2013

# Mining climate data for shire level wheat yield predictions in Western Australia

Yunous Vagh
*Edith Cowan University*

# Edith Cowan University

# Copyright Warning

**Edith Cowan University**

# Mining Climate Data for Shire Level Wheat Yield Predictions in Western Australia

**A research thesis submitted as fulfillment of the requirements for**

**the degree of Doctor of Philosophy**

**Yunous Vagh BSc Hons**

**Faculty of Computing, Health and Science**

**School of Security and Computer Science**

**Principal Supervisors: Dr. Jitian Xiao**

**Professor Craig Valli**

**Dr. Mike Johnstone**

**2013**

# USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

# ABSTRACT

Climate change and the reduction of available agricultural land are two of the most important factors that affect global food production especially in terms of wheat stores. An ever increasing world population places a huge demand on these resources. Consequently, there is a dire need to optimise food production.

Estimations of crop yield for the South West agricultural region of Western Australia have usually been based on statistical analyses by the Department of Agriculture and Food in Western Australia. Their estimations involve a system of crop planting recommendations and yield prediction tools based on crop variety trials. However, many crop failures arise from adherence to these crop recommendations by farmers that were contrary to the reported estimations. Consequently, the Department has sought to investigate new avenues for analyses that improve their estimations and recommendations .

This thesis explores a new approach in the way analyses are carried out. This is done through the introduction of new methods of analyses such as data mining and online analytical processing in the strategy. Additionally, this research attempts to provide a better understanding of the effects of both gradual variation parameters such as soil type, and continuous variation parameters such as rainfall and temperature, on the wheat yields.

The ultimate aim of the research is to enhance the prediction efficiency of wheat yields. The task was formidable due to the complex and dichotomous mixture of gradual and continuous variability data that required successive information transformations. It necessitated the progressive moulding of the data into useful information, practical knowledge and effective industry practices. Ultimately, this new direction is to improve the crop predictions and to thereby reduce crop failures.

The research journey involved data exploration, grappling with the complexity of Geographic Information System (GIS), discovering and learning data compatible software tools, and forging an effective processing method through an iterative cycle of action research experimentation. A series of trials

was conducted to determine the combined effects of rainfall and temperature variations on wheat crop yields. These experiments specifically related to the South Western Agricultural region of Western Australia. The study focused on wheat producing shires within the study area. The investigations involved a combination of macro and micro analyses techniques for visual data mining and data mining classification techniques, respectively.

The research activities revealed that wheat yield was most dependent upon rainfall and temperature. In addition, it showed that rainfall cyclically affected the temperature and soil type due to the moisture retention of crop growing locations. Results from the regression analyses, showed that the statistical prediction of wheat yields from historical data, may be enhanced by data mining techniques including classification.

The main contribution to knowledge as a consequence of this research was the provision of an alternate and supplementary method of wheat crop prediction within the study area. Another contribution was the division of the study area into a GIS surface grid of 100 hectare cells upon which the interpolated data was projected. Furthermore, the proposed framework within this thesis offers other researchers, with similarly structured complex data, the benefits of a general processing pathway to enable them to navigate their own investigations through variegated analytical exploration spaces. In addition, it offers insights and suggestions for future directions in other contextual research explorations.

# DECLARATION

*I certify that this thesis does not, to the best of my knowledge and belief:*

*(i.)* *incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher education;*

*(ii.)* *contain any material previously published or written by another person except where due reference is made in the text of this thesis;*

*(iii.)* *contain any defamatory material; or*

*(iv.)* *contain any data that has not been collected in a manner consistent with ethics approval.*

*I also grant permission to the Library at Edith Cowan University to make duplicate copies of my thesis as required.*

*Student signature*

*Date* *22 November 2013*

# ACKNOWLEDGMENTS

I am grateful in the first and last instance to God, the Creator and Fashioner who has endowed me with all faculties; physical, mental and spiritual. This is both general and specific.

I realize that my family has supported me despite my research consuming so much time and attention that would have otherwise been reserved for them. Their patience and support were invaluable in an arduous journey that inevitably comes with undertaking doctoral research. This is a debt that can never be repaid and which warrants no less than my eternal gratitude.

The primary acknowledgement is to Dr. Jitian Xiao who, as a supervisor, was a academically kindred spirit in this project. Special thanks accrue to him for agreeing to take me despite other demands on his research time, and for providing guidance, encouragement and assurance, particularly when the end of the project was near. Very special thanks are also directed to Professor Craig Valli for the belief and support he afforded me especially during the challenging times in 2011. In addition, Dr. Tapan Rai, the Faculty of Health, Engineering and Science statistician provided valuable assistance with design and analysis issues.

I am also grateful to Phil Goulding from the Department of Agriculture and Food WA (DAFWA) for supplying information needed for the project, as well as to Dr. Dean Diepeveen for facilitating cooperation from DAFWA. Thanks go to Edith Cowan University for hosting this research and specifically to the Graduate Research School, for their numerous orientation and training services and general support for researchers. In this regard, special thanks are afforded to Dr. Greg Maguire for his invaluable editorial advice.

Finally, thanks must be directed to my fellow students who provided much needed companionship on this sometimes lonely journey. In this regard, I specifically acknowledge Dr. Usman Farooq Kayani, Dr. Sunsern Limwiriyakul, Dr. Panida Subsorn, Vinh Dang, Rajeswari Chelliah, Pervaiz Ahmed, and Samaneh Rastegari.

# LIST OF SUPPORTING PUBLICATIONS

1. Vagh, Y. (2012). The application of a visual data mining framework to determine soil, climate and land-use relationships.3rd International Science, Social Science, Engineering and Energy Conference ISEEC-2011, Thailand, Feb 2-5, 2012, NPR University, 37-42. Upgraded to journal article in Procedia Engineering, 2012, 32(2012), Elsevier, 299-306.

2. Vagh, Y. (2012). An Investigation into the effect of stochastic annual rainfall on crop yields in South Western Australia. International Conference of Knowledge Discovery 2012, Bali, May 26-27, IACSIT, 227-232. Upgraded to article in International Journal of Information and Education Technology, 2012, 2(3), IACSIT, 227-232.

3. Vagh, Y., Armstrong, L., & Diepeveen, D. (2010). Application of a data mining framework for the identification of agricultural production areas in WA. Proceedings of the 14th Pacific-Asia Conference Pacific Asia Knowledge Discovery and Data Mining 2010, Hyderabad, Jun 21-24, 2010, 11-22. Upgraded to journal article in Edith Cowan University Research Online, 2010, 1(2010), ECU Publications 1-11.

4. Vagh, Y., & Xiao, J. (2012). Mining temperature profile data for shire-level crop yield prediction. International Conference on Machine Learning and Cybernetics, ICMLC-2012, Xian-China, Jul 15-17, IACSIT, 77-83. Upgraded to article in IEEE journal, 2012*, 39*(2), IACSIT, 301-309.

5. Vagh, Y., & Xiao, J. (2012). A data mining perspective of the dual effect of rainfall and temperature on wheat yield, International Conference on Information and Computer Technology, ICICT 2012, Beijing, Sep 15-16, IACSIT, 358-364. Upgraded to article in International Journal of Computer and Communication Engineering, 2012, 1(4), IACSIT, 358-364.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1

# INTRODUCTION

This chapter provides a concise view of the justification and warrant for the research endeavour. It contains the overview of the thesis wherein the specific aims of the research are outlined. It also establishes the historical background, upon which this research was founded. It raises the significance of the study in terms of niche, objectives, scope and contributions. It then concludes by stating the relevant research questions necessary in order to focus the study and direct the answering of them towards the research outcomes.

## 1.1 Research need

Agriculture contributes significantly to the Australian economy through its contribution to food exports, for instance almost 90 percent of wheat grown in Western Australia is exported (Craik & MacRae, 2010; Pimentel, 2009; Rola-Rubzen, Storer, & Pringle, 2005). The main driving forces that influence food production are climate change and the reduction in the availability of agricultural land (von Braun, 2007). These factors have to be accounted for in food production to reduce their negative effects. These optimisations are necessary for both the present state and the future of the crop growing industry.

Investigation of how to improve the productivity and sustainability of agricultural production systems is one strategy to tackle the challenges of climate change and available agricultural land. Numerous research experiments and field trials have been undertaken to analyse how land use, soils, climate and agronomic practices influence farm production systems in Australia (Anderson, 2010; Asseng & Pannell, 2012; John, Pannell, & Kingwell, 2005; Zaicou-Kunesch et al., 2010). Such studies

described by Zaicou-Kunesch et al (2010) have demonstrated that the outcomes are highly contingent upon the analytical tools employed. Until just over a decade ago, only statistical techniques were used (Cooper, Brennan, & Sheppard, 1996; A. Smith, Cullis, & Gilmour, 2001). However, the use of tools such as data mining and other analytical techniques has become increasingly important in crop prediction and decision making (Drew, 2010). This is especially crucial in relation to farmers dealing with short-term seasonal variability.

Agricultural agencies such as the Department of Agriculture and Food of Western Australia (DAFWA) have terabyte sized datasets sourced from various databases and data warehouses (Hsu, 2002). Data Mining (DM) will provide them with the capability to find patterns and trends hidden in large amounts of data through the use of specialised algorithms (Han & Kamber, 2011). The discovery process may also be supported by Visual Data Mining (VDM) where data is explored visually by humans, especially in instances where algorithms fail to uncover meaning (Keim, 2002).

For this purpose, the use of a systematic approach, that incorporates a number of technologies and methods such as statistics, DM, VDM, data processing, Data Warehousing (DW), Online Analytical Processing (OLAP), frameworks and data dissemination, was considered a viable approach. Moreover, it represents the logical next step to analysis improvement in terms of crop predictions based on historical data.

## 1.2 The background to the research

Collection and analysis of data are integral components of agricultural science research, and the historical progression of relevant analytical methods underpins this research.

Currently, data analysis is achieved through traditional statistical methods within Australia (Armstrong, Diepeveen, & Vagh, 2007). However, the identification of potential and latent patterns from different and multi-

sourced agricultural datasets has been challenging (Bocchia & Castrignanò, 2007). Agricultural data is recorded as collections from Crop Variety Trials (CVT) and grain breeding programs in Western Australia and elsewhere (Burgess & Lamond, 2010; NIAB, 2010; Zaicou-Kunesch et al., 2010). In addition to plant characteristics and traits, location information, climate specific datasets (rainfall, temperature, humidity) and attributes of soil profiles are routinely gathered. Each trial yields location specific information that is used in the prediction and crop variety recommendation process (Lauer, 1995; Matthews & McCaffery, 2012). DAFWA also gathers yield figures for specific crops such as lupins, barley, wheat, oats and canola, grown in each shire in the South West Agricultural zone (van Gool, 2011).

These efforts by DAFWA and other agricultural agencies, have concentrated on yield monitoring and crop quality assessment methods similar to those of Colvin and Arslan (2000). There has been some work on quantifying the effect of soil variation through the use of Variable-Rate Application (VRA) of treatments and inputs worldwide (Adamchuk, Hummel, Morgan, & Upadhyaya, 2004; Godwin & Miller, 2003). However, such investigations have been limited in WA (Wong, Corner, & Cook, 2001). Precision Agriculture (PA) is a strategy of Site Specific Crop Management (SSCM) that farmers employ to ensure that data, information, knowledge and best practices are managed for their specific locations, regions and zones (McBratney, Whelan, Ancev, & Bouma, 2005). Studies using PA techniques have focused on the analysis of production in terms of paddock scale zone management (McBratney et al., 2005).

It is within the spectrum of the aforementioned research that this study is situated. The emphasis of the research in this thesis shifts from yield monitoring and quality assessment to insights into SSCM especially for predictions of crop yields at specific shires. This is because there have been crop failures in the past when farmers followed crop recommendations and statistical predictions based on historical records produced by DAFWA. This research therefore, endeavours to fill the gap in the research spectrum described above. It is achieved by the

compilation and analysis of composite information on crop prediction based on site specific soil composition, rainfall and temperature, and crop yield measurements at a shire and farm level.

## 1.3 The significance of the research

The importance of this research is framed within the context outlined above and its potential to influence decision making, planning and crop planting recommendations. Specifically, this significance is demonstrated in several areas in terms of research niche, objectives, scope, benefits and contributions.

### 1.3.1 Research niche and objectives

Analysis of previous studies in the agricultural research spectrum has revealed that there was a *need for a new approach* to complex, data-driven analyses in crop planting recommendations. Historically, various analytical approaches have been used to obtain coherence and '*best fit*' results in agricultural information systems (Besemer, et, & al., 2004). Initially, there was an emphasis on yield monitoring e.g. (Calvin & Arslan, 2000) through the use of traditional statistical methods (Armstrong et al., 2007). Data mining was then introduced to improve quality assurance (Cunningham & Holmes, 2001). These methods were supplemented by OLAP for both yield monitoring and quality assessment of future crops (Abdullah, 2009; Abdullah & Ansari, 2005a). The OLAP data cube was also used to aggregate and summarise the data for sustainability of farming practices (Dunstan, 2009). The analysis of soil profile variations (Adamchuk et al., 2004) and agricultural zone management (Castrignanò, 2010; McBratney et al., 2005) are examples of other widely used approaches.

However, little work has been done to incorporate farming strategies into formal Decision Support Systems (DSS), thus rendering such an approach historically significant (McBratney et al., 2005). A new approach, along

these historical lines, that introduced advancements in technology such as DM, OLAP and DW to the analyses, was needed. This approach is novel, in that it builds on previous studies in Australia, Pakistan and India through use of a combination of DM, OLAP statistics, GIS and DW technologies to predict wheat crop yields. Another innovation is the specific visualisation and interrogation of agricultural data to determine the effect of rainfall, temperature and soil type upon the crop yields in specific shire locations. The main objectives of the thesis are to minimise crop failures resulting from farmers following inappropriate recommendations based on predictions using statistical projections.

### 1.3.2 Scope and study area

The aim was to investigate the effect of stochastic rainfall and temperature data on the wheat crop yields for shires within the study area through the use of innovative analytical technologies. Whilst the objectives outlined in the previous sub-section establish the perimeter and generality of the research, the parameters and scope of the research define its specificity.

The physical setting of the research was the agricultural crop growing region of the South West Agricultural region of Western Australia. The specific study area was a region that spanned from Esperance in the east, to Busselton in the west and south and to Perth in the north. This study area extent is shown in Figure 1-1. The climate data sourced from DAFWA, comprised a sparse dataset that contained rainfall, temperature, radiation and evaporation measurements obtained from local weather stations. The GIS for the study area terrain was in the form of soil composition, land-use, vegetation and elevation.


**Figure 1-1. Google satellite map showing study area rectangle**

## 1.4 Benefits

Several benefits should accrue from of this research. The research has implications in its benefits to farmers, its impact on global food production, and its significance when viewed as a response to climate change. These are described in the following sub-sections.

### 1.4.1 Benefit to farmers

The study is significant for the agricultural industry in general and for farmers specifically. The quality and relevance of agricultural information is crucial for farmers who need accurate predictions of crop yield to help make strategic decisions. Farmers frequently receive conflicting information about answers to the following questions (Beard, Gray, & Carmody, 2010; Hayman, 2004).

1. What is the probability of success of a given crop for a particular region?
2. How does climatic change affect the likely success of an existing crop in a particular region?

3. Which regions and crops are most sensitive to climatic change variables such as reduced rainfall? and
4. Which areas need to be equipped with irrigation, salinity control or other artificial inputs to make them suitable for a given crop?

Consequently, it is important to improve the accuracy and quality of the information used to answer these questions, especially for the first three. Answers to them will allow farmers to make better and more informed decisions about crop-region suitability (Thomas, Skinner, Fox, Greer, & Gubler, 2002), as well as crop quality and yield.

Currently, there are few formal DSS's, especially for tailored strategies that underpin management practices (McBratney et al., 2005). One innovation to help farmers manage climatic uncertainty through production forecasting and tactical management practices is Yield Prophet (Brennan et al., 2007). Other DSS tools are used in WA, including WA Wheat (yield), TACT (soil type and rainfall) and SYN (nitrogen). Information on other tools used by DAFWA can be found on the Department of Agriculture and Food's website under *farming systems* for other aids (DAFWA, 2009).

Research studies have also suggested that although farmers are experienced in *short-term*, tactical forecasting of seasonal weather conditions, they can lack in the ability to do *long-term*, climatic strategic forecasting of changing climate conditions (Asseng, 2010; Ziervogel, P, Matthew, & Mukheibir, 2010). The use of this proposed DM system of analysis may assist in crop yield predictions using both weather and climate variables for both short-term tactical and long-term strategic responses to climate change.

### 1.4.2 Response to climate change

Climate change is a key issue that Western Australian farmers are confronted with (Pitman, Narisma, Pielke Sr, & Holbrook, 2004). However,

the ability to predict certain agricultural trends is hindered by the paucity of historical rainfall, temperature and humidity profiles (Hughes, 2003). Any variation in rainfall and temperature predicted for this region is likely to have a significant impact on agricultural production (Ummenhofer, Gupta, Pook, & England, 2007). Adaptation to new agricultural methods and practices is required to survive the negative effects of volatile seasonal and climatic changes (Howden et al., 2007). It follows therefore, that changes in rainfall and temperature throughout the study area will influence wheat yields. The knowledge and understanding of these climate changes, and the cognisant adjustment to crop production strategy effects, constitute a practical response to climate change.

### 1.4.3 Impact on food production

This study has certain implications for global food production. According to the United Nations, the current world population is over 7 billion people (Dadax, 2013). As "*grain crops supply approximately 80% of the total food produced worldwide*" (Pimentel, 2009), it is very important that farming practices and technologies are optimised throughout the world to help meet this demand for both current and future populations (Foulkes et al., 2011). Initiatives such as precision agriculture, site specific crop management and decision support systems should be employed holistically to meet the challenges and demands of world food production and sustainability. Given the gravity of the supply and demand imbalance, any improvements to the quality and quantity of food crop yields will contribute to solving the food shortage problem (McInerny, 2002).

## 1.5 Main Contributions

This research contributes to the knowledge area as a result of the techniques and approaches used and developed. This is manifested in the construction of the DM framework, the associated evaluation metric,

the investigative experiments and associated crop models and processing algorithms.

- **The DM framework and evaluation metric**

The development of *a* systematic framework constituted a new approach to dealing with data complexity and its processing into usable information. Frameworks have been used before in the analysis of agricultural datasets (van Ittersum et al., 2008). However, this framework went beyond the linkage of components to be an algorithm employed as a trend and crop prediction tool, and for use in Site Specific Crop Management at a specific shire level.

In conjunction with the framework, an evaluation metric was developed to test its theoretical effectiveness and helpfulness. Several theories, approaches and usage fundamentals were used to formulate the evaluation metric including the interconnectedness of data, the basics of gauging usefulness, aspects of Human Computer Interaction (HCI), and education theory. Researchers may use the metric to evaluate other frameworks, or use it as a model to develop other methods of evaluation.

- **Data driven crop models**

The contributions to industry knowledge have been through the various analytical experiments and processing algorithms introduced in this research. These analytical experiments and associated methods are novel in that they used techniques of data mining in a mix of correlation, regression and image processing. Crop models were developed using climate variables and wheat yield, based on crop producing locations. The crop models have added to the knowledge of agricultural experts in providing improved crop planting recommendations at specific locations.

The first crop model was the isolated rainfall/soil model which showed that dominant soil types may be associated with rainfall bands in particular shires within the study area.

This was followed by the rainfall/wheat yield model which demonstrated the effect of rainfall on the prediction of wheat crop yields within the shires in the selected study area. The rainfall/wheat model revealed patterns between the high rainfall and wheat crop production in these shires.

Another model was the temperature/wheat yield model which showed the effect of temperature variability on the prediction of wheat crop yields within the shires in the selected study area. Specifically, wheat crop production was sensitive to small changes in temperature at each shire location.

The final model was the rainfall-temperature/wheat yield model which demonstrated the combined effect of rainfall and temperature, on the prediction of wheat crop yields within the shires in the selected study area. This crop model showed that the high yielding shires performed better in the years when rainfall was high, partly because of the rainfall effect on temperature. Furthermore, certain locations responded differently when variations in temperature and rainfall occurred.

These crop models provide new information to the agricultural industry on crop production performance in selected shires. Consequently, they serve to refine and improve the crop planting recommendations as provided by DAFWA.

- **New processing algorithms**

Furthermore, new processing algorithms were necessary to deal with the sparse datasets. Purpose and custom-built scripts were used to generate the GIS surface into 100 hectare grid cells, and interpolate rainfall and temperature measurements for the study area. These algorithmic scripts

are dynamic and re-usable as they can be made to match any grid cell size in any area within the world.

## 1.6 Research questions

The following research questions were determined to be pertinent to the current research direction in pursuit of a directed outcome.

### 1.6.1 Main research question

In what ways might the development and application of a system of analyses, that incorporates DM, improve the interrogation of agricultural land use datasets for the purpose of crop yield predictions?

### 1.6.2 Sub questions

1. What are the *necessary techniques and methodologies* that can be used to develop a systematic approach to predict crop yields within shires in the agricultural region of South Western Australia?

2. Which *issues* need to be considered in the application of DM for predicting crop yields from agricultural land-use and climate data in WA from a Western Australian shire level perspective?

3. Can the application of DM in crop yield predictions be used to determine appropriate *future land-uses* in Western Australia?

These research questions were used to formulate research activities designed to address the issues raised by them. These associated activities are outlined in Chapter Three. Prior to commencing the literature review, a brief explanation of the succeeding chapters is given in the following section.

## 1.7 Thesis structure

This research document is organised into a compendium of interconnected chapters. The author has endeavoured to conform to the rules of *Aristotelian statis theory* where the four basic categories of definition, causation, evaluation and policy were explored (Carter, 2000). The remaining 8 chapters of this work are therefore progressively organised as follows:

**Chapter 2.** In this chapter all the facts, information, tools and previous work related to the study are explored in concise detail as the literature review. The key concepts explored, relate directly to the research aims. They cover topics of relevance including data mining, statistics, online analytical processing, data warehousing, geographical information systems and processing frameworks.

**Chapter 3.** This chapter covers the essential methodology. It espouses the mechanics of how the research was structured for implementation. It is basically a guideline and a roadmap of the activities and processes and justifies the choices made in the research design.

**Chapter 4.** This chapter expands on the structural and research design through the prism of a conceptual processing framework. The design process was encapsulated in a visual graphic of the knowledge discovery and data mining process similar to the horizontal and vertical symbolism of the cross (Geunon, 2004). It covers the components and the workings of the framework.

**Chapter 5.** This chapter covers the convergence of the continuous and gradual variation entities of rainfall, land use and soil structure within the context of a geographical information system environment. It specifically examines the effect of rainfall on soil types within the crop producing shires. It explores the approach used to develop the succeeding crop models.

