision support system based on neural networks and a genetic algorithm. *Expert Systems with Applications*,

42(2015):6177–6183.

- Burda, M., and Wyplosz, C., 2009. *Macroeconomics*. 5th ed. New York: Oxford University Press.
- Chen, P.C.L., and Zhang, C.Y., 2014. Data-Intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275:314–347.
- Co, H.C., and Boosarawongse, R., 2007. Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks. *Computers and Industrial Engineering*, 53(4):610–627.
- Crone, S.F., and Kourentzes, N., 2010. Feature selection for time series prediction: A combined filter and wrapper approach for neural networks. *Neurocomputing*, 73:1923–1936.
- DAFF, 2014. Trends in the agricultural sector 2013. *Department of Agriculture, Forestry and Fisheries, Pretoria*.
- Davenport, T.H., Barth, P., and Bean, R., 2012. How Big Data is Different. *MIT Sloan Management Review*, 54(1):21–24.
- Enders, W., 2010. *Applied econometric time series*. New Jersey: John Wiley & Sons, Inc.
- Engelbrecht, A.P., 2007. *Computational intelligence: An introduction. Studies in Computational Intelligence*, West Sussex: John Wileys and Sons.
- Evans, C., Pappas, K., and Xhafa, F., 2013. Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation. *Mathematical and Computer Modelling*, 58:1249–1266.
- Ghwanmeh, S., Mohammad, A., and Al-ibrahim, A., 2013. Innovative artificial neural networks-based decision support system for heart diseases diagnosis. *Journal of Intelligent Learning Systems and Applications*, 5:176–183.
- Goes, P.B., 2014. Big Data and IS Research. *MIS Quarterly*, 38(3):iii–viii.
- Irwin, S.H., Sanders, D.R., and Merrin, R.P., 2009. Devil or Angel? The role of speculation in the recent commodity price boom (and bust). *Journal of Agricultural and Applied Economics*, 41(2):377–391.
- Jabjone, S., and Wannasang, S., 2014. Decision support system using artificial neural network to predict rice production in Phimai district, Thailand. *International Journal of Computer and Electrical Engineering*, 6(2):162–166.
- Jordaan, H., and Grove, B., 2010. Factors affecting forward pricing behaviour: Implications of alternative regression model specifications. *South African Journal of Economic and Management Sciences*, 13(2):113–122.
- Jordaan, H., Grové, B., Jooste, A., and Alemu, Z.G., 2007. Measuring the price volatility of

- certain field crops in South Africa using the ARCH/GARCH approach. *Agrekon*, 46(3):306–322.
- Khamis, A., Nabilah, S., and Binti, S., 2014. Forecasting wheat price using backpropagation and NARX Neural Network. *The International Journal of Engineering and Science*, 3(11):19–26.
- Khashei, M., and Bijari, M., 2010. An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with Applications*, 37(1):479–489.
- Khashei, M., and Bijari, M., 2011. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing Journal*, 11(2):2664–2675.
- Larose, D.T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. 2nd ed. New Jersey: John Wiley & Sons, Inc.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- Mayer-Schonberger, V., and Cukier, K., 2013. *Big Data*. London: John Murray.
- McAfee, A., and Brynjolfsson, E., 2012. Big Data: The management revolution. *Harvard Business Review*, 90(10):61–68.
- Mofokeng, M., and Vink, N., 2013. Factors affecting the hedging decision of maize farmers in Gauteng province. In: *4th International Conference of the African Association of Agricultural Economists, Tunisia*.
- O’Neil, C., and Schutt, R., 2014. *Doing Data Science*. 1st ed. Sebastopol: O’Reilly Media, Inc.
- Power, D.J., 2014. Using ‘Big Data’ for analytics and decision support. *Journal of Decision Systems*, 23(2):222–228.
- Provost, F., and Fawcett, T., 2013a. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59.
- Provost, F., and Fawcett, T., 2013b. *Data science for business*. Sebastopol: O’Reilly Media, Inc.
- Qi, M., and Zhang, G.P., 2008. Trend time-series modeling and forecasting with neural networks. *IEEE Transactions on Neural Networks*, 19(5):808–816.
- Ruta, D., 2014. Automated trading with machine learning on Big Data. In: *2014 IEEE International Congress on Big Data*. Anchorage: IEEE Computer Society.
- Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operation Research Society*, 66(8):1352–1362.

- Trostle, R., 2008. *Global agricultural supply and demand: Factors contributing to the recent increase in food commodity prices*. [online] Available at: <<http://www.ers.usda.gov/publications/wrs0801/>>.
- Tsadiras, A.K., Papadopoulos, C.T., and O’Kelly, M.E.J., 2013. An artificial neural network based decision support system for solving the buffer allocation problem in reliable production lines. *Computers & Industrial Engineering*, 66(4):1150–1162.
- Tsay, R.S., 2010. *Analysis of financial time series*. New Jersey: John Wiley & Sons, Inc.
- Venter, M.M., Strydom, D.B., and Grové, B., 2013. Stochastic efficiency analysis of alternative basic grain marketing strategies. *Agrekon*, 52:46–63.
- Wilamowski, B.M., 2009. Neural Network architectures and learning algorithms. *IEEE Industrial Electronic Magazine*:56–63.
- Wiles, P.S., and Enke, D., 2014. Nonlinear Modeling Using Neural Networks for Trading the Soybean Complex. *Procedia Computer Science*, 36:234–239.
- Wright, B.D., 2011. The economics of grain price volatility. *Applied Economic Perspectives and Policy*, 33(1):32–58.
- Wright, B.D., 2014. Data at our fingertips, myths in our minds: recent grain price jumps as the perfect storm. pp.538–553.
- Yu, S., and Ou, J., 2009. Forecasting model of agricultural products prices in wholesale markets based on combined BP neural network-time series model. *2009 International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2009*, 1:558–561.
- Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175.
- Zou, H.F., Xia, G.P., Yang, F.T., and Wang, H.Y., 2007. An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. *Neurocomputing*, 70:2913–2923.