**Chapter 6.** This chapter is the first in the series of crop models. It is an analytical chapter that explored the effects of the first of the two main climatic variables on the wheat crop yield in the selected study area. The analyses undertaken were specifically designed to determine the effect of the stochastic average monthly rainfall on the wheat crop yield.

**Chapter 7.** This is the second in the series of analytical chapters. It explored the effects of the second of two main climatic variables on the wheat crop yield in the selected study area. The analysis was specifically tailored to determining the effect of the stochastic monthly temperature on the wheat crop yield.

**Chapter 8.** This chapter covers the final investigation and experiments for the climate crop model. The combined effect of both stochastic rainfall and temperature on the wheat crop yield at the shire level was measured. Consequently, it represents the *gestaltian* embodiment of *the whole is greater than sum of the individual parts,* as a platform for predicting future wheat predictions.

**Chapter 9.** This concluding chapter of the thesis summarises, reviews and explores any limitations and future directions.

# Chapter 2

# LITERATURE REVIEW

This section reviews literature relevant to the research study. It describes the development and use of theoretical frameworks. Alternative techniques for data analysis and interpretation, including statistical and other prediction techniques as well as data mining and online analytical processing are discussed. This is followed by sections on data warehousing and methods of reporting and information extraction. Subsequent sections detail the importance of statistics, OLAP, GIS, spatial data mining, data mining and visual data mining in an agricultural context. The focus is on the analysis of appropriate data mining and GIS software tools. Following that, a review of literature related to similar case studies is undertaken. Finally, the findings of the literature review and justification for the conducted research is outlined.

## 2.1 Introduction

There have been a number of research studies undertaken that focus on the importance of using data mining as a supplementary tool in transforming large volumes of agricultural data into meaningful information (Chien & Chena, 2008); Abdullah & Ansari,2005; Cohen, 2004; Ekasingh et al, 2005; Holmes et al,1998). Some early studies have experimented using data mining as an extension to statistical analysis; such as Holmes et.al. (1998). Later data mining studies used a standard data mining tool (Cunningham & Holmes, 2001). For example, Abdullah & Ansari (2005) reported on the use of AgroAdvisor, a specialized analysis tool for crop evaluation. Yet another study by Ekasingh et al., (2005) examined how data mining can be used in crop production predictions. Most of the early studies that used data mining as a supplement to statistics were concerned with crop yield management and crop quality assessment. Researchers

such as Dunstan and Qiang (2009) explored the possibility of finding land-use implications that matched specific requirements using genetic algorithms and OLAP techniques (Dunstan, Despi, & Watson, 2009). Dunstan, Armstrong and Diepeveen (2009) have extended the research using OLAP, by investigating the effective selection of land-use in a catchment area (Dunstan, 2009). In addition the anomalies associated with the complexities of the data were explored using OLAP data cubes (Dunstan et al., 2009). Other studies have involved investigating the spatial and temporal variability in managing agricultural zones (Castrignanò, 2010).

The abovementioned previous related research provided a background upon which to found the present study. The foundation themes were found to be DM, OLAP, data warehousing, agriculture, crop prediction, GIS and site management. In examining the previous related works, it became obvious that a set of analytical processes could be devised which could be used to investigate research problems that had a similar data complexity. A number of trial experiments with the agricultural climate and terrain characteristic profile data of the South Western Agricultural region of Western Australia were conducted. These trails indicated that the methods used could be generalised into frameworks. The frameworks could then be used with other similarly profiled data with continuous variation attribute and gradual variation geographic dimensions.

## 2.2 Frameworks

In order to understand framework formalisation, it is necessary to understand that system data do not exist in isolation but are related to other data by sharing common features (Garlin & Notkin, 1991). Although the data from different systems may have common features, they appear to be outwardly unrelated, or related in uncharacteristic and undescribed ways (Luck & D'Inverno, 2001). A framework is considered a well-structured and refinable specification that permits the identification and understanding of

common properties for the purpose of creating models from common abstractions (D'Inverno, Justo, et., & al., 1996). Frameworks conform more visibly to the constructive approach in education theory (Piaget, 1970) in that modular elements are assembled and connected together. For example, the position and orientation of the building blocks of the spatial based framework are considered to be significant in terms of the interface parameters. Alternatively, the relational approach to frameworks deals with logical and abstract considerations (Bakewell & Garbutt, 2005). Apart from these approaches to the construction of frameworks, other fundamentals and theories are important considerations for their design.

## Theoretics of frameworks

The theory and fundamentals of framework design are important for their significance in devising the metrics of evaluation of frameworks. Frameworks may be defined in terms of interfaces and human computer interaction where systems and or components combine with respect to interpretations of the spatial, relational and constructive domains (Ullmer & Ishii, 2001). In the spatial approach, the position and orientation of the building blocks of the framework are significant in terms of the interface parameters.

Frameworks should also be considered with respect to the theories of instructional design and authentic learning environments that begin to address the divide between theory and practice (Herrington & Oliver, 2000). This is due to the fact that generally, advisory bodies such as DAFWA, are essentially theorists whilst their clients, such as farmers, are the practitioners.

Situated learning, as a part of education theory has been defined by Collins (1988) as learning knowledge and skills that reflect their usefulness in real life contexts. Consequently, useful knowledge may best be acquired when a framework displays nine situated learning elements according to the general surmise of Herrington & Oliver (2000). These elements are

authentic contexts (Collins, 1988), authentic activities (J. S. Brown, Collins, & Duguid, 1989), expert performances (Lave & Wenger, 1991), multiple roles and perspectives (Bransford, Sherwood, & et.al., 1990), collaborative construction of knowledge (Young, 1993), reflection (J. S. Brown et al., 1989), articulation (Bransford et al., 1990), coaching and scaffolding (Herrington, Oliver, & Herrington, 2007) and authentic assessment (McLellan, 1993). In addition, the frameworks must be re-usable (Garlin, 1990). In order for them to achieve this property within a context, frameworks should incorporate levels of abstraction, granularity and specificity as depicted in Figure 2-1 (Greenfield & Short, 2003).

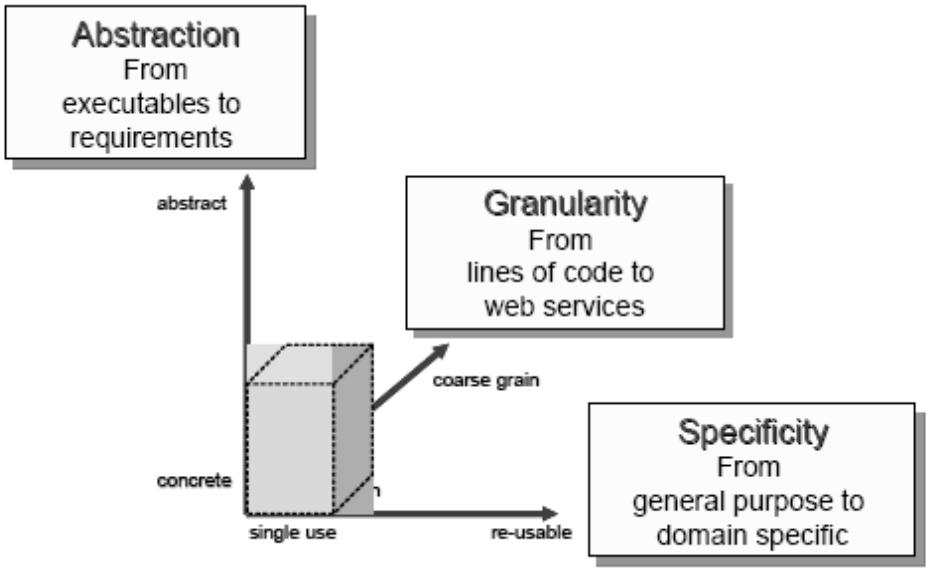



**Figure 2-1. Aspects of framework and software design (Greenfield & Short, 2003)**

Furthermore, users of the framework should find it helpful in analysing their own datasets. Kitchenhaum et al. (2005) proposed a metric for gauging the helpfulness of a framework. According to them, frameworks should be *usable* within a specific context, allow the evaluation techniques to be *validated* and should also have some *value* in terms of the benefits that may be derived for using the framework (Kitchenham, Linkman, & Linkman, 2005). Given that generally, frameworks could be characterised in the context of the nine elements of situated learning, the levels of

abstraction and the degrees of helpfulness, it is possible to identify several types of frameworks.

## Framework usages

Frameworks have been used in a number of different research application areas including software (Sadowski, 1997), in the field of pattern recognition (Kittler, 1998), for defining classification rules in a network IDS context (Agarwal & Joshi, 2000), the exploration of the parameter of 'transversal endurance' through the use of a classification hierarchy (Vassilios S. Verykios, Bertino, et., & al., 2004), geographic measurement (J. Miller & Han, 2009), agricultural policy option comparisons (van Ittersum et al., 2008) and visual data mining (Schulz, Nocke, & Schumann, 2006).

## Type of frameworks

Each framework is highly specific to the application and the domain (L. Cao & Zhang, 2007). However, it is possible to discern the type within each framework application. Frameworks can be classified as either theoretical (Mannila, 2000), component based (Berzal, Blanco, Cubero, & Marin, 2002), processing (Payyappillil, 2005), distributed data mining (Sdorra, Hafez, & Raghavan, 2001), ontological software based (Marinho et al., 2010) or any other key descriptor that defines the function of the framework.

An example of a framework in an agricultural context is the SpatioTemporal Integrated Forecasting Framework (STIFF) which is essentially a processing framework (Z. Li, Dunham, & Xiao, 2003). The logical and abstract considerations were the steps to identify the target and siblings, a time series model for location forecasting, the construction of an artificial neural network, a statistical regression method, case studies and a comparison metric. The building blocks of the current framework are

statistics, data mining and data warehousing. Essentially the current research is an example of a processing, component style framework that incorporates, amongst other things, statistics and analytical tools.

## 2.3 Traditional analytical tools

There are number of traditional analytical tools and methods used to help interpret and give meaning to data. These include statistics, probability, decision making, data analysis and various statistical methods.

Statistics is a subject that is concerned with inductive inference (Savage, 1972) made from "*the understanding of structure in data*" (Venables & Ripley, 2002). Whether one subscribes to the dictum of "*lies, damned lies, and statistics"* as attributed to Dilke (1891) or simply employing statistical methods of analysis in order to make inferences from experiments or surveys (Lymann Ott & Longnecker, 2010), it is important to recognise that assertions, statements and evidence require statistical support. However, there could be a number of reasons why statistical support could be misinterpreted. These have been enumerated by Otts (2003) in his article entitled, "*What educated citizens should know about statistics and probability*". They include making spurious cause and effect relationships, practically insignificant findings due to the difference between statistically significant and practically significant results, insufficient sample size, insidious bias in surveys, low probability in diagnostic tests and natural variability (Utts, 2003).

### Decision Making

Both statistics and probability are tools needed in order to make decisions, especially when those decisions are about making a choice from a set of possibilities in situations of uncertainty. Decision making according to Bell and Raiffa et al. (1995) depends on the philosophy and disciplinary background of the decision maker. They contend that decision makers may

be classed into three different interest groups. They list them as mathematicians or decision theorists, psychologists and methodologists. The groups are described as normative, descriptive and prescriptive respectively. Furthermore, these classifications are linked to the three related disciplines of statistics, mathematics and economics (Bell, Raiffa, & Tversky, 1995).

However, this study will be mostly concerned with the prescriptive method. This is due to its pertinence as a theory of choice dedicated to intelligent action and decision making made with the help of reason and technology using data analysis (March, 1983). There are many statistical methods employed in data analysis. Some of the methods are traditional, having been employed over the last 30 years, while others are regarded as modern methods.

## Data Analysis

Some of the traditional techniques of data analysis include normal distributions, t-distributions, chi-squared and F-distributions as well as analysis of variance, multivariate analysis, and linear regression (Manly, 2005). Multivariate analysis methods are in turn made up of a number of variations including factor analysis, discriminate function analysis, cluster analysis and canonical correlation. Included in this category of multivariate analysis methods are the methods of ordination where variables are plotted on axes such as principal components analysis, multidimensional scaling, principal coordinates analysis and correspondence analysis (Manly, 2005; Michalski, 2000).

## Statistical methods

The modern methods include techniques such as dynamic graphics, non-linear estimation, regression, re-sampling and other simulation type inferences (Duckworth & Stephenson, 2002). A large number of the

modern statistical methods are regression types and algorithmic. Some of these are the general additive model (Ekasingh, Ngamsomsuke, Letcher, & Spate, 2005), Generalized Unbiased Interaction Detection and Estimation (GUIDE), multivariate adaptive regression splines, Projection Pursuit Regression (PPR), Classification And Regression Tree (Carter), Neural Networks (NN) and a host of others (Wilcox, 2010). Other modern methods deal with patterns such as point process pattern analysis and synthesis (Illian, Penttinen, et., & al., 2008). There are also a whole host of methods that are related to Analysis of Variance (ANOVA) and Analysis of co-variance (ANCOVA) (Hill & Lewicki, 2006).

Apart from the aforementioned statistical methods employed in previous research studies, there have been instances where researchers have sought to augment the analytical process using a combination of methods. In this regard, there have been many studies and research articles which have relied on algorithms, complexity, statistics and probability to analyse data (Cohen, 2004). Two of the most recent supplements to statistical methods of analysis are data mining and online analytical processing. While DM is largely an automated process, OLAP is a presentation tool designed to enable manual knowledge discovery (Abdullah, Brobst, Pervaiz, Umer, & Nisar, 2004). The following sections will therefore elaborate on DM and OLAP and their relevance in this research.

## 2.4 Data mining

There are many analytical approaches to the knowledge discovery process such as machine learning, data mining, neural networks, and genetic algorithms (Wiemer & Prokudin, 2004). In essence, DM may be regarded as having formed as a result of a convergence of the three type of technologies which include computing power, statistical tools and data warehousing (Hill & Lewicki, 2006). Data mining and knowledge discovery through databases (KDD) are considered to be related terms, with data

mining being a generalization and KDD a specific reference to the technology (Rodriguez, Carazo, & Trelles, 2004). In data mining jargon, it is possible to make the cyclic inference that KDD is the “antecedent” term and data mining is the “consequent”. Nevertheless, in many ways according to Rodriguez et al. (2004), they are considered to be synonymous. Generally, KDD is “*the process of extraction and abstraction of any type of pattern, perturbation, relationship or association from analysed data*” (Rodriguez et al., 2004, p 494). Another definition according to De Falco et al. (2005) is the description of data mining as the process of searching through large databases in order to extract useful information that would have otherwise escaped investigation. These large databases typically contain extremely large amounts of multivariate data and a search is done through automatic means (Abdullah, et al. 2004). In addition, the useful information usually manifest as patterns from which inferences may be made.

In summary, data mining is an automated prediction and analytical process which is involved in the transformation of data into useful information. This is achieved through uncovering latent patterns which are hidden in enormous amounts of related data available in various databases and data warehouses. The artefacts which are produced from the data mining process may then be utilised either in automated decision support mechanisms, or assessed manually by decision makers (Fernandez, 2003). The exponential growth of available data has been made possible by an ever increasing capability for data collection and storage capacity in computer hardware advancements in technology. This development has occurred over the last two decades and is due to the availability, affordability and effectiveness of computer storage devices, in particular, the hard-drives and the associated read-write access times (DeFalco, Della-Cioppa, & et.al., 2005).

These advances in data availability through facilitated storage, data availability and data diversity have resulted in an increase in the complexity and inter-relationships of the data entities. Consequently, traditional user-driven analysis of the storehouses of data by statisticians has become

increasingly difficult, creating a demand for an automation of the process (DeFalco et al., 2005). Furthermore, the latent patterns hidden within the data may not be uncovered through simple statistical means (X. Zhang, Pan, & Wang, 2008). Data mining solves these problems by providing a richer set of tools capable of discovering the patterns and inferring new knowledge. This fact has seen the proliferation of data mining as an emergent pattern matching and prediction tool (Seifert, 2006).

Data mining is a developing science and it can be defined and categorized in a number of ways depending on the specific knowledge domain. For example, this has manifest in the domain of biological science where the technology of data mining has been applied successfully and categorized as bioinformatics (Yang, Adelstein, & Kassis, 2009). Various techniques have been employed within bioinformatics to filter out the useful data to gain valuable information (Frédérique Lisacek, Cohen-Boulakia, & Appel, 2006). These include the processes of comparing sequences, constructing evolutionary (phylogenetic) trees, detecting patterns in sequences, determining 3D structures from sequences, inferring cell regulation, determining protein function and metabolic pathways, assembling DNA fragments and using script languages (Cohen, 2004).

Of these, only the first three relating to comparing sequences, detecting patterns in the sequences and the construction of evolutionary trees are relevant for the present DM study. The interpretation and evaluation of these patterns allows the user to gain new knowledge as depicted in Figure 2-2 where raw data is selected, processed, transformed and analysed. This is in order to reach a point where the data is progressively converted into information and knowledge as part of a continuum of knowledge discovery and data mining (KDDM). The diagram in Figure 2-2 is a creation of one of the two most renowned authors in DM as denoted in the caption. These two groups of authors are Witten & Franke and Han & Kamber who have both recently completed 3rd editions of their respective books in 2011.

In order to understand the internal mechanics of the DM and the KDD process, it is necessary to have a perspective of the underlying data mining techniques.

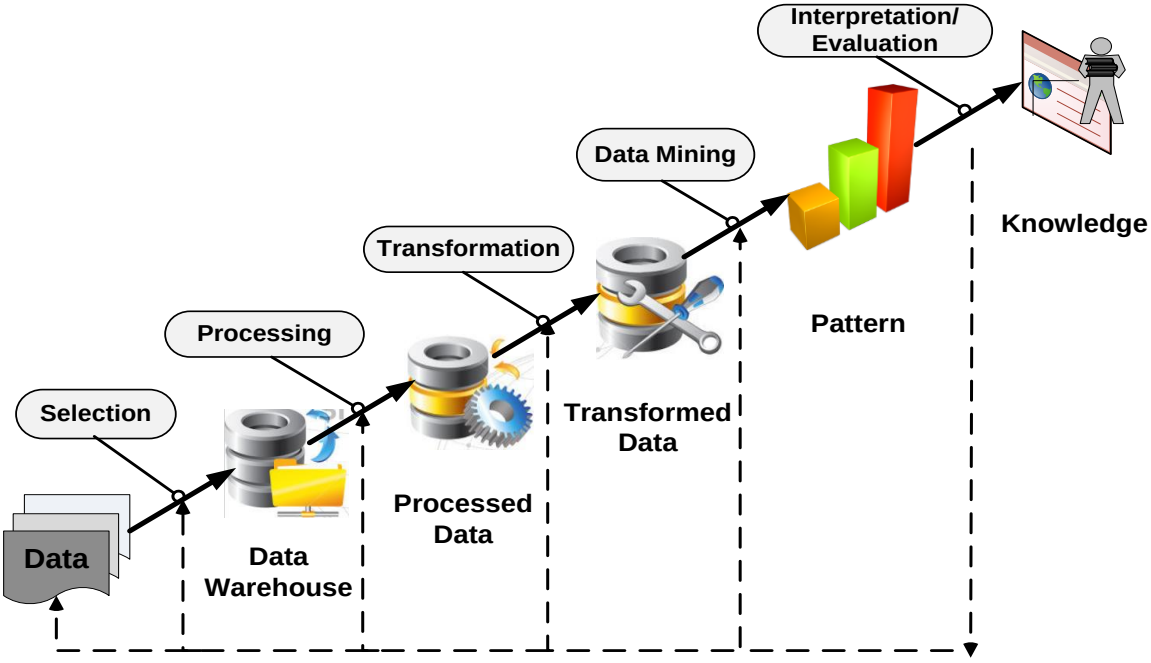



**Figure 2-2. A re-drawn pictorial representation of the KDD process (Han & Kamber, 2011; Witten, Franke, & Hall, 2011)**

### 2.4.1 Data mining techniques

Data mining techniques can be grouped into the categories of the problem solving approach such as statistical methods, Case Based Reasoning (CBR), Neural Networks (NN), decision trees, rule induction, Bayesian Belief Networks, Generic Algorithms (GA) and evolutionary programming, fuzzy sets and rough sets (Goebel & Gruenwald, 1999).

The process of data mining can be further categorized by the steps used to carry out the process. They are the four techniques of association rules, classification rules, clustering and numeric prediction (Witten et al., 2011). The technique of association will be dealt with first.

## Association

Association rules are used to "*show the relationship between a set of antecedents and its associated consequents*" (Rodriguez et al., 2004, p 494) and are used to "*predict any attribute, not just the class*" (Witten & Frank, 2005). Association rules, unlike classification rules are not combined for use as a set. Furthermore, association rules have validity that can be expressed through its support, which is the number of instances for which an association rule applies and predicts correctly (Witten & Frank, 2005). The support or coverage can be shown to have an accuracy or confidence. The accuracy is the ratio of the number of correct predictions of the rule to all the instances for which the rule applies (Witten & Frank, 2005).

In instances where the accuracy exceeds the normal percentage, there is the measurement of improvement. This indicates the strength of the shift from normalcy (Rodriguez et al., 2004). A classical association rule example from Market Basket Analysis is "*89% of the customers that purchase bread and milk also purchase sugar.*" In this case, it could be inferred that the purchase of bread and milk also means the purchase of sugar with 89% confidence (Rodriguez et al., 2004, p 495). Here the proportion of the purchases that relate to the rule is its support, and if the measure is more than what is statistically expected then this is considered to be the improvement.

A large variety of algorithms have been developed to mine association rules. Nevertheless, the first step in this process is to determine the frequency of item sets that exist above the user-specified minimum support (Rodriguez et al., 2004). Instances arise where there is a need to "*find rare and sparse data associations*" for which the support requires lowering, thereby making for inefficient algorithms (Rodriguez et al., 2004, p 494).

Normal mining of association rules is likely to produce redundancy (Zaki, 2004). As the number of frequent item sets grows larger, there is a corresponding increase in the number of rules produced with a high percentage being redundant. This holds true even in the case of sparse

datasets (S. Kotsiantis & Kanellopoulos, 2006; Zaki, 2004). Several early attempts have been made to eliminate redundancy. These include the pruning of the discovered association rules by forming rule covers, or mining rules of interest only, through the use of user constraints like metrics of interestingness (L. Liu, Y, Shan, & Yin, 2008). These, according to Zaki (2004), do not successfully address the issue of rule redundancy. The approaches to reducing redundancy have been in terms of reducing the size of the item set (closed frequent item set) and in enhancing the pruning (Lucchese, Orlando, & Perego, 2006).

A closed frequent item set is a smaller subsection of a frequent item set (Mielikainen, 2003). Closed frequent item sets are explored through the use of algorithms. Studies by Zaki (2004) have concluded that using closed frequent item sets reduces the number of redundant rules exponentially even when applied to large dense datasets. This conclusion was reached following work by researchers like Zaki who proved that algorithms like CHARM, outperform others like AClose and Apriori for mining all closed frequent item sets (Zaki, 2004). The CHARM algorithm has been shown to surpass other closed frequent item set "*algorithms like Closet (Pei et al., 2000), Mafia (Burdick et al., 2001) and Pascal (Bastide et al., 2000)*" (Zaki, 2004, p 225).

Although researchers like Bastide et al. (2000a) have succeeded in extracting minimal association rules, extraction of non-redundant rules "*leads to different, mutually complementary, notions of smaller association rule sets*" (Zaki, 2004). Non-Redundant Association Rule Discovery (NRARD) significantly improves the efficiency of association rule discovery as it eliminates redundancy. However it is not the optimum. This is because the requirements for redundant rules have to be strict (J. Li, 2006) for further improvement to be possible. Optimal rule discovery produces "*rules that maximize an interestingness measure*" (J. Li, 2006) through further pruning. The Optimal Rule Discovery (ORD) algorithm by Li (2006) has been shown to exceed the performance of association rule discovery and NRARD algorithms. ORD also has significantly less computational complexity (J. Li, 2006). Optimal pruning utilizes mechanisms like support

pruning and closure pruning. Support pruning has been shown to work well in sparse density workspaces or when the minimum support for a rule is high (J. Li, 2006). However support pruning does not work well when the density is high or when the minimum support is low (Bhattacharyya & Bhattacharyya, 2007; J. Li, 2006).

While association is considered symmetric in that no attribute is assigned unequal importance in matching the rule set, classification on the other hand, is asymmetric in that a single attribute is defined to be the class (Freitas, 2000).

## Classification

Classification uses a set of pre-classified examples which are basically the precondition or antecedent. A model is developed from the preconditions which generate a set of grouping rules. These rules form the consequent from which a new object is characterized (De Falco et al., 2005).

Several classification methods are used to extract meaningful relationships in the data. They range from the most commonly used data mining technique of classification rules (Witten et al., 2011), symbolic learning implementation (Alur, Madhusudan, & Nam, 2005) to neural networks (Liao & Wen, 2007). One of the most popular classification methods is decision trees used in the case of trivial datasets, wherein the cases are successively partitioned until a single class is derived from all the subsets. In machine learning terms, decision trees are classified as predictive models (Witten & Frank, 2005). Observations about a particular entity are mapped against conclusions to the items target value. Each node apart from the root corresponds to some variable. The arc to a child node represents a possible value of the variable. Each leaf represents the predicted value of the target variable traversed from the values of the variables in its path to the target from the root (Witten & Frank, 2005).

## Classification - decision trees

There are three types of decision trees based on the predicted outcomes. If the predicted outcome is the class to which the data relates it is known as a classification tree. In cases where the predicted outcome is a real number for a numeric prediction it is referred to as a regression tree. Lastly if both a class and a real number are the predicted outcomes, then the method is a Classification And Regression Tree (Carter) analysis (Witten & Frank, 2005). Decision trees have certain advantages over other data mining techniques. Decision trees are simple to understand and interpret. They are capable of handling both nominal and numeric data, and are easily explainable through Boolean logic in typical white box model fashion. They allow validation of the model through statistics and are capable of good performance on large data sets (Luger, 2005).

Another technique used in DM is when the relationship between the data is required to be understood in terms of proximity. This technique is known as clustering.

## Clustering

Generally clustering is a process of grouping data items into sets or clusters based on some measure of similarity (Jain, Murty, & Flynn, 1999). As some details may be omitted for clarity and simplification, clustering may be regarded as "*concise summaries of data*" (Berkhin, 2006, p 5). The output usually takes the form of a diagram that depicts the instances as clusters. Clustering is an *unsupervised* classification of patterns that is based around observations, data items or attributes (Xiong & Kumar, 2006).

The clustering process can be viewed as having distinct phases as shown in Figure 2-3. They are feature extraction, similarity, computation and grouping. Grouping can be performed in a way in which the output clusters are partitioned into groups (Moore et al.), or fuzzy where each pattern has

a varying degree of membership to each of the output clusters (Jain et al.,1999).

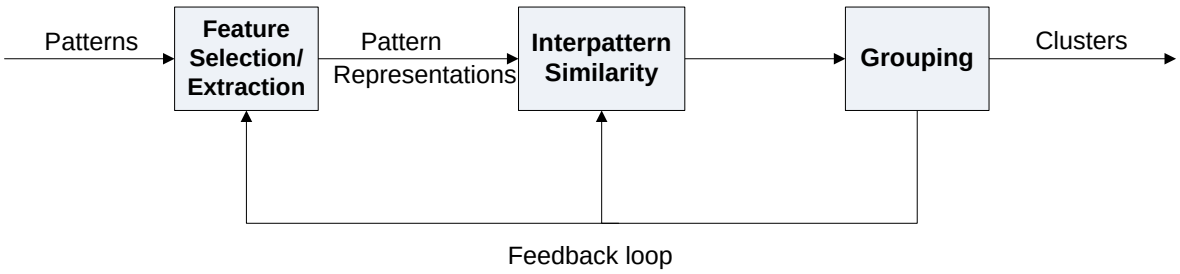



**Figure 2-3. The distinct phases of clustering (Jain et al., 1999)**

There are many types of clustering techniques, and these can be broadly classified into hierarchical, partitional or locality methods based on the type of algorithm used (Jain et al., 1999). Their use allows for different approaches to the clustering process to be implemented. This is detailed in Figure 2-4. Whatever the approach to the clustering process, the effect is similar; they all result in reducing the search space for data mining algorithms, thus making them more efficient (Abdullah & Ansari, 2005a).

The diagram in Figure 2-4 has been modified and updated from the original by Jain et al. (1999) for the purposes of clarification and currency so that it is brought in line with recent developments. Whilst the focus of clustering techniques is mainly pattern recognition through similarity, the next section deals with techniques of prediction where essentially target variables are estimated from known object variables.

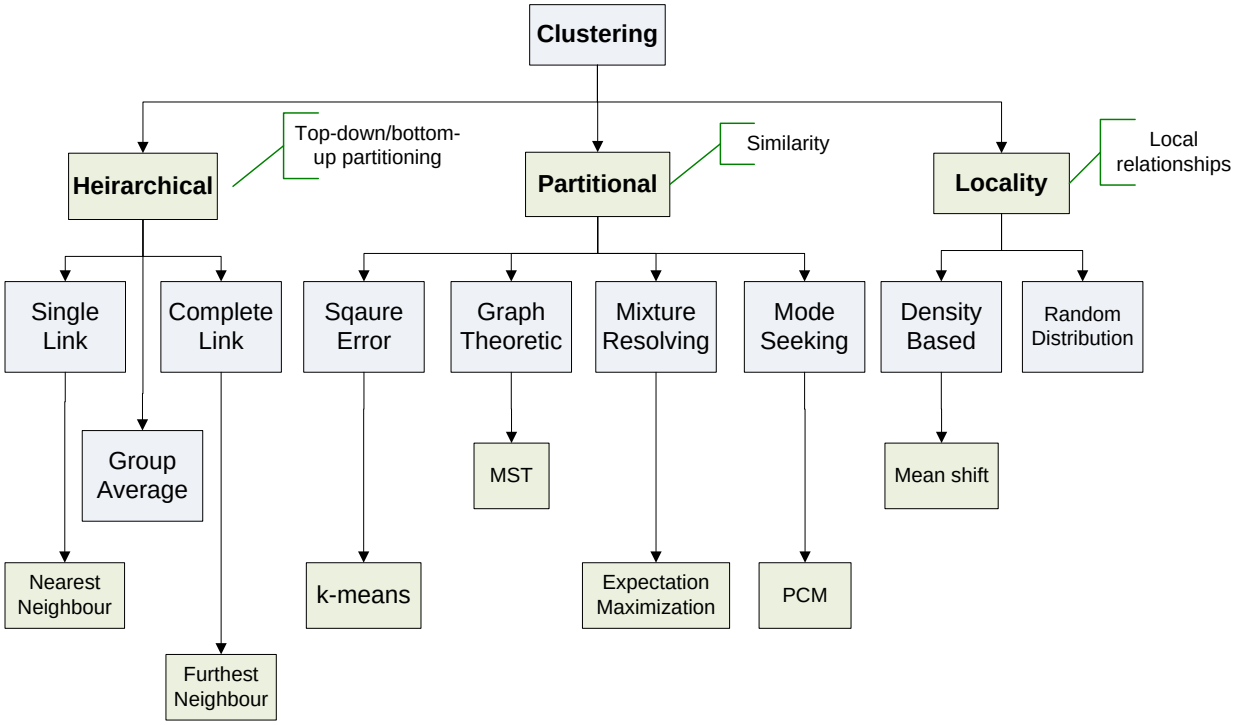



**Figure 2-4. The different approaches to clustering adapted from (Jain et al., 1999)**

## Predictive modelling

Predictive modelling or *supervised* learning is somewhat different to clustering. In this method, models are constructed with the ability to predict the value of a target attribute (dependent variable) from the given values for a set of input attributes (independent variables). In other words, the description of an object is used to predict a target property of an object (Zenko, Dzeroski, & Struyf, 2006).

There are many predictive modelling methods, most of which produce some degree of interpretable models where possible. The different models suit different types of data. They range from linear equations to logic programs (Konen, 1999). The most common of these methods that manifest in interpretable models are decision tree learning and rule learning (Zenko et al., 2006). An example of rule learning is the Generalized Linear Modelling (GLM) technique (Mosley, 2005).

The GLM technique is used to predict the value of a dependant variable from a series of independent variables. In an insurance setting, for example, the impact of insurance class of motor vehicle on loss costs could be determined. Two independent variables that do not have a constant relationship to each other are said to have interactions. The GLM model allows the discovery of these interactions automatically (Mosley, 2005). Predictive modelling is essentially one-dimensional in that only the properties of the target variable are predicted. In order to cater for prediction in more than one dimension of the target variable, another method called predictive clustering is used.

## Predictive clustering

Predictive clustering is the extension to the previous two methods of clustering and predictive modelling where the clusters are predicted in two dimensions of the target properties as well as the description of object. Figure 2-5 provides a graphical representation of the various methods of clustering.

Clustering in terms of prediction and modelling are specific to data analyses where the variable to be predicted is categorical. Quite often in analytical research, the target variable is of a numeric nature. For these instances, another type of DM called numeric prediction is used.

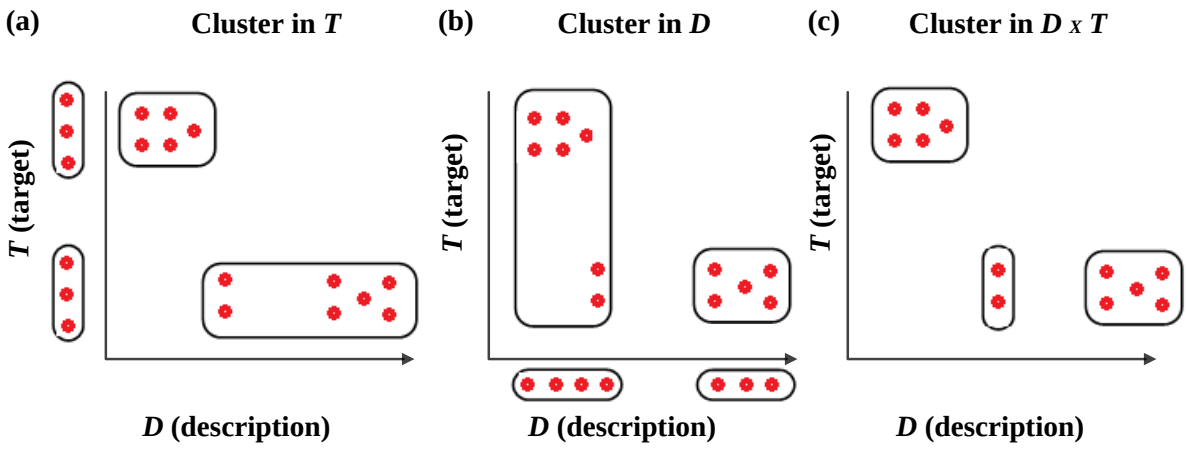



**Figure 2-5. Re-drawn predictive modelling (a), clustering (b), and predictive clustering (c) - (Zenko et al., 2006).**

### Numeric prediction (regression)

In numeric prediction, the outcome to be predicted is a numeric quantity rather than a discrete class or value (Witten et al., 2011). Numeric prediction is considered as a special type of decision tree that stores a class value that is the average value of instances that reach the leaf node. This type of tree is also called a regression tree. If on the other hand, the leaf node predicts the class value of the numeric instances that arrive there, it is called a model tree (Witten & Frank, 2005).

This now concludes the section of the data mining techniques. As the quality of output depends largely on the quality of input, it is important to refine the data without introduction of bias, prior to input. In the data mining paradigm, terminology has been introduced to cater for these eventualities. The main descriptions are “data warehousing” for the pre-selection and relevance, and “dirty data” for pre-processing.

## 2.4.2 Issues relating to effective data mining

There are a number of issues relating to effective data mining and they deal mainly with the pre-processing of the data. The quality of any analysis of data depends on the quality of the raw input data according to the dictum of Garbage In, Garbage Out (GIGO) (Baesens, 2009; Pyle, 1999; Y Vagh, Armstrong, & Diepeveen, 2010). Dirty data cleansing, data selection and data reduction are all processes applied to the data prior to it being analysed by the DM process in order to produce quality output.

- ### Dirty data

As Business Intelligence (BI) is derived from stores of data, it is important that the data is of a high quality. Unfortunately, a great deal of data is “dirty”. Dirty data can generally be characterized as missing, wrong or non-standard representations of the same data (Han & Kamber, 2011; W. Kim et al., 2001). There are also instances when data from legacy sources have inadequate or no metadata descriptors (Francis, 2005). In addition, issues

of pollution, outliers and noise (M. L. Brown & Kros, 2003) may also exist. Kim et.al (2001) described a taxonomy of the different classes of dirty data. Their original descriptions were transformed into a new table by the author. This new table details the general type of dirty data, the constraint, enforceability, symptom, specific class and example.

The data problem symptoms were categorised as classes. This is displayed in Table 2-1. The actual data problems encountered within the course of this research were of the missing and unusable nature. For example, the missing data was in relation to the sparseness of the weather stations and the associated climate variable readings. Another data problem related to the measurement units was that the yield data was recorded in tonnes per shire. This measurement had to be scaled-down to tonnes per hectare in order to standardise the wheat yield across the shires of differing size. Dirty data can be dealt with on the basis of data selection approaches.

- ## Data selection

Data selection is not a mere random process of choosing a subset of available data. Rather, the approaches to data selection have to be designed with the context of the data in mind, as if all the data was used (deHaro-Garcia, delCastillo, & Garcia-Pedrajas, 2009). According to deHaro-Garcia et al. (2009), although other approaches such as sampling (Cochran, 1997), Modified Selective Subset (MSS) (Barandela, Valdovinos, Sanchez, & Ferri, 2004), entropy based uncertainty instance selection (S. Kim, Song, Kim, & Lee, 2006), Intelligent Multiobjective Evolutionary Algorithm (Chen et al., 2005) and the LQVPRU method (Li et al., 2005), two fundamental approaches to data selection have emerged. They are the scaling-up of algorithms and the scaling-down of data (H. Liu & Motoda, 2002). Scaling down of data can be regarded as data reduction. An important aspect of scaling-down of data is the selection of relevant data prior to input to the data mining algorithms. Scaling-up of algorithms

**TABLE 2-1. TAXONOMY OF DIRTY DATA**
**ADAPTED FROM (W. KIM, CHOI, HONG, KIM, & LEE, 2001)**

| TYPE | CONSTRAINT | ENFORCEABILITY | CLASS | VIOLATION/DESCRIP | EXAMPLE |
|---|---|---|---|---|---|
| Missing | No-Null | Not existent | Missing data | | |
| | | Not enforced | Missing data | | |
| Wrong | Integrity | Automatic | | | |
| | | RDBS supported | | | |
| | | User specifiable | Wrong type | Type or value | |
| | | | Dangling data | Referential Integrity | |
| | | | Duplicated data | Non-null uniqueness | |
| | | | Mutually inconsistent data | Condition trigger | |
| | | Guaranteed through Transaction | Lost update | Concurrency control | |
| | | | Dirty read | Concurrency control | |
| | | | Unrepeatable read | Concurrency control | |
| | | | Lost transaction | Improper crash recovery | |
| | | Not RDBS supported | Wrong category | | |
| | | | Outdated temporal data | | |
| | | | Inconsistent spatial data | | |
| | | Not automatic | Single table & field | Erroneous Entry | |
| | | | | Misspelt | |
| | | | | Extraneous | |
| | | | Single table & multi fields | Wrong field entry | |
| | | | | Wrong derived field | |
| | | | | Inconsistency | |
| | | | Multiple tables/files | Value mismatch | Employee no |
| Not wrong; Unusable | | Different | Different data | | Same entity in multiple tables |
| | | Ambiguous | Abbreviation | | dr - doctor/driv |
| | | | Context | | homonyms |
| | | Non-standard | | | |
| | | Non-compound | Not auto transformable | Abbreviation | Hwy for highway |
| | | Different representation | | Alias name | |
| | | | Auto transformable | Encoding formats | |
| | | | | Representations | |
| | | | | Measurement Units | |
| | | | | Abbreviation | |
| | | | | Different orderings | |
| | | Compound | Concatenated data, | Abbreviation | |
| | | Different representation | | Special characters | |
| | | | | Different orderings | |
| | | | Hierarchical data | Abbreviation | |
| | | | | Special characters | |
| | | | | Different orderings | |

is about refinement and optimisation in relation to insufficient data. When run in parallel to the scaling-up of algorithms, a bi-directional approach is created for extracting information nuggets from the mining process (Tsang, Kwok, & Cheung, 2005).

- **Data reduction**

Generally, data is held in a flat file format with descriptors called attributes or features. The individual lines of attribute-value fields form an instance, that is known variously as a record, tuple or a data point (Cereghini & Ordonez, 2002). The scaling-down or data reduction process can be achieved through several means. They include the selection of features, discretising values, instance selection. These methods serve to lower the number of columns, the possible values and the rows in the data set respectively (Cano, Herrera, & Lozano, 2006). Thus, data reduction is a way of reducing the dataset through a constriction of the features, values and instances. In the context of the study area within the scope of this research, the features selected were the land uses, soil composition, elevation; the values were for rainfall and temperature and the instances were codes 340 and 341 which were the designated codes for cropping and cereals.

- **Feature and instance selection**

Feature selection involves pre-selecting only those attributes of interest and excluding others. This has the effect of reducing the number of columns in each record or tuple (Kim et al., 2001). With respect to the study area under investigation, feature selection was used by only selecting the grid cells that had designated land uses for the planting of crops.

Another method of data reduction is instance selection. Instance selection involves selecting an independent sample of data that models the whole, without any performance deterioration for accomplishing tasks (Liu & Motoda, 2002). Thus the central idea of instance selection is approximation

of the whole data with a view to achieving better results by removing noise through data irrelevancy (Brighton & Mellish, 2002). This process can be divided into three overlapping tasks viz., enabling instance selection, focusing and cleaning (Kim et al., 2001). Enabling achieves data reduction thereby resulting in algorithm efficiency. Focusing allows the emphasis to only relevant parts of the data that is of interest. Cleaning is achieved through the selection of relevant instances of data. This usually means that irrelevant instances together with noise (incorrect) and /or redundant data are removed, resulting in high quality data inputs to the data mining process (S. B. Kotsiantis, Kanellopoulos, & Pintelas, 2006).

The abovementioned processes of feature selection and instance selection constitute the mechanisms for scaling down of data. Scaling-up of algorithms on the other hand, deal with insufficient data and sparse databases (Bayardo, Ma, & Srikant, 2007). The issue of sparse databases is discussed in the following section.

- ## Sparse databases

Sparse databases occur where there is limited amount of data available for forming associations between the various tables. This situation is commonplace in bioinformatics where there is considerable interest to establishing the connections between rare data nodes (Rodriguez et al., 2004). According to Cui, Zhao & Yang (2010) sparse databases are typically horizontally represented and are generally characterised by limitations and deficiencies in schema evolution, column numbers, storage and performance. Although these characteristics render working with sparse databases difficult, there have been laudable efforts to minimise the limitations. Rodriguez et al (2004) have proved that deeper insights into the data are possible by finding association rules with low support but high confidence and potential maximum benefit. Their study demonstrated that these problems can be solved by algorithms through well-organized data structures being used in the search space, thereby making the explorations more efficient and intelligent (Rodriguez et al, 2004). Cui, Zhao & Yang

(2010) found that sparse databases can be effectively interrogated. They did this by using both horizontal and vertical representations of data, together with sub-space splitting in their HoVer framework. Within their framework, the SQL queries were executed vertically and the results returned horizontally.

The climate data used in this research was limited to recordings from widely dispersed weather stations thereby characterising the rainfall and temperature profiles as sparse datasets. The strategy employed to overcome this limitation was the use of ordinary krigeing which effectively interpolated the values for these variables in each of the 100 hectare cells over the entire grid surface. Thus the variables for both average monthly rainfall and temperature were scaled up using stochastic values generated through ordinary krigeing executed in the Evolution R statistical software package.

As can be seen from the preceding sections, a data analysis task may be characterised by combination of both of the down-scaling of data as well as the up-scaling of the algorithms. In addition, the data used in subsequent DM tasks may first be required in different forms or scales to the stored format. This conversion of the data into the required formats and scales is partly the domain of OLAP.

## 2.5 Online analytical processing

Online analytical processing organizes aggregate queries on data, such as sums and averages, for use in decision support algorithms. Information is collected from detail tables so that alternatives, trends and projections are able to be viewed through axis pivoting and changing aggregation (Ossimitz, 2009). The difference to data mining is that the answers to the queries are stored to provide a Graphical User Interface (GUI) that is intuitive. Although OLAP is powerful and fast, its strength depends on human intelligence. Furthermore, domain expertise by the user is necessary for the extraction of data and the conversion to valuable

information (Abdullah, Brobst, Pervaiz, et al., 2004). This requisite makes OLAP predominantly user-driven, as opposed to data mining which is data-driven. The limitation to using OLAP is therefore its dependency on the expertise of the user.

Overall, OLAP is considered to be complementary to data mining and is a presentation tool that is used to further explore results and findings generated from data mining into more granular detail (Berry, et., & al., 1997). In order to facilitate OLAP, data warehousing is used as a platform for data mining where multi-dimensional data of varying resolution is capable of being suitably extracted from different locations for further analyses (Han & Kamber, 2011).

## 2.6 Data warehousing

The consolidation of data from different sources into a central repository is known as data warehousing (W. Kim et al., 2001). Consequently, a data warehouse (DW) is an integrated and time-varying collection of data (Inman, 1996). It is intended to be used for the support of management decision making, through easy access and manipulation by analysts and decision makers (Sean, 1997). In addition, a data warehouse is an enterprise oriented, non-volatile collection of read-only data that is stored at several levels of detail (J. Miller & Han, 2009).

Although large amounts of data are contained within it, "*a data warehouse is not just a store of data*" (Inmon, 1996). This is because the major characteristic defining a data warehouse lies in the process of analytic enquiry that is optimized through normalized relationships of data elements and implementation (Abdullah & Ansari, 2005a; Abdullah, Brobst, Pervaiz, et al., 2004). The optimization overcomes inconsistencies in multi-formatted data and the analytical enquiry allows for archived operational data, that is periodically refreshed, to be extracted (Fernandez, 2003). In support of, and with these guidelines in mind, Ruiz & Becker et al (2005) proposed an architecture for the building and operation of a data

warehouse. According to them, a DW should comprise of an application integration layer, a data integration layer and a presentation layer (Ruiz, Becker, et, & al, 2005). Furthermore, if the data warehouse is a single and central organisation wide data warehouse, it is then referred to as a an Enterprise Data Warehouse (EDW) (Ossimitz, 2009).

Data extracted from data warehouses is increasingly being used by corporations to develop business advantage in dynamic business environments. This is done through monitoring activities using dashboards for business intelligence and competitive edge (Watson & Wixom, 2007). In April 1988, the National Agricultural Statistics Service (NASS) which administers the U.S. Department of Agriculture's statistics implemented their first data warehouse (Yost, 2000). Since then data warehouses have become commonplace in various industries including telecommunications, insurance, financial services, retail, healthcare and taxation (R Kimball, Ross, Thornthwaite, Mundy, & Becker, 2011). Many software products have facilitated the creation of data warehouses, for the purposes of analysis and mining of data (Kim et al., 2001).

### 2.6.1 Dimension modelling

In order to reduce data within a data warehouse to a simpler design that aids retrieval efficiency, preliminary data modelling is performed on the data (R. Kimball, 1997; R. Kimball & Ross, 2002). The dimensional modelling proposed for this study consists of entities such as region, climate, soil and vegetation. Essentially, dimensional modelling is equivalent to the process and regime of data selection prior to directed enquiry through the use of fact and dimension tables (Sen & Sinha, 2005). An established technique of dimensionality reduction and multivariate analysis used in unsupervised learning is Principal Component Analysis (PCA) (Labib & Vemuri, 2006).

### 2.6.2 Principle Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical technique used in many scientific disciplines where several inter-correlated and dependant variables are examined and used to extract a new set of variables (Abdi & Williams, 2010). PCA is a method of reducing the dimensionality of a dataset by establishing a set of variables that are pertinent in summarising the attributes of the data. It is classically used before clustering. These summary variables are known as the Principal Components (Ma, Chou, & Yen, 2000) which, although not correlated are ordered in terms of variance (Jolliffe, 1986).

Traditionally, there is an assumption that the first few PCs should capture most of the variation in the original dataset, while the last few PCs in the analysis may then be assumed to cover the outliers (Yeung & Ruzzo, 2001). However, this may not be the case as Yeng & Ruzzo (2001) have concluded. Although there are heuristics for choosing the number of first few PCs to retain, most of the selection is done ad-hoc and the rules are informal (Jolliffe, 1986). In the context of this research, the principle components were deemed to be the rainfall and temperature profiles extracted from the climate dataset. The other principal components were the soil type and land use characteristics of the study area.

## 2.7 Types of data

Data is characterised by different degrees of complexity and with specific application domains. For example, geographic spatial information is multi-dimensional data with many facets, including associations and measurements of the topography. The following paragraphs explain the complexities associated with GIS and measurements of complex trait data.

Data can be considered to have both characteristic and referential components. The characteristic component refers to observations and measurements, while the referential component specifies the context of

where the observations and measurements were made (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). The referrers can be moments of time, coordinates of a location, or sets, associations and combinations (Andrienko & Andrienko, 2006). For example, temperature is a one dimensional data item which only holds one unrelated feature. However, if the temperature data was associated with a location, the dimensionality of the data item would increase to 2 dimensions. Additionally, if it is associated with both a time and a location, its dimensionality extends to 3 dimensions.

Data may also have many attributes in the form of data structures spread over time (temporal) and space (spatial). Consequently, a data object could be either plain (neither temporal or spatial), temporal, spatial or spatio-temporal (Parent, Spaccapietra, & Zimány, 1999). The special nature of temporal and spatial data warrants an investigation into spatio-temporal data (Roddick, Hornsby, & Spiliopoulou, 2001). However, "*spatio-temporal features may be associated to objects, attributes or relationships*" (Parent et al., 1999). In this regard, the following associations are possible.

1. Spatial data is measurement data or data structures that can be associated with physical or geospatial positions.
2. Temporal data is measurement data or data structures that are described as events across particular points in time, whether that time interval is a linear continuum or a cycle. The retrieval, recording and processing of the data can be done at different temporal precisions such as seconds, hours, days, months, years etc. (Andrienko & Andrienko, 2006).
3. Data combining both spatial and temporal references may be used in analysis and processing in geo-spatial contexts.

Data that has elements of geo-spatial dimensions can generally be classed as geographic information systems.

### 2.7.1 GIS

There are a number of components to a Geographical Information System (GIS), the foremost of which is the actual geographic data and the next major component is the set of data-processing functionality (Tomlin, 1994). The functions include collecting, storing, retrieving, transforming and displaying spatial or geographically referenced data within purpose-built software such as the open source QuantumGIS or the proprietary ArcGIS software suite. Additionally, the objects within the specified Euclidean space interact, based on simple distance and proximity relationships (H. J. Miller & Wentz, 2003). Within a GIS, the spatial data is geographically referenced to the surface of the earth and is referred to as *geospatial* data (Worboys & Duckham, 2004). In addition, a GIS is composed of a static view of geospatial data known as spatiality, as well as how the geospatial phenomena evolve in time, referred to as temporality (Koperski, Han, & Adhikary, 1999).

GIS allows users to "*question, interpret and visualise data*" in a way that facilitates quick understanding and promotes sharing (Markovic, Stanimirov, & Stoimenov, 2009, p 1).

The GIS model is based on the conception of the geographic world as objects and fields, termed the object model (Goodchild, 1992). Within this classification the world is represented as a surface which can be populated by discrete and identifiable entities that each have specific representation and distinguishable properties (Couclelis, 1992). They may not be related to specific geographic phenomena such as climate and vegetation. In addition, man-made features such as roads, buildings and other improvements are denoted as objects.

An alternate view, known as the field model, mirrors the physical geographic features of the world as a set of spatial images distributed over the terrain, wherein terrestrial phenomena such as climate and vegetation are modelled as fields (Y. Liu & Goodchild, 2008). Consequently, both the object model and the field model have been adopted in providing a workable frame of reference for current GIS technology (Camara,

Monteiro, et., & al., 2000). The nature of the spatial phenomena involved, determines the type of data that is to be processed in a GIS (Leduc, Bocher, Fernando, & Moreau, 2009). It is therefore important to understand the different GIS data types.

- ## GIS data types

Spatial data comes in two basic forms for describing spatial features and attribute data. These forms are the raster or grid cell data and vector or polygon data (Bolstad, 2005; Congalton, 1997). Raster data is stored as numerically coded grid-cell or pixel data in the form of *n*-dimensional bit or pixel maps (Han, Kamber, 2006). Vector data is composed of a series of points, lines and polygons represented as unions or overlays as well as the partitions and networks that are formed by these components (Bolstad, 2005). In this way any geographical object can be represented digitally, where for example, objects such as traffic lights and buildings can be represented as points. The lines are used to denote things such as rivers, roads and pipelines whereas the polygons are used to represent any bounded object such as farms, parks and reserves (Gray, 2008; Ormsby, Napoleon, Burke, Groessl, & Feater, 2004). Vector or Euclidian space entities present a problem in terms of computation whereas raster spaces are more conducive to computability (Y. Li, Li, Chen, Li, & Lin, 2003). The computability of the raster space is due to the fact that invariability is not lost when vector data is converted to raster data. Consequently, both GIS raster and vector data types are employed in industry for different purposes. In this research, the soils, vegetation and land-use profiles were in vector format, while the elevation data was in raster format.

- ## GIS Usages

As a field of science, GIS deals with the location, description, explanation and prediction of patterns and processes of scaled geographical data together with principles, techniques, analysis and management of spatial

information (Longley, Goodchild, et., & al., 2005). GIS systems are used for measurement, mapping and analysis of geographical terrain and may be domain specific and location based, such as in criminology where it is referred to as Geo-ICT (van Schaalk & van der Kemp, 2009). GIS is used for decision making in town planning for mapping new road networks, devising future agricultural strategies and discovery of mineral resources for mining (Worboys & Duckham, 2004). More generally though, significant advances in spatial data acquisition, storage and retrieval have led to the establishment of Spatial Data Mining and Knowledge Discovery (SDMKD) (D. Li & Wang, 2008).

### 2.7.2 Complex data

In order to set about preparing data for subsequent analysis, a data warehousing approach is often used where data integration and dimension modelling is used for the creation of an appropriate analysis context (Boussaid, Tanasescu, Bentayeb, & Darmont, 2007). These analysis contexts are made up of models where data is presented using indicators and observation axes, termed measures and dimensions respectively (R. Kimball & Ross, 2002). However, as the emphasis of the created models in data warehouses is for supporting DSS, the models are not normalised. According to Darmont et al. 2005), data that emanates from various sources and is heterogeneous is known as complex data (Darmont, Boussaid, Ralaivao, & Aouiche, 2005). Accordingly, they have characterised complex data as being multi-format, multi-structure, multi-source, multi-modal and multi-version.

In the case study presented in this thesis, the data is easily recognisable as complex data according to this definition. This is because the raw data exists in the form of vector, raster and text formats (multi-format), they are differently structured in terms of shapefiles and comma delimited text (multi-structure) and arising from different source databases held at DAFWA (multi-source). In addition, the data was constituted of different profiles of soil, vegetation and climate (multi-modal) as well as changing in

terms of time and value (multi-version). The complexity of the respective datasets is represented in the Table 2-2.

**TABLE 2-2. THE COMPLEX CATEGORISATION OF THE DATASETS WITH REFERENCE TO DARMONT ET AL. 2005**

| RAW DATA | COMPLEXITY |
|---|---|
| Vector, raster, ASCII text data | Multi-format |
| Shapefiles, delimited text | Multi-structure |
| Arising from different databases and weather | Multi-source |
| Different profiles e.g. Soil, vegetation, | Multi-modal |
| Time-series data e.g. Range of months and | Multi-version |

In addition to the above mentioned characteristics of complex data, digital images are another form of complex data where the complexity is a combination of the characteristics listed in Table 2-2. The next section therefore deals with the nature and complexity of digital images.

### 2.7.3 Digital images

A digital image is made up of a number of points or picture cells (pixels) in a matrix (Guarneri, Vaccaro, & Gaurneri, 2008). There are many types of digital images such as Bit Maps (BMP), Tagged Information File Format (TIFF), Joint Photographic Experts Group (JPEG) and Portable Network Graphics (PNG). Consequently it can be seen that digital images are multi-format, multi-structure, multi-modal and essentially multi-source as well, due to their origin. Digital images may also be analysed both manually and through some form of image processing. The activity of manual analysis through visual exploration is known as visual data mining (Keim, Mansmann, Schneidewind, Thomas, & Zeigler, 2008).

## 2.8 Visual data mining

Visual Data Mining is a process involving human intuitive interaction with data (Simoff, Bohlen, & Mazeika, 2008; Y. Vagh & Xiao, 2012) where it has been described as *overview first*, *zoom and filter*, followed by *details on demand,* otherwise known as Schneiderman's (1996) information seeking mantra. This approach to DM combines the use of DM algorithms with techniques of information visualisation (Carter, 2000) due to the visual nature of the data. Visualisation in turn, depends on scientific, geospatial and information analytical methods (Vaus, 2001).

The visual nature of data can be quite complex due to size and heterogeneity. This visual complexity is in addition to the structural and the dimensional complexity, which then significantly adds to the computational complexity (E. J. Wegman, 2003). Wegman (2001) suggests that there are critical way points in the size of data, usually in the order of between $10^6$ and $10^7$ bytes where the modes of analysis separate. Furthermore, he adds that the distinction is marked by a failure of complex algorithms or unreasonably long Ethernet data transfer rates. At these points the concerns shift from statistical optimality to computational optimality as the computational complexity is in the order of $0(n^2)$ or higher (Papadimitriou, 2003). A task oriented view of the visual analysis process is given in Figure 2-6. Each stage of the mappings and transformations require some human interaction or specific task to be undertaken by the human analyst (Garcia, DeOliveira, Maldonado, & Mendonc, 2004).

Images may be analysed by either visual inspection or automated analysis through various methods of image processing (Yarrow, Perlman, Westwood, & Mitchison, 2004). However, according to Yarrow et al (2004) images examined through the process of visual inspection can be "*limited by subjectivity, operator fatigue, and the lack of quantifiable metrics*", especially if the number of images that are being examined are numerous. In these instances it is better to develop an automated method of analysis.

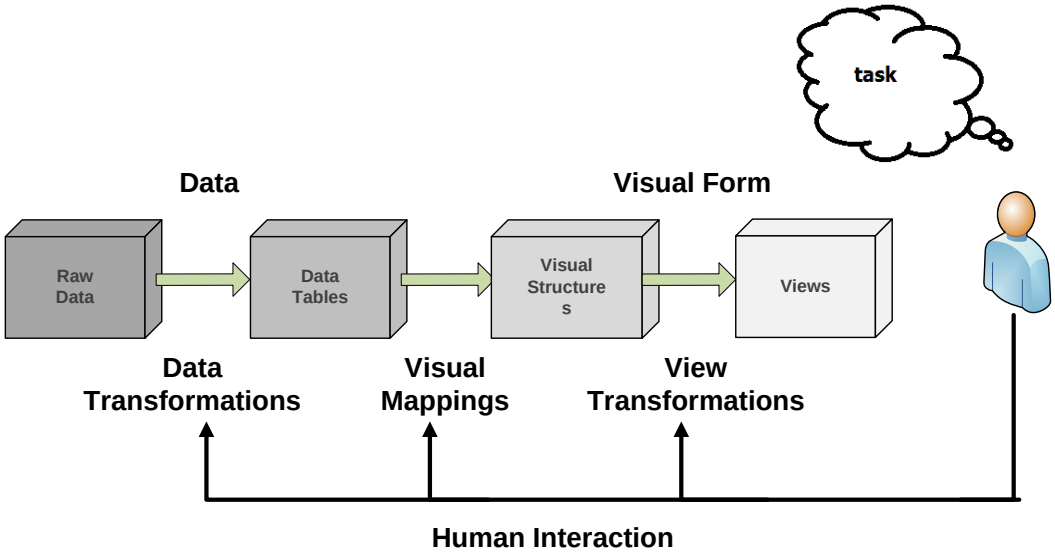



**Figure 2-6. The Visual Analysis Process redrawn from (S. Kim et al., 2006)**

However, for the purposes of the case study in this thesis, manual visual inspection has proved quite sufficient. This is because it was used primarily as a primer, and for exploration prior to subsequent confirmatory data mining analyses. As the research data involved an element of geographical location it was necessary to explore aspects of spatial data mining and associated knowledge discovery.

## 2.9 Spatial data mining

Spatial Data Mining and Knowledge Discovery (SDMKD) is about uncovering the implicit relationship and characteristics that may exist in large spatial databases (Ng & Han, 2002). The process involves discovering interesting and previously unknown patterns that may be potentially useful (Shekhar, Zhang, Huang, & Vatsavai, 2003). The basic methods of analysis used in SDMKD are classification, association rules, characteristic rules, discriminant rules, clustering and trend detection (Kuba, 2001) as well as serial rules, predictive rules and exceptions (D. Li & Wang, 2005). In addition mining, extraction and prediction of data may

be achieved through different techniques such as probability theory, spatial statistics, evidence theory, fuzzy sets, rough theory, neural networks, algorithms, decision trees, exploratory learning, spatial inductive learning, visualisation, spatial online analytical mining (SOLAM) and outlier detection (D. Li & Wang, 2005).

A graphical depiction of the techniques is given in Figure 2-7 where the left hand side of the diagram represents the specialised technique and the right hand side represents how the technique is implemented. As an example, the technique of visualisation is based on evidence theory. In conjunction with this, the viewpoints and methods of knowledge acquisition as modified from tables presented by Li & Wang (2005) is given in Figure 2-8. The diagram is constructed in a similar fashion to Figure 2-7 where the left hand side represents the method of knowledge acquisition and the right hand side of the figure represents the viewpoints upon which the methods are based. As an example, the association method is based on a logical viewpoint as annotated in Figure 2-8.

SDMKD is not only used for intelligent analysis of GIS data through uncovering both spatial and non-spatial patterns as well as general characteristics, but for knowledge acquisition of remote sensing image classification (Li , Di, & Li 2000). Spatial data mining is considered to have a number of facets and is comprised of spatial statistics, spatial analysis, GIS, GPS, machine learning, image analysis data warehousing and data mining (Shekhar et al., 2003).

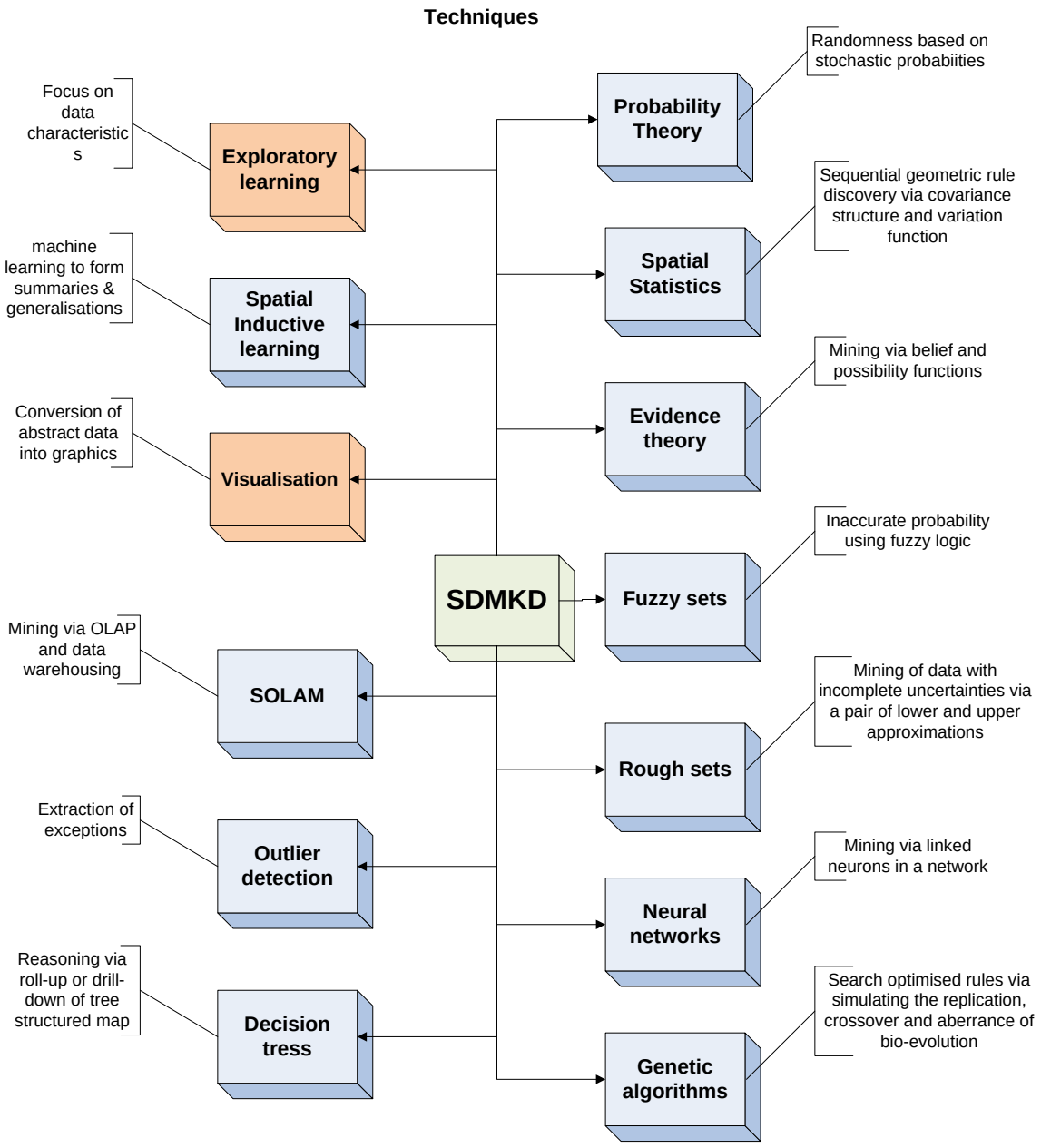



**Figure 2-7. Techniques used in SDMKD modified from descriptions and tables (D. Li & Wang, 2005)**

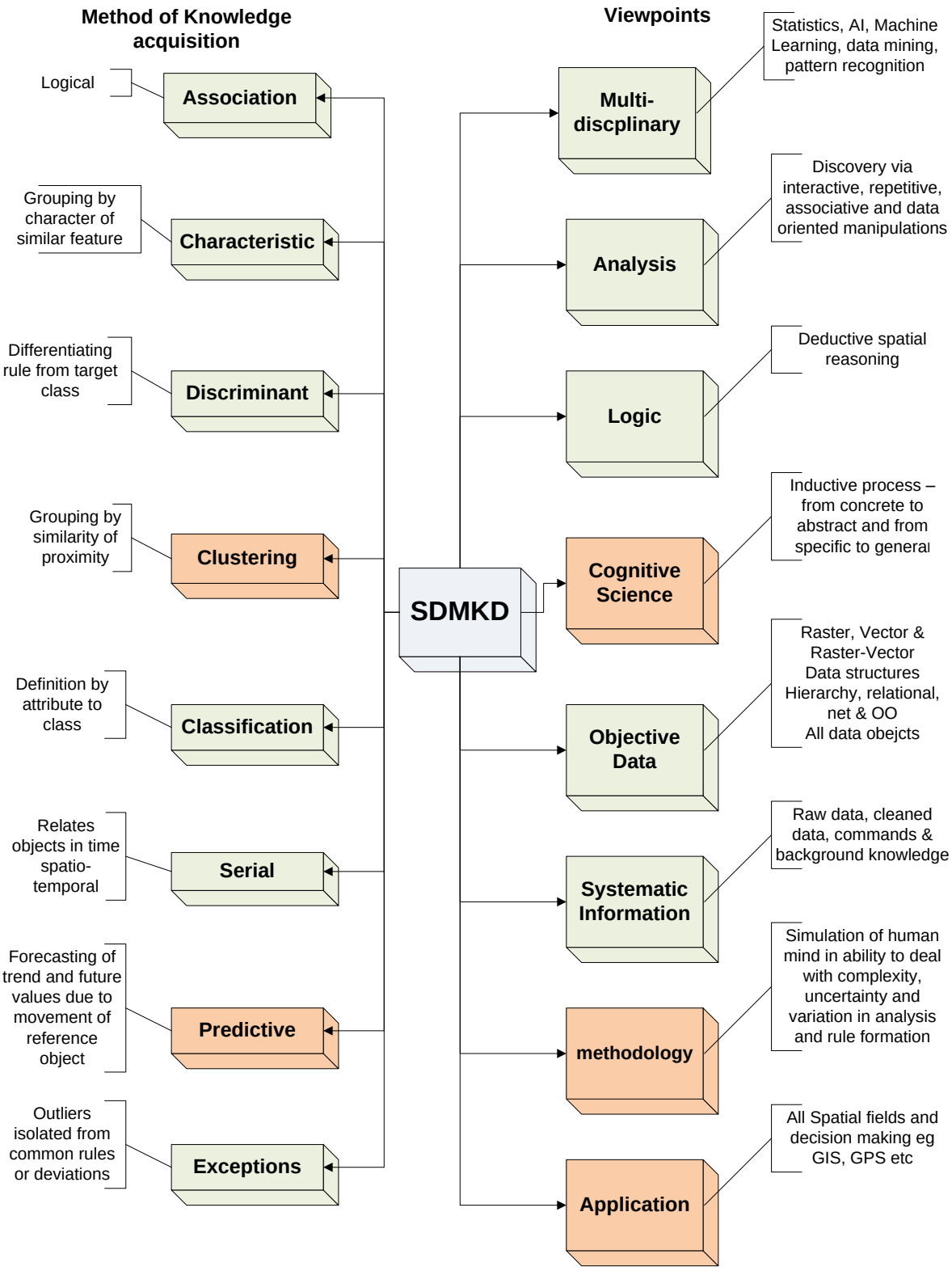



**Figure 2-8. Spatial data mining and knowledge discovery viewpoints and techniques modified from descriptions and tables (D. Li & Wang, 2008)**

The spatial data is represented either in raster (pixel) or vector (point) graphics. In the vector model of representation, there are many data structures such as the Quad tree, k-d-tree, R-tree, R*-tree, R+-tree, Vp-tree, M-way Vp-tree, Multi-vantage-point tree and M-tree (Kuba, 2001). The monitoring of the individual activity patterns in general, and land-use in particular, have been made possible due to advancements in global positioning system (GPS) technology as well as the appearance of mobile GIS devices (C. Li & Maguire, 2003). However, SDMKD is considered to be more complex than the discovery process for relational data. This is due to the fact the mining algorithms have to consider the relationships involving location (Keim, Panse, Sips, & North, 2004) and the effect of neighbouring objects (J. Miller & Han, 2009). Ordinary statistical methods are not suitable due to the interdependence of objects. Nevertheless, mining granularity comes in two flavours which are spatial object granularity and pixel granularity (Li et al., 2000). Consequently, the display of dense point sets on maps have been explored using combinations of clustering and visualisation (Keim et al., 2004).

A new approach of combining inductive learning using the C5.0 algorithm together with image classification through Bayes was used to improve classification of land-use in China by Li & Di et.al, (2000). Other previous approaches in coping with dense geographic data is 2.5D visualisation where data points are aggregated up to map regions (Keim et al., 2004). Visual Insight's In3D and ESRI's ArcView are examples of this technique in commercial systems. However, important information may be lost because of aggregation and only the coarsest patterns are observable (Keim et al., 2004). More detail is possible when the individual data points are visualised as bars on a map. Applications such as SGI's MineSet and AT&T's Swift3D embody this technique. The problem with this approach is that too many data points plotted in the same location causes occlusion where all the data is not seen simultaneously. A method devised to prevent overlap in a two-dimensional display without aggregation is to reposition the pixels locally as in Gridfit (Keim & Hermann, 1998). This approach does

have its own drawbacks, in terms of the placement of the points being quasi-random depending on the ordering of points in the database.

New approaches are thus evidently formed from using a combination of methods, techniques and viewpoints that closely match the datasets under analysis. Nevertheless, all data that have multiple facets, especially if they are location specific, have an inherent diversity that necessarily categorises them into what is termed complex data. In the context of this research, the datasets were location specific but the diversity was inherent in the separate profiles of soil, land use and climate variability. Inevitably, it becomes necessary to employ consistent algorithms in the effective processing and mining of complex data and this is usually achieved through the means of proprietary or open-source data mining software tools.

## 2.10 Data mining software tools

Data is a representation of the results of the observation or measurement of phenomena (Andrienko & Andrienko, 2006). Through analysis, the phenomena may be studied to answer questions about the phenomena. As the data, represented as measurements or descriptions of phenomena, are arranged in various structures in relation to the subject, tools are required for its interrogation and investigation. A framework is a formal method that describes the processes that need to be performed on the data in order to achieve a desired outcome. The whole process involves using a set of prescribed tools specific to a designated task to attain a solution for the defined problem.

As the significance of data mining emerged as a valid data analysis approach, many tools were developed to address the general, as well as the specific issues within the field. For example, AgroAdvisor is tailored specifically for the agricultural field, as the name suggests. Although WEKA (Witten & Frank, 2005) can be configured to be utilized with any domain, it has primarily been used in agriculture. Commercial tools like

MineSet, offer additional analysis tools of searching, sorting, filtering and drilling down, as well as 3D graphics capability. Pryke (1998) explored a number of tools like Clementine, DBLEARN, EFD, Explora, INLEN, Knowledge Discovery Workbench and RX and provided a useful analysis of the strengths and weaknesses of each. In addition, Goebel and Gruenwald (1999) provided a comprehensive survey of software tools that were prevalent in the 1990s. Many of the earlier model tools have been superseded by more inclusive models such as WEKA which have the added advantage of being open-source (Hornik, Buchta, & Zeileis, 2009). There are also more statistical software packages such as SAS, SPSS, S-plus, R and Revolution R in current use. Some are analysis packages with statistical libraries such as MATLAB and Mathematic, and others are more general software languages with statistical libraries such as JAVA, C++ and PERL (E. Wegman & Solka, 2005).

These software tools are described for their permutational possibility akin to building blocks, for use in a component DM framework. In the course of the data exploration and analyses undertaken for this research, a number of software packages with some relevance and application to GIS were explored and subsequently utilised. These were the specialised GIS and statistical software packages of QuantumGIS, GRASS and the Revolution R statistical package.

### 2.10.1 R Statistical Package

The Revolution R statistical package is an Open Source software package with a large statistical capability organised into several packages. Each of the packages contain a set of different statistical functions or algorithms (E. Wegman & Solka, 2005). Some of the statistical techniques available within the R package include both linear and non-linear modelling, time series analysis, classification, cluster analysis, prediction, as well as re-sampling and survival analysis (DuDoit, Gentleman, et., & al., 2003). The R package also contains graphical techniques for visualisation and has a high degree of extensibility through packages (Gentleman & Ihaka, 2010).

In addition to the eight packages supplied with R, the CRAN family of internet sites offer a wide range of modern statistical techniques for use with R (Hunter & Cheung, 2005). Although using R can be difficult for users not familiar with Command Line Interface (CLI), it does have mechanisms such as the intersystem interfaces for interacting with software from other languages. Furthermore, R has been recently upgraded to feature a GUI interface (Revolution R). Consequently, the R software package has been used extensively during the course of this research, especially in regards to processing and interpolation of climate data. The algorithms for the interpolation of the rainfall and temperature profiles, the creation of the study area grid and the projection of the resultant interpolated data were all developed in the R software environment.

### 2.10.2 QuantumGIS

QuantumGIS is an Open Source GIS that is free to download and use. It is a feature rich environment complete with applications and third-party plug-ins. These tools offer various facilities for the creation, editing, mapping and conversion of a variety of raster and vector file formats that include ESRI shapefiles, ASCII grids, text and images (Stefanakis & Prastacos, 2008). The user friendly software package runs on a range of operating systems including Linux, Unix, Windows and Mac OSX (Boulos & Honda, 2006).

QuantumGIS has been used in this research for the purposes of loading the point data and for the subsequent visualisation of the weather station points on a surface map. It has also been used for the initial test interpolations of the point data onto the study area grid surface in order to generate 1000 metre pixel size data for rainfall and temperature readings. Although, the interpolations were subsequently superseded by interpolations done in the Revolution R software package, the interpolations done in QuantumGIS were an important part of staging the research visually and for preliminary analytical exploration.

### 2.10.3 GRASS

GRASS is an acronym for Geographical Resources Analysis Support System and is essentially a public domain geographic information system (D. C. Miller & Salkind, 2001). It is used for the data management analysis and visualisation of GIS related data and is available for download and free usage under the terms of the GNU General Public License (GPL). The GRASS software package is used essentially for the processing of raster, vector and point data and additionally contains image processing modules (deHaro-Garcia et al., 2009).

GRASS has been used in this thesis for the visualisation of the geographic data profiles of land-use and for the production of the PNG images. The PNG images constructed in the GRASS software package enabled the visualisation process especially in regards to the overlays of separate profiles of elevation, soil composition, land-use and vegetation. Its variable setting feature of the overlays were particular useful in discerning underlying detail of composite images.

### 2.10.4 WEKA

The WEKA software and machine learning toolkit is an excellent general-purpose environment for applying techniques of classification, regression, clustering and feature selection in bioinformatics research. Its range of algorithms and data pre-processing methods makes it an invaluable aid to data mining (Witten & Frank, 2005). The in-built data mining algorithms have the capability for categorical and numeric machine learning. Another of the tools in its arsenal is the meta-classifiers to enhance the performance of classification algorithms such as boosting and bagging. The data pre-processing routines allow the raw data to be manipulated and transformed into suitable form for input (Cunningham & Holmes, 2001).

WEKA's feature selection tools permits irrelevant attributes to be identified and excluded. In addition, it comes equipped with experimental support for "*verifying the comparative robustness of multiple induction models (for*

*example, routines measuring classification accuracy, entropy, root-squared mean error, cost-sensitive classification, etc.)"* (Cunningham & Holmes, 2001, p 1). Lastly, it has benchmarking tools useful for comparisons of relative performance of a variety of learning algorithms on a number of datasets (Cunningham & Holmes, 2001).

Weka is equipped with graphical user interfaces for data exploration and experimental comparison of different machine learning techniques. The main aim of WEKA is to assist users to extract useful information and for the identification of a suitable algorithm from which a predictive model could be generated (Frank, Hall, Trigg, Holmes, & Witten, 2004).

- **WEKA usage**

The only input requirement in WEKA is that the data be in a single relational table in Attribute Relation File Format (ARFF). However data from a spreadsheet is easily convertible to the ARFF format. Most of the data in an ARFF file "*consists of instances with attributes for each of the instances separated by commas*". This format is known as Comma-Separated Value (CSV) format. Most spreadsheet and database programs allow conversion to this format for export. The conversion to ARFF format from CSV format can be done manually or automatically through WEKA (Witten & Frank, 2005).

For a manual conversion, the CSV format must be edited with a text-editor or word processor to add three types of tags. The name of the dataset is added using an @relation tag. Attribute information is then added by using the @attribute tag which shows all the possible values in curly brackets. Lastly, all the actual CSV data is preceded by the @data tag (Witten & Frank, 2005). WEKA allows input to be from files, from a source URL on the web, or a database using JDBC objects and SQL queries to select the records from the database. Furthermore, WEKA can convert data imported in C4.5 names format or in binary serialized instances format automatically to ARFF format (Witten & Frank, 2005).

- **WEKA filters and features**

The pre-process tab in WEKA allows data to be filtered via supervised and unsupervised methods. Both these methods allow attributes or instances to be selected. Many permutations are selectable under either the attributes or the instances. Attributes can also be simply removed from the process to eliminate their effects (Witten & Frank, 2005).

WEKA has other features that are selectable from the main tabs on the interface. They are the classify, cluster, associate, select attributes and visualize tabs. All of these together incorporate the different techniques used in data mining. The user simply has to test which method and technique produces the desired results. Most of these require some knowledge of what the techniques accomplish and their relevant strengths and weaknesses (Witten & Frank, 2005).

The combination of visual and data analyses specific to geographic location in an agricultural context inevitably gives rise to the idea of refinement in the data analysis to what may be more appropriately termed precision agriculture.

## 2.11 Precision agriculture

Precision agriculture is a general term that has arisen due to the use of advanced GPS and sensor technology being utilised in the agricultural sector for the purposes of crop and soil management and optimisation of crop yields (Ruß, Kruse, Schneider, & Wagner, 2008). Specifically PA or precision farming is concerned with "*the sampling, mapping, analysis and crop management*" with reference to the spatial and temporal variability of the agricultural growing regions (Ruß, Kruse, Schneider, & Wagner, 2009).

One form of precision agriculture is site-specific crop management in relation to the time and space variability. In order to scientifically test and validate the concept of site specific crop management, the proposal and testing of the null hypothesis is necessary (Gray et al., 1997). The null

hypothesis in relation to precision agriculture assumes large temporal variability in crop yield relative to the scale of a single field and asserts that the optimum risk avoidance is uniform management. Consequently, Whelan and McBratney (2001) argue that precision agriculture is validated when the null hypothesis is refuted. In their studies, they examined the degree and cause of the variations together with the suitability for management intervention.

Precision agriculture in terms of sustainability, is about managing the inputs and treatments applied to a crop according to its location, in order to manage the crop yield in an environmentally friendly way (von Braun, 2007). Sustainability has been defined with a consideration for the economic, environmental and sociological impacts taken holistically (Gulati, Joshi, & Cummings, 2007). Corwin (2005, p 13) describes agricultural sustainability as based on a "*delicate balance of maximising crop productivity and maintaining economic stability while minimising the utilisation of finite resources and detrimental environmental impacts*".

It is evident from these definitions that precision agriculture is not only about crop production and economic management, but about risk management with reference to climate and the environment in the long term. In addition, although SSCM strategies, which is another name for PA, may be used to optimise crop yield with minimal environmental impact, it is largely dependent on the variability and the accuracy of the analyses (Piatetsky-Shapiro, Brachman, Khabaza, Kloesgen, & Simoudis, 1996).

Furthermore, farmers have difficulty in correlating their yield maps to a management strategy (Wang, 2003). Figure 2-9 depicts the process of SSCM in terms of matching the variables in order to produce the desired outcomes of increased crop production with minimal environmental impact (Casa & Castrignanò, 2008).

## 2.12 Case studies similar to the current work

A number of recent studies were found to have a similar theme. These included the research of Castrignano (2010), Dunstan (2009) and deOliveira (2009). Castrignano's research sought to investigate the "*different approaches to delineate agricultural management zones*" (Castrignanò, 2010). This was done by collecting raw agricultural data, performing factorial krigeing to decompose them into regionalised factors for estimation, performing principal component analysis and then studying the variation in attribute and geographic space. The objective was to study the spatial variability of soil and determine the response of durum wheat to spatial and temporal variability. The method used by Castrignano (2010) is depicted in Figure 2-9. Factorial krigeing was carried out on the raw data and these were decomposed into the regionalised variable data. On the other hand, the environment data was used in its original form. Together, these variables represented attributes with continuous variations for the regionalised data, whereas the environment and geographic variables displayed gradual variation. These contrasting variables were then setup in a matrix for covariant function analyses.

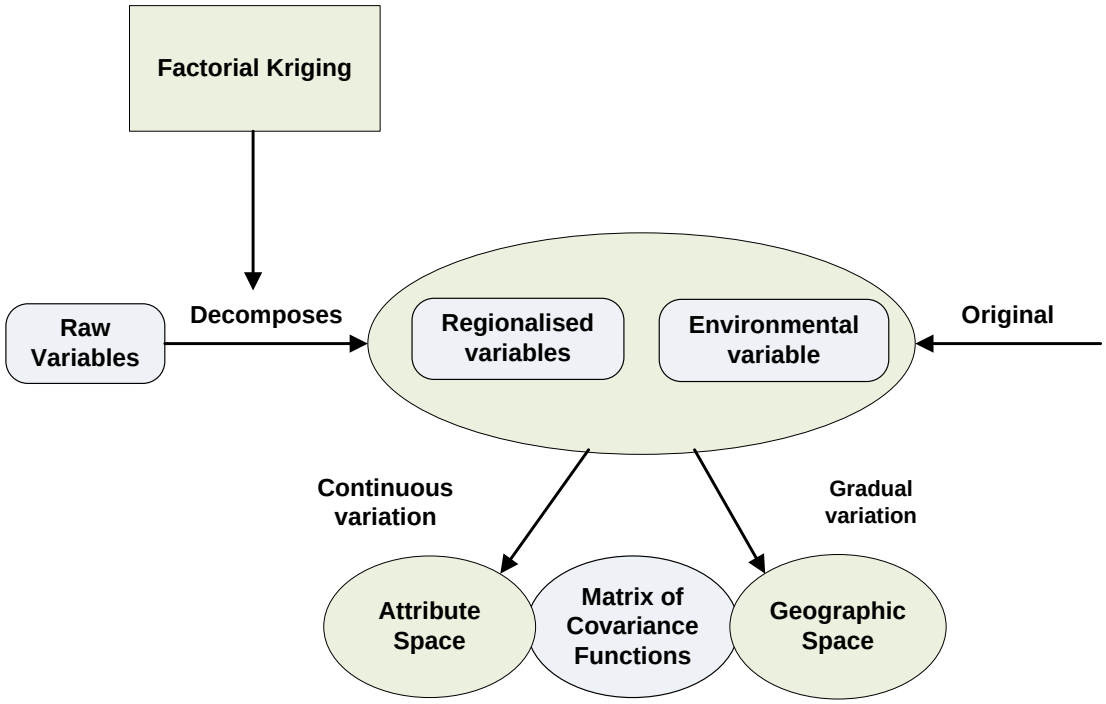



**Figure 2-9. Process to study variation in attribute and geographic space (Castrignanò, 2010)**

Dunstan (2009) in his study of "*the hierarchies of sustainability in a catchment*" formulated LUCC models to predict the effect of land-use. He extracted a dataset of topography and land-use measurements from which subsets were selected. The selected data was then processed for the purpose of making predictions about the effects of runoff in a catchment area. The data was also extrapolated into nominated time intervals using an R program as well as a customised Perl/C++ program designed to specifically process the subset. The output was saved into a data-cube for summary and aggregate interrogation and reporting. Dunstan's (2009) data-cube was generated "*on-the-fly*" by calculating aggregation measures across the selected cuboid cell - e.g. average depth of the water table for all cropping regions above 300 metres. In addition, a visualisation tool was used to display the output. This process is depicted in Figure 2-10.

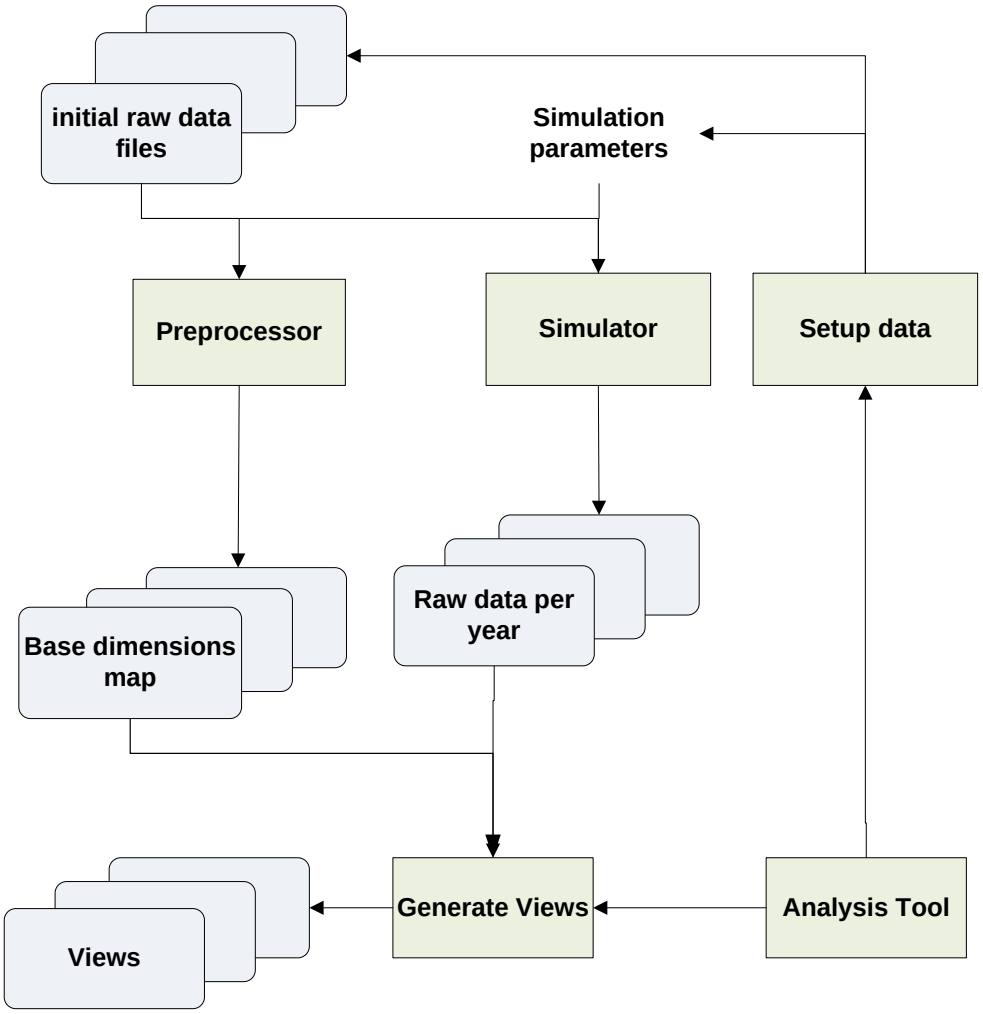



**Figure 2-10. System view of LUCC methodology (Dunstan, 2009)**

The research of deOliveira (2009) was in the use of decision support systems to address issues in precision agriculture using site specific crop management data and technology in an Australian context. The researcher offered a decision model in support of SSCM adoption especially in regards to the delineation of management zones.

In the past, other similar research studies have been undertaken in the usage of rainfall to predict crop yields through the use of empirical models (Parthasarathy, Kumar, & Munot, 1992). The focus of each of these studies was different in the emphasis of the importance of the climate relationships to crop yield (Mearns, Rosenweig, & R. Goldberg, 1997; Semanov & Porter, 1995), and the relevance of climate variables such as rainfall and temperature on the phenological growth stages crops (Wheeler, Craufurd,

Ellis, Porter, & Prasad, 2000). The present study was situated in conformity with and in resonance with this type of previous research.

## 2.13 Critique and findings from the literature review

A review of the literature has revealed that the progression of analytical studies from statistical analysis to DM analysis has been gradual albeit ongoing. In addition, there have been limited studies in agriculture where OLAP has supplemented or complemented DM. This situation was more pronounced in Australia. The research studies have also been limited to the areas of yield management and crop quality assessment. Little research has been done in the areas of precision agriculture and site specific crop management in terms of crop yield predictions, although deOliveira (2009) made some contributions in the area of DSS in PA for the management of SSCM zones. Studies of the effects of soil variation and climate change on land use have also been limited. Furthermore, the agricultural data available was very diverse, complex and from different sources. For example, agricultural data typically consists of plant and soil characteristics, plant and soil treatments, climate and weather data as well as topographical data in different formats.

Even though the research of Castrignano (2010) followed the pattern of the simultaneous investigation of continuous and gradual variation of climate and geographic variables, there were limitations to the complexity and analysis of the variance. The method used was ambiguous and black-box like especially with regards to the matrix of covariance functions. Therefore, it was not conducive to abstraction as was evident from the diagram in Figure 2-10. A more lucid and systematic approach was needed for the admixture of continuous and gradual variability of the component variables.

As such, there appeared to be no research done where data was gathered from various sources and systems and stored in a common database. This

meant that there was a need for the creation of a data warehouse where the data may be stored, and from which the selected data could be confidently and regularly interrogated. In addition, there was evidence that such complex data may require multiple strategies such as statistics, DM and OLAP as well as an algorithmic framework in order to competently and consistently interrogate and analyse the data.

Although the research of Dunstan (2009) employed the use of a data cube for the purposes of OLAP, this process was not effectively communicated or expounded. In terms of a systematic process of analyses, the review of the literature indicated that although other agricultural frameworks did exist, there was a need for a framework that catered for GIS and topographical data as well as an evaluation metric together with a way of comparing other existing frameworks. This research is therefore focused on covering some of these shortfalls in previous studies in a similar research space and it holds the promise of contributing to knowledge and best practices in this field from both an agricultural and scientific perspective. A list of the related research in this field is presented in Table 2-3.

**TABLE 2-3. LIST OF RELATED RESEARCH**

| Author | Year | Subject | Purpose | Advantages | Disadvantages |
|---|---|---|---|---|---|
| Gray et al | 1997 | Precison agriculture | Testing of null hypothesis | Field management | Large variability |
| Whelan and McBratney | 2001 | Precision agriculture | Refutation of null hypothesis | Investigated causes of variability | Single field |
| Corwin | 2005 | Sustainability | Definition of Sustainability | Investigated crop yield maximisation | Relationship Complexity |
| Von Braun | 2007 | Precision agriculture | Input management of treatments | Input/output effect | Specificity |
| Gulati, Josh, & Cummings | 2007 | Sustainability | Economic, environmental and sociological impacts | Holistic approach | Measurement of impacts |
| Ruß, Kruse, Schneider, & Wagner | 2008 | Use of sensor technology | Crop yield optimisation | Use of technology | Parameters not applicable to Present study |
| Dunstan | 2009 | Prediction of land use | Use of models for the prediction | Sustainability of agricultural land | Catchment area modelling |
| deOliveira | 2009 | Decision Support | Site Specific management | Australian context | Application |
| Castrignano | 2010 | Agricultural management | Investigation of different approaches | Investigation of continuo and gradual variation parameters | Limited scope |

# Chapter 3

# RESEARCH METHODOLOGY AND DESIGN

This chapter commences with an overview of the research methodology and covers relevant topics such as research paradigms, methods (empirical research and case studies), the Exploratory Data Mining (EDM) approach, the Visual Data Mining (VDM) approach and the research activities and analyses undertaken. This research was a process of iterative refinement, where certain avenues of investigation and tools were accepted, rejected or modified based on their suitability for effective research design and analysis outcomes. The main focus was to develop algorithms, experimental procedures and models to enhance understanding, visualisation and interpretation of the results.

## 3.1 Overview

The blueprint of this research design was based on which questions to ask, identifying relevant data, data collection and analysis of results to address the questions posed. The research was more deterministic rather than probabilistic in nature. It also attached more to theory building, in terms of experimental prediction not based on any hypothesis, than to theory testing a prior hypothesis (Vaus, 2001). The methodology employed involved a set of experimental research activities centred on a single multi-faceted complex case study. The aim was to answer the main research question of *"In what ways might the development and application of a system of analyses that incorporates DM, improve the interrogation of agricultural land use datasets for the purpose of crop yield predictions?*

## 3.2 Research methods

“Research is an inquiry process consisting of several specific components that include reflective inquiry, procedures, data (gathering, processing, analysis), issues of reliability and validity of study, and presentation of research findings” (Hernon, 1991, p 4). The reflective inquiry may be defined as the problem statement, literature review, theoretical framework, logical structure, objectives, research questions and hypothesis. The procedures are described as the design and methods of data collection (Hernon, 2009). This research is further defined as primary research in that it is original data analysis, as opposed to secondary research which is only the summary, collation and/or synthesis of other similar research (Luo, 2012). However, certain sections such as the literature review and the research methods are secondary research components. The process model, adapted from Oates (2007) and Hernon (1991), illustrates the research progression and is shown in Figure 3-1.

Part of the process of reflective inquiry, as denoted in Figure 3-1, involves building a skeletal framework. This is done for focus and sensitisation of the research material through logical structure and objectives, as concepts are not yet fully formed and a researcher’s knowledge is still limited (Morse et al., 2002). The literature review is then used to conduct the concept analysis, construct a scaffold of previous knowledge and establish boundaries so that the focus and scope guide the data collection and allow the analysis to proceed (Morse & Mitcham, 2002).

Hernon (1991, p 4) lists the three aims of research to include “theory building, theory testing and problem solving for decision making”. Hoadley (2004, p 203) defines empirical research as ”trying to predict and model the real world”. This research meets the problem solving criteria of these research aims, in that it seeks to model and predict phenomena in a real world situation.

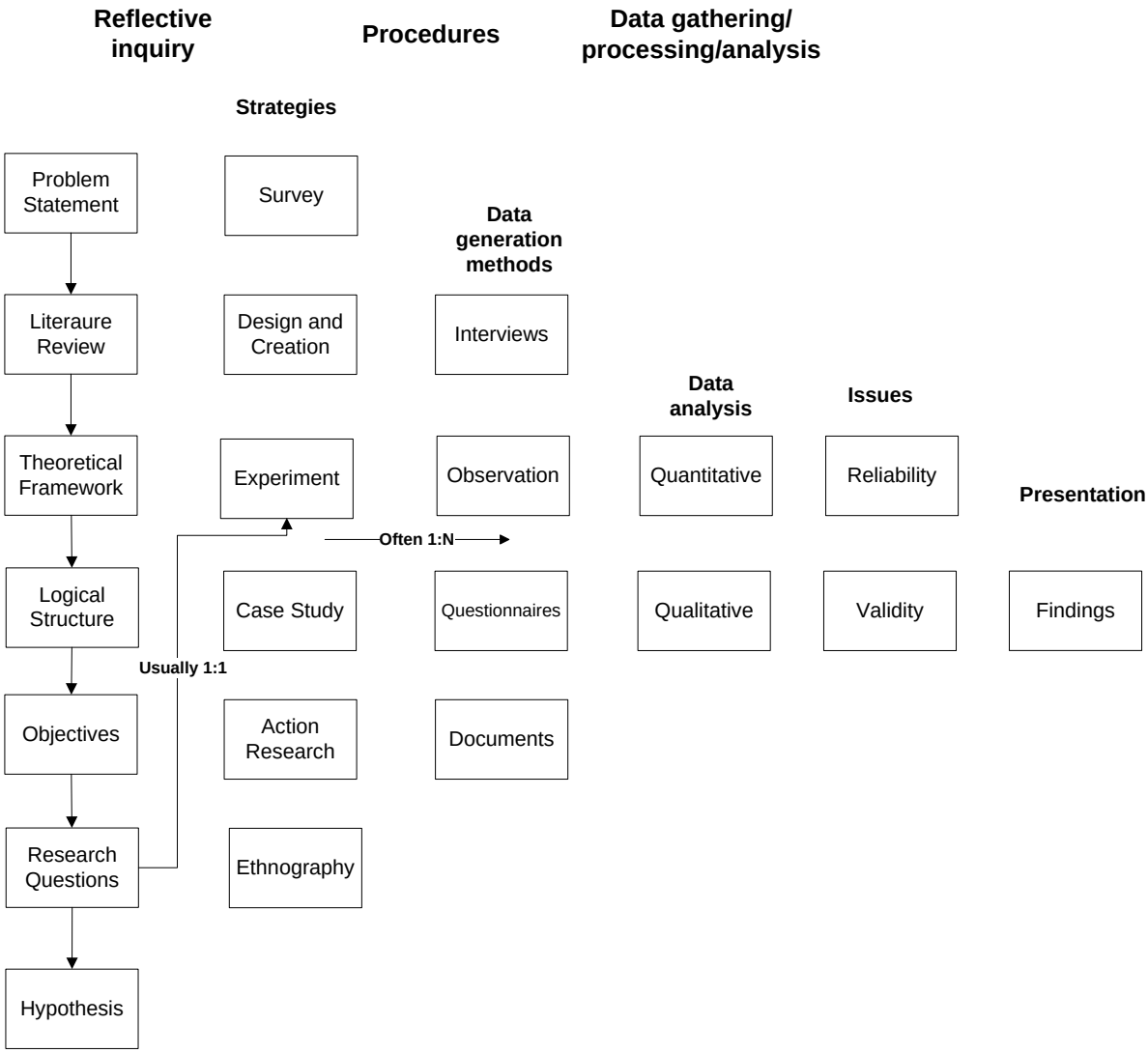



**Figure 3-1. Adapted process model showing research variety and process (Hernon, 1991; Oates, 2007)**

Qualitative research is research conducted in a natural setting based on participants' explanations (Denzin & Lincoln, 1994), while quantitative research is concerned with quantifying a relationship or deals with group comparisons (Creswell, 1994). This research may be considered a mix of qualitative and quantitative methods as it not only examines natural phenomena such as rainfall and temperature at a specific location but uses historical and numerical data in the analysis.

Issues of validity are important in any type of research whether the method used is quantitative, qualitative or mixed. There are several types of

quantitative validity including statistical validity, construct validity, internal validity and external validity (Shadish, Cook, & Campbell, 2002). Lincoln & Guba (1985) have adapted these labels for qualitative methods into credibility (internal), transferability (external), dependability (reliability) and confirmability (objectivity). Chris & Coryn (2007) define trustworthiness as the extent to which research results are believable by the participant and transferability as the degree to which it can generalised (Chris & Coryn, 2007). Dependability deals with the criterion of the variability of the context and confirmability is the degree to which results can be confirmed by others (Trochim, 2006). Issues of validity in mixed method research are more fully explored by Onwuegbuzie & Johnson (2006).

Bouma & Ling (2004, p 1) state that, “the research process must consist of linked activities whereby the research questions are connected to both the aims and results of the research through several intermediary phases”. The research questions facilitate the research strategy and the methods of data collection and analysis. In this research, the activities are interconnected and sequentially progress the experiments in answering the relevant research questions.

Having established the aims and the type of research, as well as research issues, the inevitable question is about rigour which may be defined as how data is gathered and interpreted in a way that quells the subjectivity filters of “our bias, our politics and our worldview” (Hoadley, 2004, p 203). In other words, rigour is the extent to which the subjective influence of the researcher is minimised. One way to satisfy the requirement for rigour is to use replicable and detailed experiments. Introducing stochasticity through interpolation (ordinary krigeing) can create bias (Abe & Smith III, 2004), however, this has been compensated for by using more sophisticated statistical techniques such as DM algorithms (Han & Kamber, 2011). These efforts of statistical and DM technique inference are means to augment the rigour of the research within the experimental paradigm.

## 3.3 Research paradigms

The way knowledge is studied and interpreted is influenced by the choice of research paradigm as it is where the intent, motivations and expectations of the research are determined (Mackenzie & Knipe, 2006). There are various research options to guide a disciplined inquiry. A paradigm may be described in terms of epistemology, ontology and methodology (Neuman, 2005). Put simply, epistemology is the theory of knowledge and how we get to know and it is interrelated to ontology and methodology (Audi, 2011). Specifically, ontology deals with the philosophy of reality, epistemology is concerned with how that reality is known and methodology refers to the specific practices used to attain knowledge of that reality (Krauss, 2005). Research paradigms are therefore different philosophical views and interrelated assumptions of reality, how it comes to be known and the methods used to acquire that knowledge (Oates, 2007).

Study and interpretation of knowledge is determined by these basic belief systems (Mackenzie & Knipe, 2006) and can be further subdivided by questions based on ontology (what is reality), epistemology (relationship between the knower and the known) and methodology (the methods used to know reality) (Brand, 2010; Lincoln, Lynham, & Guba, 2011). Positivist (value-free), interpretivist (value-laden), transformative (value-changing), pragmatic (value-relevance) and critical Research (value-cognisant) are some examples of different paradigms (Krauss, 2005). Due to non relevance to this research and their predominant entrenchment in the social sciences, the transformative, pragmatic, interpretivist and critical paradigms are not elaborated upon here. Instead only positivism and empirical research are briefly outlined.

The positivist paradigm stems from a realist ontology in that reality exists and is governed by natural laws and that science seeks to discover the truth and workings of that reality objectively (Guba, 1990). It is a conventional inquiry that has an objectivist epistemology through absence of influence on the outcomes. Its methodology is typically experimental

(manipulative) in that experiments are conducted to refute or confirm an established hypothesis under controlled conditions. Positivists are considered reductionists in that the studied phenomena are broken down to construct objective generalisations used to predict meaningful changes in circumstances (Brunner, 2006). Positivism is mostly associated with the quantitative method.

The empirical study is a scientifically defined and planned method wherein the research questions are postulated, followed by a process of data collection, analysis and presentation of the results (Perry, Sim, & Easterbrook, 2004). Empirical studies may be conducted using a variety of method classes. Easterbrook, Singer, Storey & Damien (2008) list five explicit classes of empirical research methods relevant to software engineering as:

- Controlled Experiments (including Quasi-experiments)
- Case Studies (both exploratory and confirmatory)
- Survey Research
- Ethnographies
- Action Research

These classes are described by Saunders, Lewis & Thornhill (2003) as strategies in their research onion diagram.

Although this study features several aspects of positivism, especially with regards to the quantitative nature, experiments and the framework, it involves a controlled experiment using a single but complex case study. Thus, it may be regarded as empirical research. There has been involvement with an industry partner viz. DAFWA with a view to improving their crop recommendation process in an iterative cycle. Furthermore, a notable outcome of this research has been the DM framework construct brought about through reflection. As a result, this research has some aspects of a cycle of action-reflection which characterises action research.

A number of research activities were carried out in this research. Each of these research activities was specifically designed in order to address and answer each of the three sub-questions expounded in Chapter One. Prior to the commencement of these research activities, a process of data extraction was necessary.

## 3.4 Data Extraction

The process of the extraction of the data was largely *data driven* with the focus established by the research questions. The complexity of the GIS data and the sparseness of the climate data necessitated iterations of the data extractions. Influential factors included the custom and functionality of the available software packages used for the processing of the datasets, the associated proprietary GIS formats, the data availability from separate databases, and the outcomes under investigation.

The datasets comprised the GIS (land use, soils, vegetation, elevation profiles), climate (data measurements from Australia-wide weather stations) and the crop (lupins, wheat, oats, canola and barley) production data streams. All datasets were sourced from various DAFWA databases within the agricultural organisation initially for the state's whole agricultural region.

The data extraction process was fraught with several challenges. Firstly, the scale of the extraction and subsequent storage task was huge. The portability of these unwieldy datasets was facilitated through the use of a 500 GB temporary storage device, while a separate 64bit workstation (laptop) functioned as a portable repository for the multiple datasets.

Secondly, a database was needed to function as a data warehouse. The open source Postgres database was installed on the workstation to facilitate this functionality.

Thirdly, the GIS datasets were inordinately large with each file several Gigabytes in size. Problems with processing logistics of GIS datasets

within Postgres required revision of the data extraction to reduce the size of the original datasets to include only the selected study area.

Fourthly, there were processing problems in relation to grid cell size resolution of the study area. The rectangular study area of 104,328 $km^2$ was created in the ArcMap software suite as a vector shape file and constructed as a grid surface. The vector polygon was divided into multiple square cells of 2500$m^2$. This conversion of vector to raster was done using the spatial tools function in ArcMap. The 2500 $m^2$ cells would have created unprecedented resolution of the datasets possibly with better prediction outcomes. However, this proved untenable in terms of storage and processing power of a moderate workstation with 4GB of memory. Thus a reduction in GIS cell resolution was necessary. Subsequent conversions to raster used cell areas of 1 ha, 5.25 ha and finally 100 ha. This final grid required a square cell size of 1000 m sides and still provided sufficiently fine resolution. The effect of the change in grid cell size on the total number of cells and the resolution is shown in Table 3-1.

**TABLE 3-1. THE GRID CELL SIZE VERSUS RESOLUTION COMPARISON**

| CELL SIZE | DIMENSION | TOTAL CELLS | GIS RESOLUTION |
|---|---|---|---|
| 50m | 15120 x 2760 | 41,126,400 | Very, very high |
| 100m | 7560 x 1380 | 10,432,800 | Very, high |
| 250m | 3024 x 552 | 1,669,248 | high |
| 1000m | 756 x 138 | 104,328 | high |

Fifthly, there were data warehouse functionality problems associated with the open source software Postgres. This challenge was solved by switching to the Microsoft software combination of EXCEL and ACCESS.

Sixthly, the climate dataset contained superfluous data that was not relevant to the study area. The full climate dataset contained measurements of rainfall, maximum temperature, minimum temperature, radiation and evaporation from over 5000 weather stations across Australia. Climate data from the 1300 Western Australia weather stations

was limited to those that fell within the selected study area for the years of interest. This reduced the number of weather stations to 513. The climate data was further restricted to the selected rainfall and temperature profiles for the crop growing seasons from April to September.

Seventhly, the climate data was limited and sparse, and data was needed for each grid cell in the study area. This challenge was solved using interpolation. Initially, this was experimented with in ArcMap, and QuantumGIS using different krigeing methods such as ordinary krigeing and cubic splines. The final interpolation was then formulated in Revolution R as a series of scripts for generating separate stochastic rainfall, maximum and minimum temperature profiles. A separate script was also created for the projection of the stochastic climate data profiles onto the grid surface of the selected study area for the grid cell size stated above.

Finally, the various attributes of data required for this research were from different sources with different formats originally. It was therefore necessary to collate them uniformly as part of the data extraction process. The soil, land use and natural and European vegetation profiles in shape file format were sourced for the whole agricultural region. The elevation dataset was available as a DEM profile with a proprietary ERMapper format. These datasets were available in sections which were integrated in ArcMap and then intersected with the study area for extraction. The elevation, land use, vegetation, soil composition GIS datasets were extracted for the study area. The elevation data and other GIS datasets were extracted in raster and vector format respectively. The GIS datasets were extracted using the proprietary ArcMap, ERMapper and GeoMedia software suites available at DAFWA. The GeoMedia software was used for data intersections, data sub-scenes and cross referencing for the integration of the GIS datasets. The annual crop production data was sourced from the DAFWA crop production database through SQL selections in .xsl format.

The whole data extraction process is indicated graphically in Figure 3-2. It formed the first part of research activity 1 in preparation for the data

processing, data modelling and data storage in the ensuing analyses. In addition, each of the separate investigations carried out for the content of Chapters Five to Eight had their data extractions tailored for the specific variables under scrutiny.

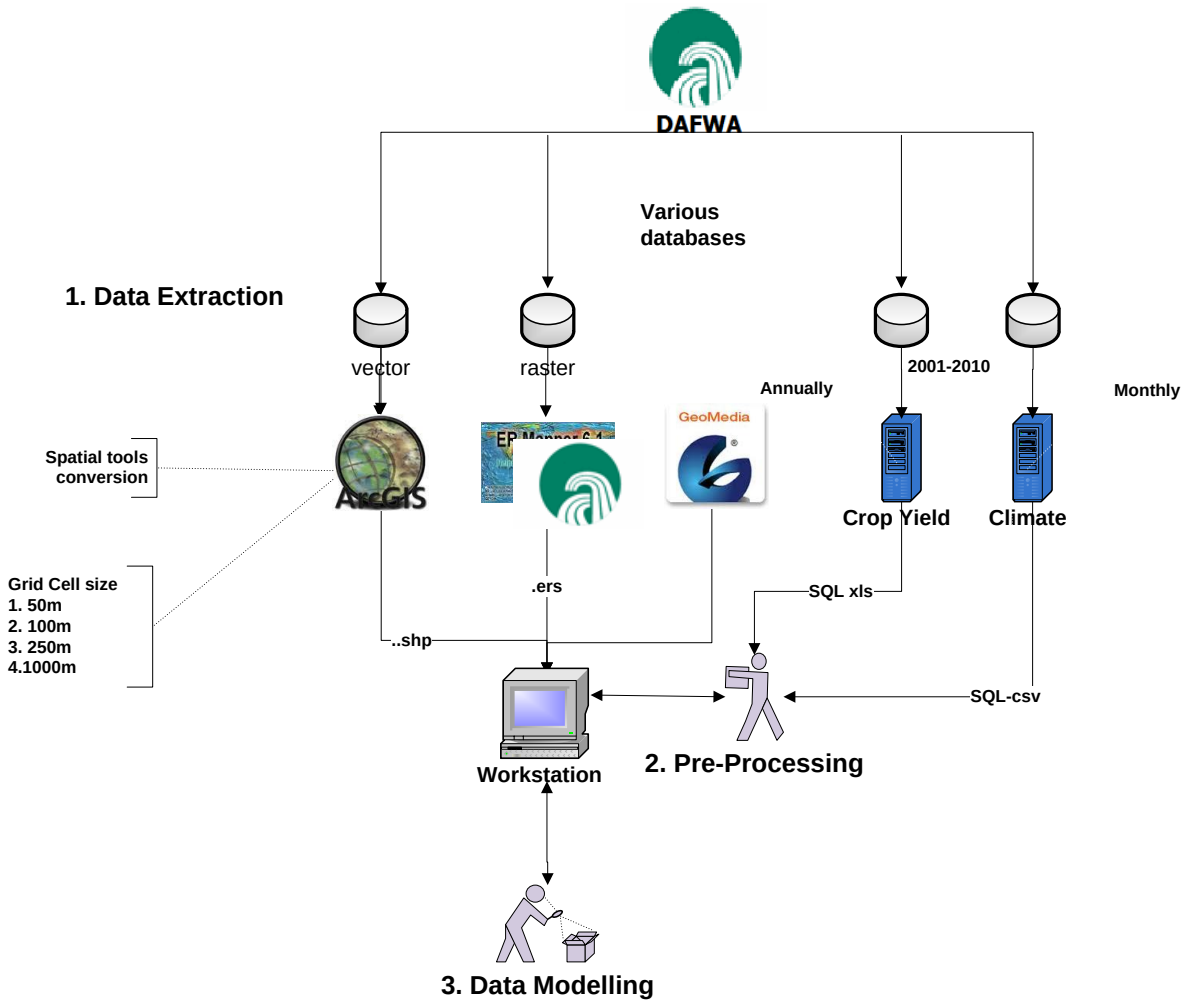



**Figure 3-2. The pictorial representation of the data extraction process**

## 3.5 Research activity 1

This research activity was designed in order to address sub question 1:

"*What are the necessary techniques and methodologies that can be used to develop a systematic approach in order to predict crop yields within shires in the agricultural region of South Western Australia?"*

This research activity comprised the selection of techniques and methods reviewed in the literature with a view to formulating a system of analyses for the processing of complex data to determine the effect of certain climate variables on crop yields within agriculturally productive shires in the South Western agricultural growing region. In essence, this research activity was an embodiment of deconstructing the literature, using the data as a skeleton and utilising prior knowledge as the scaffold (He et al., 2007) as previously outlined in Section 3.2.

In deconstructing the literature, related and similar research and case studies were scrutinised for the techniques and methodologies used. In this way, prior knowledge was built up and this aided the second stage of using the data as a skeleton. Together with data collection, it served as a guide in determining which techniques and methods were suitable for subsequent analyses. In deconstructing the literature from the related and similar previous research in the field, building blocks emerged. The building blocks were in the form of components of a perceived system of analysis. At the top-level they were evident as concepts and methods such as statistics, DM, OLAP and data warehousing.

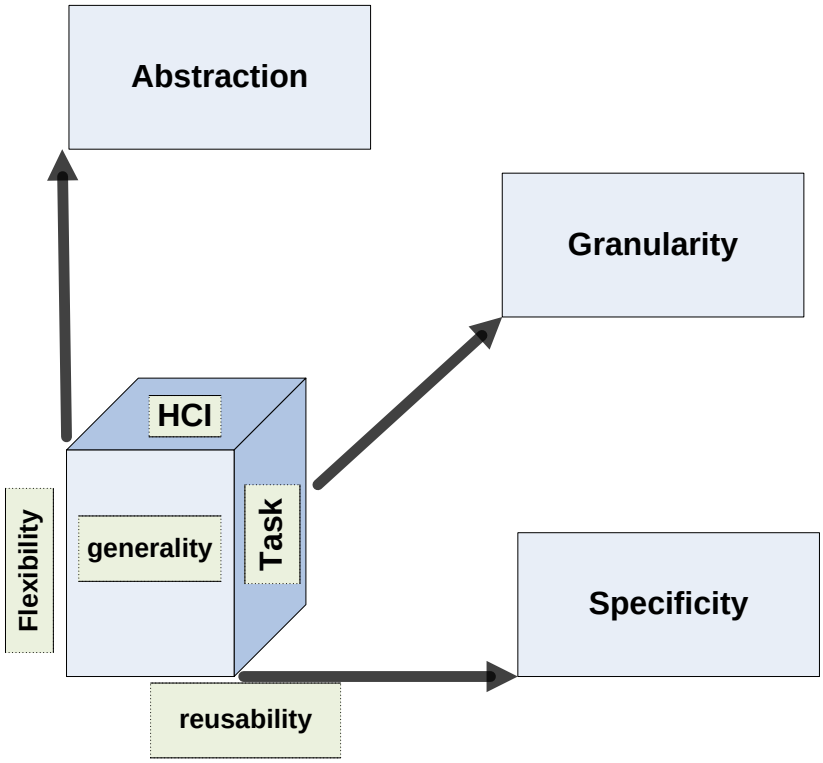
This research activity was therefore exploratory and qualitative in nature as it was basically an evaluation task. In addition, it was about discovering which techniques and methods of analysis within these disciplines actually suited the data and its inherent complexity. During the course of this research, it became apparent that several measures would need to be established in order to facilitate the analyses. Two of these were the formulation of system of analysis into a framework as well as the system of metrics for its evaluation.

### 3.5.1 Metrics of evaluation

The evaluation of the techniques and methods was made qualitatively through comparisons of the components and infrastructure. The main

reason for this was that appraisal of these techniques and methods were inherently subjective and contextual.

The architecture was determined by the contextual background having datasets and tasks that were highly variable (Schulz et al., 2006). The metrics for the evaluation was based on Greenfield and Short's (2003) three axes of critical innovation, as well as the five design criteria for a VDM framework for structures as expounded by Schulz et al. (2006). The three aspects of abstraction, granularity and specificity were blended into the five design criteria of generality, flexibility, usability, efficiency and task orientation. The hybrid schematic for the basis of the metric is depicted Figure 3-3. HCI in the diagram refers to human computer interaction.




.

**Figure 3-3. Re-drawn schematic of metrics of framework evaluation based on Greenfield & Short (2003) and Schulz et al. (2006)**

### 3.5.2 Attributes and scales of measurement

The characteristics and attributes depicted in the schematic representation in Figure 3-3 were used to construct a future evaluation metric for use in comparing and assessing frameworks. The attributes used to assess the different frameworks are listed in Table 3-2.

The architecture attribute is a descriptive comparison and was simply an indication of the nature of the system of analysis used. In addition, the architecture attribute was assessed in terms of structural appropriateness, tasks combination, the dependencies and the splits or joins (Rozinat, de Medeiros, Gunther, Weijters, & van der Aalst, 2008).

**TABLE 3-2. THE EVALUATION METRIC SHOWING ATTRIBUTES AND DESCRIPTIONS OF MEASUREMENTS**

| ATTRIBUTE | TYPE OF MEASUREMENT | DESCRIPTION OF THE MEASUREMENT |
|---|---|---|
| Architecture | Descriptive Type | Categories represented by codes, component(Cfar), assessment (af), theoretical (Lu, Jong, Rajasekaran, Cloughesy, & Mischel), |
| Abstraction | Liekert Scale | 1= low; 4 = very high: criteria for low, medium, high, very high |
| Granularity | Liekert Scale | 1= low; 4 = very high: criteria for low, medium, high, very high |
| Specificity | Precision | Based on the ratio of relevant components to total components |
| Quality | Precision | Based on completeness and understandability |
| Reusability | Recall | Single use or generalisable |
| Task Orientation | Precision | Based on the ratio of relevant components to total components |
| DM Techniques | Descriptive | Codes representing either association, classification, clustering, regression |
| Algorithmic | Indicative | Yes or No indication for the process being stepwise |
| Visualisation | Indicative | Yes, No indication + type of visualisation |
| HCI | Liekert scale | 1 = poor; 2 = average; 3 = good; 4= very good; 5= excellent |

| | | |
|---|---|---|
| Efficiency | Ratio | The ability and use of resources to achieve the outputs (Phillips-Wren, Hahn, & Forgionne, 2004) using data envelopment analysis (DEA). In other words, the ratio of the conversion efficiency. |
| Effectiveness | Scale (liekert) | Concerned with decision outputs (Phillips-Wren et al., 2004) measured as 1 = poor; 2 = average; 3 = good; 4= very good; 5= excellent |
| Components | Descriptive | List of the components |
| Prediction Accuracy | Percentage | Mean Absolute Error (MAE) where the average absolute deviation between the actual yield (DAFWA) and the tested framework prediction is measured. |

Abstraction was a ranking attribute and it was ranked on a Liekert scale of 1 to 4 where 1 = low, 2 = medium, 3 = high and 4 = very high. The specificity was a measure of whether the method of analysis was only applicable for a specific domain and was either single use or generalisable for any context. It was based on the measure of precision.

The quality of a framework or model was based on the measure of completeness and understandability of its algorithmic process. Completeness is the degree of coverage between the user's requirements and the model, whereas understandability was the ease of interpretation of the model by the user (Akoka et al., 2007). The criterion of DM techniques was descriptive and was simply noted as association, classification, clustering or regression. The algorithm attribute was an indicative measure and was simply denoted as dichotomous.

Visualisation was also an indicative measure and flagged dichotomously as either a yes or no. HCI and effectiveness were scale measurements measured as a Liekert scale with 1 = poor; 2 = average; 3 = good; 4 = very good; 5 = excellent. The components attribute was also descriptive, and the last attribute of prediction accuracy was evaluated using Mean Absolute Error (MAE) where the average absolute deviation between the actual crop prediction (DAFWA) and the test prediction was measured.

## 3.6 Research activity 2

This research activity was designed to address the following sub question 2:

"*What issues need to be considered in the application of DM for the prediction of crop yields from agricultural land-use data in WA from a Western Australian shire level perspective?*"

The main focus of the activities in this part of the research was on the Extract, Transform and Load (ETL) phase and the analysis and usage phase. The ETL phase was the pre-processing and formatting of the data whereas the analysis phase dealt mainly with the data mining techniques that were used within the software program WEKA. The ETL phase was the activity that focused on dealing with data structures and data representation, as well as storage and retrieval aspects of data warehousing.

The data mining activity explored the relevance of classification and clustering algorithms that were applicable to the relevant data. The usage phase of the system of analysis was the interpretation of the predictions from the DM mining process and its subsequent reuse. All the pre-processing activities were evaluated with a view to designing a system of analysis that was both algorithmic (stepwise) as well as re-usable.

### 3.6.1 The datasets

Three separate datasets were used, namely a geographical information system (GIS) dataset, a climate characteristic dataset and an agricultural production dataset. The respective datasets consisted of land use topography and mapping elements (GIS), rainfall and temperature elements (climate), and specific crop yields for shires within the agricultural growing region (production).

The first two datasets were used to form the base layer upon which shire and crop yield data were constructed. The datasets were used to first construct the knowledge layer and then establish best practice in terms of predicting the shires with the highest crop yields. This process was in keeping with the data, information and knowledge continuum.

### 3.6.2 GIS Dataset

The GIS dataset had to be first constructed from a series of other datasets with different topographical profiles such as the elevation, soils, vegetation and land-use profiles. The construction of the total GIS dataset was based on the selected study area with a grid cell resolution of 1000 m. In essence, the study area was a geographical surface grid with each cell in the grid made up of an area of 100 hectares. This constructed grid surface was a cadastral map of a subset of the agricultural growing region. The GIS dataset consisted of the following individual profiles:

- **Elevation profile**

The elevation profile was obtained from a Digital Elevation Model (DEM) profile. As the DEM was already in raster format, there was no need for any vector-raster conversion. The DEM was imported into GeoMedia and projected into the study area as a surface grid.

The resultant elevation profile exposed some anomalies as not all the 104328 cells within the study area had height measurements. These were found to be within a section of the study area that was not digitised for elevation as it was not part of the growing region. Other sections which had zero heights were found to be logically attributable to areas of ocean within the study area. Both of these anomalies were ignored.

- ## Soils profile

The soils profile was constructed from three subsets of soil mappings originally done as ESRI shape files. ESRI shape files are a format of vector data that is made of points, lines and polygons of geographic coordinate data. The three subsets were the north, south and central sub-systems of the agricultural region of South Western Australia. The selected study area encompassed most of the central soil sub-system but had some overlapping with both the north and south sub-systems. Consequently, all three soil sub-systems had to be geographically intersected with the study area to produce a merged soil sub-system profile in the form of a shape file that had to be converted to a raster format to produce the grid cell detail for projection onto the surface area grid. This activity was done in the proprietary software GeoMedia.

The resultant soils profile had east/north coordinates and a code for the soil classification. The soil classification was quite complex as each soil group had an apportioned percentage. The dominant soil group with the highest percentage ratio was selected to be the final soil group. As the soil groups were coded, a lookup table was used to translate the codes into the matching descriptions.

- ## Land-use profile

The land-use profile was used to determine the dominant purpose of the land. The land was classified into primary, secondary and tertiary land use purposes in a range from the general to the specific land-use, with the Tertiary Land-Use (TLU) being the most specific. Consequently, the TLU was used for analyses. There were 135 separate land-uses which were variously numerically coded from 110 to 634. A lookup table was used to convert the codes into descriptions in the MS ACCESS database. The land uses of interest, with respect to crop yields, were the cropping and cereal land-uses designated by the codes 340 and 341 respectively.

The land-use profile was first converted into the UTM GDA94 zone 50 coordinate reference system. The shape file was subsequently converted from a vector shape file into a raster file. This was then projected onto the grid surface within the GeoMedia software suite.

- **Vegetation profile**

There were two vegetation profiles supplied in the initial data specification and they included the designated European and pre-European general classes. The pre-European vegetation profile was simply a dichotomous indication of the absence or presence of vegetation at the specified location. Conversely, the European vegetation characteristically noted the actual plant species of the naturally occurring vegetation. This was originally intended for used due to its connection to rainfall, but was later eliminated from the subsequent analyses. This was due to non-relevance to crop yield and because vegetation was not a principal component.

### 3.6.3 Issues with the GIS Dataset

The GIS data had its own inherent challenges, the first being the associated software needed for handling it. To this end, several software packages were evaluated to find the most suitable. The packages included a selection of both proprietary and open source software. The former included ArcMap, ERMapper and GeoMedia while the latter included Revolution R, Udig, QuantumGIS, PostgresGIS, GRASS, GeoDA and SAM. Software feasibility experimentation of the open source packages reduced this to only QuantumGIS, GRASS and Revolution R. The data association with software in a GIS context meant that each dataset emanated from a distinct and separate source thereby attributing to it a multi-source nature characteristic of complex data.

The next issue, related to the software, was the problem with differing formats. The selected open source software packages had the ability to

deal with data from different sources as each proprietary GIS software package produced their own specific data formats. The different profiles of GIS data which were created in the different software packages resulted in different formats which warranted the multi-format classification of complex data.

Size and processing feasibility studies were done through a process of trial and error, as the study area was first formed at a cell size resolution of 50 m by 50 m. When the data extraction at this resolution proved too large, unwieldy and extremely slow, the cell size was progressively increased to 100 m, 250 m and finally to 1000 m. At each trial, the decision to reduce the resolution through an increase in the cell size was made on the basis of processing efficiency.

Another issue associated with GIS data was the Coordinate Reference System (CRS). Prior to dealing with GIS data in any GIS software, the data needs to be converted to the desired CRS. The study area was based on the CRS of using eastings and northings as the x and y coordinates. After first experimenting with the Albers GDA 94 CRS, the Universal Transverse Mercator (UTM) GDA 94 Zone 50 was eventually used in order to match the South Western region of WA. This selection was used to convert the coordinates from latitudes and longitudes to eastings and northings.

### 3.6.4 Climate dataset

The other major data component of this research study was the climate dataset. The climate dataset was obtained as set of coordinate point data that was essentially sparse data due to the fact that the climate variables of rainfall, maximum and minimum temperature, evaporation and radiation that were sourced from the DAFWA, were only specific to the widely dispersed locations of the weather stations at which they were historically recorded. Due to the constraints of scope and focus imposed in a research study, it was deemed appropriate to reduce the climate dimensionality to only rainfall and maximum and minimum temperatures. Nevertheless, the

sparseness of the rainfall and temperature datasets meant that the data had to be transformed from actual and observed measurements to stochastic measurements, whilst at the same time retaining the actual data.

The grid surface of the selected study area had a total of 104328 cells each of which was bounded by a length and breadth of 1000 m, thereby necessitating that rainfall and temperature measurements be generated for each of the cells within the study area. This was done through a process of interpolation using ordinary krigeing. After much trial and error experimentation, the interpolation was finally achieved in the Revolution R statistical package. The custom script for the interpolation is provided in Appendix A1.

### 3.6.5 Issues with the climate dataset

The climate datasets presented several problems generated by the size and complexity of the processing requirements for krigeing, as well as the number of climate attributes such as rainfall, maximum temperature, minimum temperature, evaporation and radiation. Data sampling of the climate data was a major issue. At the outset, the data was sampled in four instances of ten year intervals with climate data for 1980, 1990, 2000 and 2009. This was subsequently changed to 1980, 1990, 2000 and 2010 in the course of data collection in 2011. At that time, it was decided that establishing a climate pattern from such a time separated sample would be difficult.

Consequently, climate data was sampled annually from 1980 to 2010. However, this goal could not be achieved in terms of processing time and available computer power. As a result, it was decided that the analysis would be better served if the climate data was extracted for a seven year period which included the years 1980, 1992, 1999, 2002, 2003, 2005 and 2010. This selection was chosen in consultation with DAFWA experts as it contained a mix of drought years (1980, 2002 & 2010), wet years (1992 &

2005) and highly productive wheat years (1999 & 2003). Due to time and resource constraints, the climate data was initially sampled for the years 2002, 2003 and 2005. This was later expanded to cover the decade from 2001 to 2010 so that a limited time-series evaluation could be carried out. In addition to this major issue other problems emerged while dealing with the climate dataset. These are detailed in Table 3-3.

### 3.6.6 Production dataset

The production dataset was also sourced from the DAFWA. This dataset contained details of the crop yields for each of five crops (lupins, barley, canola, oats and wheat). For the purposes of the aims of this research, only the wheat yields for the shires within the study area were used. The raw data was in tonnes produced by each rural shire within the agricultural region of the South Western Australia. This was subsequently converted to tonnes per hectare of delivery area. The extraction of the crop production data underwent several iterations. This was due to the fact that the first extraction was done using only the three selected years of 2002, 2003 and 2005. When the extraction was expanded to cover the full decade from 2001 to 2010, the formats were different. Consequently, this necessitated a third extraction to keep the formats consistent.

## 3.7 Research activity 3

This research activity was designed to address the following sub question 3:

"*Can the application of DM in crop yield predictions be used to determine future land-uses in Western Australia?"*

This question was answered by using the series of disparate but related datasets to construct a composite and complex case study which was subsequently used to validate the algorithmic system of analyses.

**TABLE 3-3. THE ISSUES WITH PROCESSING THE CLIMATE DATASETS**

| ISSUE | IMPACT | RESOLUTION |
|---|---|---|
| Size of datasets | Out of memory instances in R | Increased the memory capacity of the computer to 4GB |
| Complexity of data | Memory constraints | Use optimised read algorithms |
| Processing time | Delay in reaching the analysis phase | Refine the processing |
| Overall size, complexity, processing time | Data would become unworkable | Scale the data down, by selecting a study area extent and doing an intersect with the data |
| Climate data only as point data | Data for the study area would be limited | Perform krigeing (interpolation) in order to produce data over the full extent of the study area |
| Climate point data for all of Australia | The number of points would be over 5000 | Reduce the number to 1318 by only limiting them to Western Australia and parts of the South. |
| Krigeing of climate point data in GeoMedia | Took approximately 1 day per climate attribute. This was too long. | Reduce the number of points to 278 so that points extend beyond the study area only. |
| Krigeing of climate point data to be done only for the study area | The krigeing would be ineffective as points on the edge of the study area would have reduced references. | Perform an initial krigeing to include point just beyond the study area. Then follow up with a subset. |
| Krigeing was beyond the desired study area | The kriged surface would not match the study area exactly. | Performed a subset of the kriged data to only intersect with the study area. |
| Performing of the subset in GeoMedia. | This was too lengthy and complex. | Found the equivalent of subsetting (clipper) of the study area in QuantumGIS. |
| Krigeing of the climate data produced only a flat greyscale image | It was difficult to determine the bands within the kriged datasets | Performed a 32 greyscale near black function in QuantumGIS to produce a more textured image. |
| The kriged surface did not show the bands graphically | The overlay would not show meaning in relation to the specified climate attribute | Performed contouring of the kriged surface in 10 meter elevation bands. |
| Raw climate data had 5 attributes including rainfall, max and min temperature, evaporation and radiation for each month in a year. | Increases the work required in processing the climate dataset. | Reduced the number of attributes to only rainfall and maximum and minimum temperature |
| Raw climate data had only site numbers of the weather stations | Krigeing requires the x and y coordinate data for position and point referencing. | Added the site reference data in longitude and latitudes. |

| | | |
|---|---|---|
| The coordinates of the site data was in long/lat format | Climate data would not be properly overlaid onto the soil and vegetation datasets | Change the climate dataset site data to have the coordinates in eastings/northings |
| Krigeing of the datasets could only be done with one attribute at a time | This restriction was mainly in GeoMedia | Separate the datasets into three; one for rainfall, one for maximum temperature and one for minimum temperature |
| Krigeing of the 1318 data points took too long | Increases the overall time for processing | Do an SQL join in the Access database to only select the related 278 points |
| Initial climate data was extracted with a four snapshot that included years 1980, 1990, 2000 and 2009. | This would mean that the data would be out of date, by the time the analysis was completed. | Requested a new extract for the year 2010. |
| The selected 4 years snapshot data of only the years 1980, 1990, 2000 and 2010 | This would mean that variations in between would be missed. | Did an extract for each year from 1980 to 2010. |
| Each year dataset was too large in size for processing | Processing for each dataset would take up to much disk space and take too long for processing | Opted for a selection of dry, wet and productive years. |
| The selected 7 years would produce a combination of 7x12x3 (252) datasets | This would make the initial run through of the processing framework too | Opted to do a 1 year trial 2002 (Stahla, Moorea, Floyer, Asplina, & McKendrya) |
| The clipped study area data was in raster format | Any vector operation could not be performed in QuantumGIS. | Convert the raster data to a vector format through the *polygonise* function in QuantumGIS |

The datasets provided GIS, climate and production information. These separate datasets were linked together both relationally and referentially through pre-processing and processing in preparation for the main DM analytical activity. An overview of the whole analytical process follows.

### 3.7.1 Overview of the analytical process

The study relied on the data mining tool WEKA and the statistical analysis packages Revolution R and SPSS particularly for ANOVA and time-series capability. The database repository was Microsoft Access. Data extractions, linkages (joins), and selections from the repository were carried out using the structured query language (SQL). Initial data

preparation and selection were carried out with the aid of other software such as Microsoft Excel and the statistical software package Revolution R. Data reconstructions and transformations and cross tabulations were done in both SPSS and Microsoft Excel. Each of the datasets supplied by DAFWA, were stored in separate folders that were classified into elevation, soils, vegetation, production, rainfall and temperature categories.

The open source software suites such as QuantumGIS and GRASS were used to import the datasets that were supplied as shape files. The study area selection and gridding was done using the proprietary GIS software such as ArcMap and ERMapper. Within these software tools, regions were selected and then prepared for export into a format suitable for conversion into other formats appropriate for WEKA. Specifically, R software was used to load the regions selected as shape files for conversion for importation into WEKA.

Having been extracted, the data was then subjected to exploratory data mining to determine the attributes of the datasets deemed relevant as part of the data selection phase. The selected data was then explored in summary and aggregate form as part of the Exploratory Data (Akoka et al., 2007). Different views from the EDA were then used to analyse the detail using the WEKA tool. Each of the datasets was examined for attribute selection to construct a new relevant dataset. The composite dataset results were then compared with the individual dataset results for integrity as part of the validation phase for the framework. This constructed dataset was subsequently used for detailed analysis within WEKA as part of the activity experiments.

### 3.7.2 The activity experiments

The following group of experiments were conducted iteratively for the effect of the climate variables on wheat yield within the crop growing shires of the selected study area:

1. Rainfall and soil type relationship.

2. Rainfall effect on wheat yield.
3. Temperature effect on wheat yield.
4. Rainfall and temperature effect on wheat yield.
5. Soil type and wheat yield relationship.

Prior to these experiments, the datasets were subjected to several transformations and restructures. The datasets were normalized in that multiple or varied entity representations, and redundancies were removed. This was the data reduction phase, where unwanted and spurious data was eliminated from the samples. This data was then validated to check for accuracy and integrity.

The next phase was the dimensional modelling where the data was somewhat expanded and relevant features and attributes were selected. The lateral expansion occurred because attributes were added from other data-collections. Both the initial reduction and expansion constituted the pre-processing of the dataset and included the removal of outliers through Exploratory Visual Data Mining (EVDM) of the mapped data.

The datasets were subjected to techniques of association, classification, clustering and numeric prediction to find the best technique and associated best fit algorithm. In the secondary analysis aggregation via OLAP tools determined the appropriate scales of measurements for equal comparisons. The results were then analysed and reported upon.

Finally, the two sets of results were pooled and graphed to determine an overall pattern that gave both a microscopic and a macroscopic view of the data. The datasets were then analysed through a series of tests which included cross tabulations, correlations, sequencing, time series and regression. The results of these tests were then scrutinised, interpreted and reported on.

Recommendations were then made as to the best fit regions within the WA agricultural region for the cultivation of crops especially the wheat crop. This whole process is depicted graphically in Figure 3-4 with four distinctive stages of data extraction, pre-processing, data modelling and data

reduction, and data analysis. In addition, the three activity experiments are also charted within the diagram.

## 3.8 Outcomes of the research process

During this research phase and the subsequent evaluation, the system of analysis employed was considered as being repeatable and generalisable. Thus the processes could be adopted as a framework for similar research investigations with minimal adaptation to structure and format, especially with the processing techniques. The research constructed a processing and component framework, named the DM Framework, as the predictive functionality stemmed from data mining.

All of the processes, tasks and activities involved in this research were abstracted and modelled into a process map through the software platform and use of Microsoft Visio. The framework was grouped into three sections: the data warehousing concepts and phases of ETL and integration; administration and monitoring; and the analysis and usage.

The framework could be viewed in terms of a horizontal axis representing the semantic transformations of the data, according to the data-information continuum, and a vertical axis representing the processes of OLAP and DM. The use of OLAP and DM simulated the expansion and contraction of the data to represent macroscopic and microscopic views.

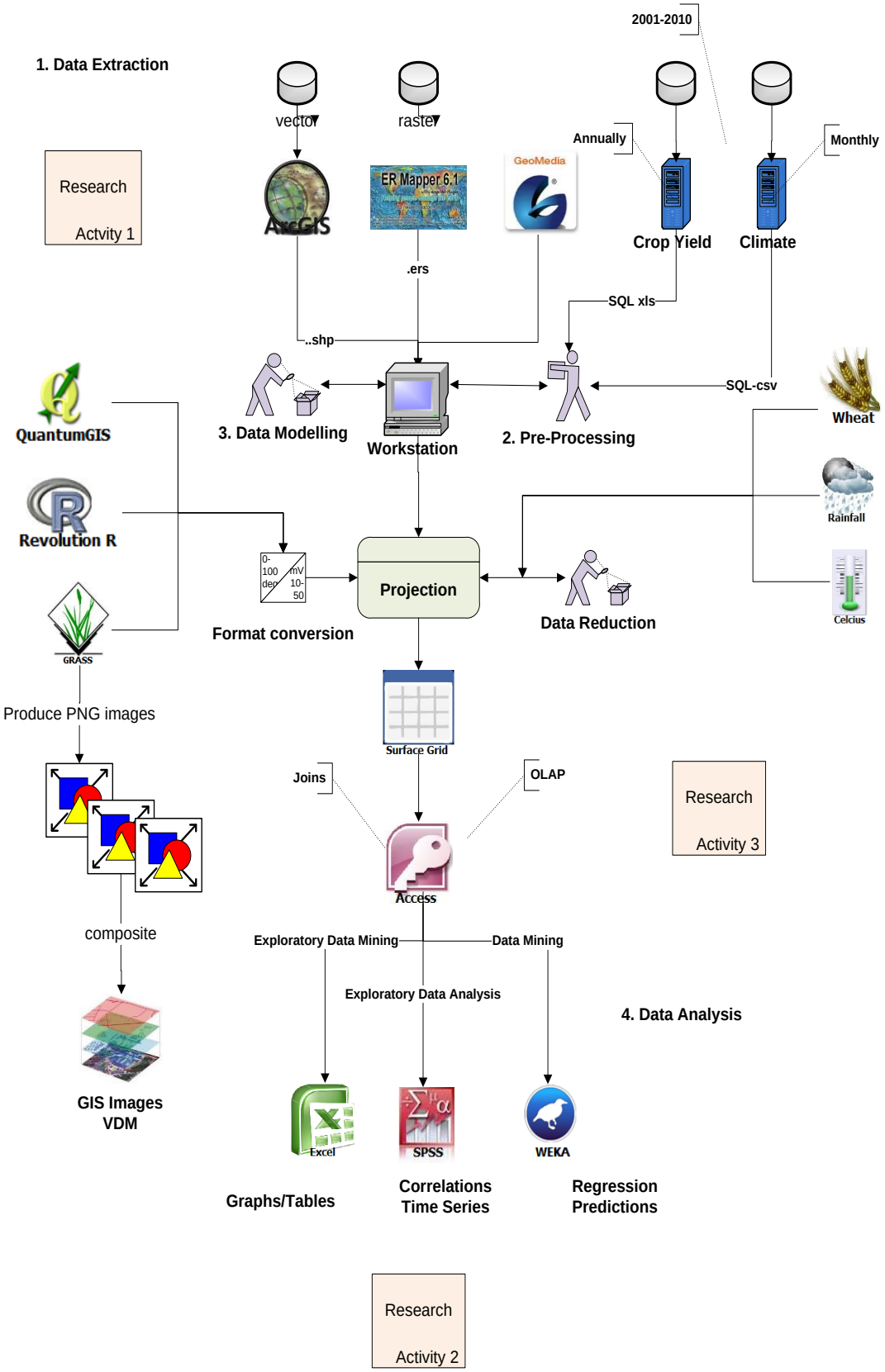



**Figure 3-4. The research design and process**

# Chapter 4

# A DM FRAMEWORK FOR AGRICULTURAL DATA ANALYSIS

This chapter represents the generalisation of conducting research where the data analysed was of a complex nature and originating from two opposing streams of continuous (climate) and gradual variation (location and soil) data. This chapter describes the research in a conceptual and practical way in order to assist in forging the focus and direction of the research. The main aspect of this part of the research was the formulation of the process methodology and a DM framework.

## 4.1 Introduction

This Data Mining framework is a set of generic algorithmic guidelines that may be used for the analysis of highly disparate (format) and variable (scale) data. The general processing framework was developed with the aim of enhancing the analysis of agricultural datasets arising from diverse and complex sources, and is a synthesis of different technologies brought together for the purpose of enhancing the interrogation of these datasets. Aspects of data warehousing phases, Exploratory Data Mining and a post-processing phase for cyclic updating of data as well as for data refinement were incorporated in its construction. The data, information, knowledge and wisdom continuum formed the horizontal axis, with DM and OLAP making up the vertical axis as shown in Figure 4.2.

The DM framework could be used to identify agricultural production areas in Western Australia specifically for crop prediction, planting and harvesting strategies. Farmers may use the results from it to better devise tactical and strategic plans brought about by short-term seasonal variability and long-term climatic changes. These outcomes form part of a

recommendation for industrial best practices in agricultural crop production.

## 4.2 Data mining justification

Many agricultural based research organizations have initiated programs which no longer depend solely on statistical analysis for crop planting recommendations but have incorporated data mining as a feasible alternative. For example, DAFWA is just in the exploratory stage of adopting data mining. Others like the Agricultural Ministry of Pakistan have made the successful transition already and have reported the benefits of increased prediction accuracy as part of their crop management strategy (Abdullah & Ansari, 2005b).

There have also been other research studies which have sought to include OLAP to enhance the data mining experience (Abdullah, Brobst, Pervaiz, et al., 2004). There have been instances of generalization based data mining where object cube models and OLAP are investigated (Han, Nishio, et., & al., 1998). A more recent study, where OLAP is used for quick analysis of aggregates of agricultural data was done in Pakistan in 2009 in the field of pest management in cotton crops (Abdullah, 2009).

This research utilises a DM framework to enhance the interrogation of agricultural data for the purpose of making recommendations to agricultural practitioners. The DM framework addresses the complexity of the agricultural data domain in two dimensions; the information continuum and the online analytical processing. The proposed DM framework also incorporates other aspects such as data warehousing and exploratory data mining.

## 4.3 The Australian context

Within Australia, the identification of potential and latent patterns of different agricultural datasets has been found to be difficult to realize. A

number of crop prediction failures have been reported when only traditional statistical methods were employed to data gathered from Crop Variety Trials (CVT) trials and the seed breeding programs (Burgess & Lamond, 2010). The Western Australian agricultural agency, DAFWA, has sought new approaches such as data mining in order to improve such predictions and their seed variety planting recommendations. New methods of analyses of agricultural data have great potential for farmers and the agricultural industry, given the huge amounts of available historical research, climate, land-use and production data.

There are both short and long-term benefits in improving these analyses. The short term benefits relate to such things as tactical forecasting and prediction as well as to day-to-day management of crops and land-use within the climatic parameters of Western Australia. The long-term aspects relate to strategic forecasting and planning and policy definition. Such benefits have been reported for other regions of the world such as Pakistan (Abdullah, Brobst, Umer, & Khan, 2004).

Although the seed breeding program and crop variety selection process, specific to growing regions in WA, as recommended by DAFWA, has enjoyed considerable success in the past, there have been instances of crop failure in terms of grain quality and grain yield predictions (Armstrong et al., 2007). These failures have been attributed to the use of averages in growth measurements, as well as the trials being limited in field sites and growing seasons. Consequently, the resultant predictions were imprecise for crucial forecasts of crop yield. Precision Agriculture (PA) is a strategy that agriculturalists need to employ so that data, information and best practices may be managed through the use of information technologies that accumulate complex data from multiple sources in the crop production decision making process (Boumaa, Stoorvogela, et, & al, 1999). The proposed DM framework provides a pathway to PA practices for the future, due to its methodical nature.

## 4.4 DM Framework development

General framework fundamentals (Section 2.2) were used in the design of this DM framework and it was developed after previous studies concluded that data mining had distinct advantages over single statistical methods of analysing data in the Western Australian agricultural context (Armstrong et al., 2007). The development of the framework was in conformity with a situation where the hypothesis testing paradigm was the norm (Mitra & Acharya, 2003) in that it began as an idea. The decision to include OLAP as part of the framework arose from other previous studies in the agricultural data mining area which have combined data mining to augment other data analytical processes (Abdullah, Brobst, Pervaiz, et al., 2004).

Metaphorically speaking, the use of DM techniques inspired the creation of the DM framework from a hypothesis generation perspective especially in terms of exploratory data mining as essentially explorations seek to form theories (Hand, Mannila, & Smyth, 2001). In other words, if a *bottom-up* approach is equated to the hypothesis generation perspective and a *top-down* approach is equated to a hypothesis testing perspective, then the DM framework occupies a pathway that is a mixture of both perspectives. In this regard, the DM framework conforms to the defining characteristic of abstraction where the concrete aspect is represented by case studies (data) and the abstract is the actual DM framework itself together with the insights (information, knowledge and recommendations) that the use of the framework provides.

## 4.5 DM Framework description

The constructed framework was based on the data, information, knowledge and wisdom continuum (Cleveland, 1982). The understanding of this continuum is depicted graphically as a natural progression from researching to reflecting (Clark, 2009) as described in Figure 4-1.

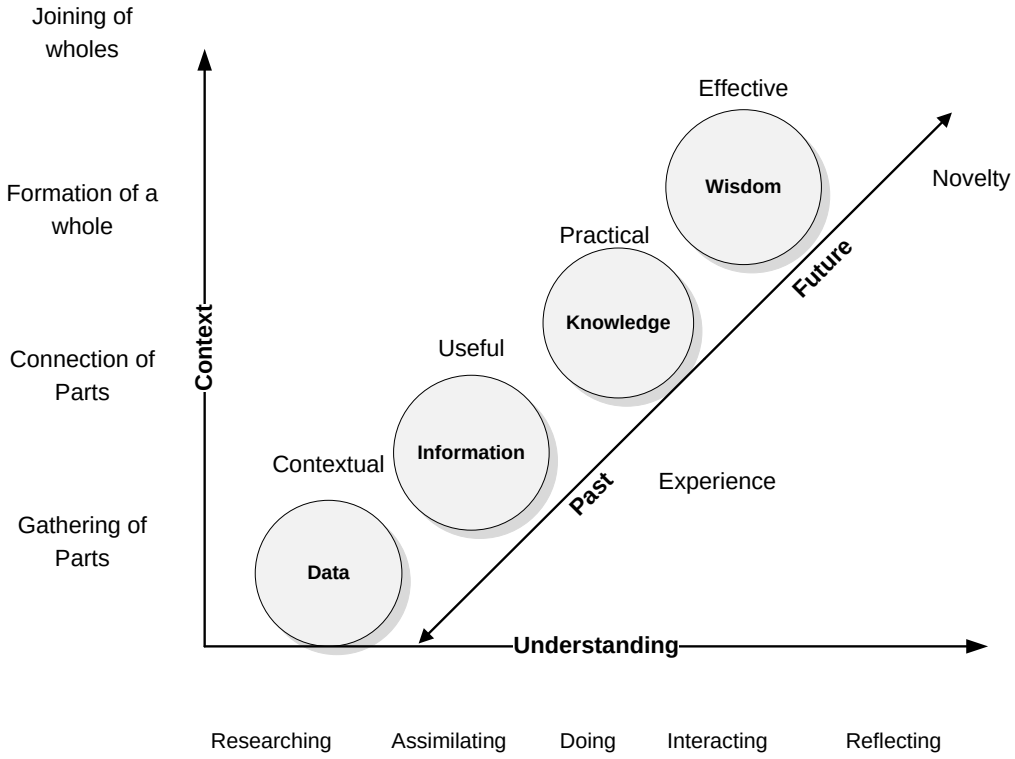



**Figure 4-1. The Continuum of Understanding with modifications (Clark, 2009; Cleveland, 1982)**

Consequently, the data, information, knowledge, wisdom continuum was used as a horizontal baseline for the proposed DM framework. Both data mining and OLAP were included to represent a vertical dimension to the framework but with contrasting viewpoints of the same data. The extraction of information for specific contexts of use, with reference to the agricultural context in this case, represented the transformation into useful knowledge. This transformation is depicted as a theoretical concentration and convergence for practical use through applications such as software tools.

The DM framework assumes a logical process of data capture, storage, processing and customized reporting to end-users. This logical progression assumes top-down significance in relation to its construction. The volume and content of the data interrogated through the use of data mining tools is used with the aim of extracting those components of the data that would be considered to be best-practices within the industry.

Agricultural Industry best practices is a term that may be used to describe the '*nuggets*' of information that could be applied to various and differing agricultural settings while still remaining essentially true.

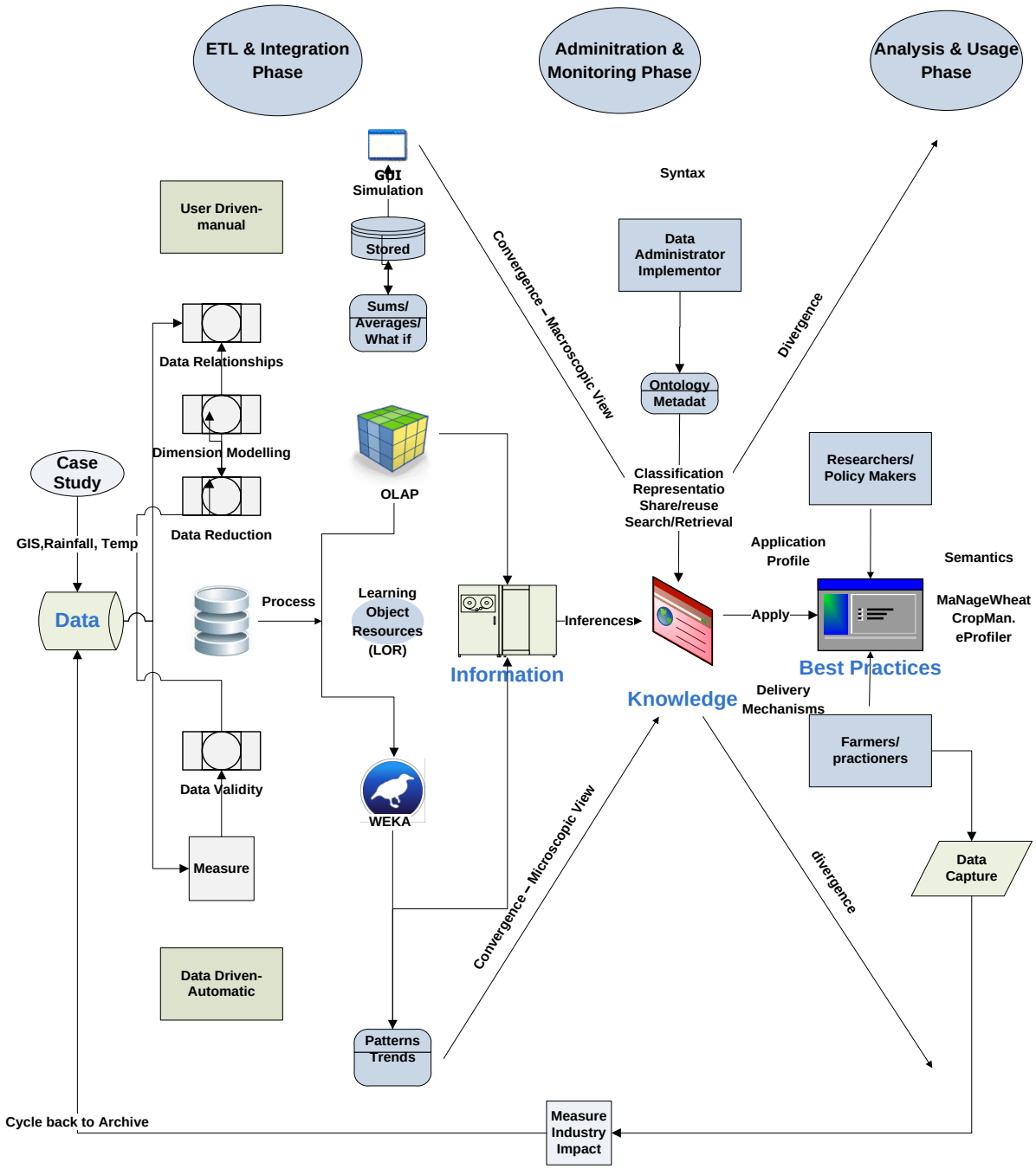



**Figure 4-2. The DM Framework**

## 4.6 Components of the DM Framework

The framework is constituted of several related components and is shown in Figure 4-2. The exploration of the associations and interactions with and between each of the components enabled a framework to be developed. The components were made up of the processes of data capture, data storage, data analysis as represented through DM and OLAP tools, customization of the resultant information, constructing the knowledge base, and formulating the best practice as a defined deliverable for effective wisdom. The principal component used for originating and testing the DM framework was the multi-faceted case study. The validation of the framework in different agricultural contexts imparts a bottom-up significance to its construction that is in keeping with the hypothesis testing strategy (Hand et al., 2001) mentioned earlier.

### 4.6.1 Data capture & storage

The data that was used as part of the multi-faceted case study for the validation of the DM framework was captured over a number of decades and stored in the data archives and databases of the DAFWA. The data existed in three separate databases within DAFWA, and was accumulated from various disparate sources and structured for use within the DAFWA agricultural domain in its capacity as an expert, scientific and industry advisory organisation. The case study for this research was extracted from the archives and databases of DAFWA and copied and stored on a laptop within a newly created database. This database acted as a pseudo or transitional data warehouse.

### 4.6.2 Data Mining and OLAP

Data mining and online analytical processing formed essential parts of the DM framework that made up the transformations of the data into active views. Both data mining and OLAP were therefore the tools of presentation

of the data into useful information. The two mechanisms offered contrasting views of the data. Data mining allowed a microscopic view whereas OLAP presented a macroscopic vision of the same data through the use of aggregations, cross tabulations and pivoting. Data mining has the tendency towards an inductive approach to problem solving while OLAP lends itself to a more deductive way to finding solutions. Validation of the framework was made through the use of such DM tools as WEKA, where hidden and latent trends were sought. General OLAP tools of aggregation, cross tabulations and pivoting were used to provide input for simulations and what-if scenarios.

### 4.6.3 Customising information

In order for the data to be of use in construction of the knowledge base, it required customization as a preliminary process in its pathway to becoming information. This customization was what was termed dimension modelling where only the relevant characteristics of the domain were extracted as part of the research relevancy. The dimension modelling was both expansive and reductive in nature, and was dependant on whether data was added or excluded respectively.

### 4.6.4 Dimension modelling and data structuring

In order to reduce the raw datasets within a data warehouse to a simpler design that aided retrieval efficiency, preliminary data or dimension modelling was performed on the data (R. Kimball, 1997). The dimensional modelling employed for this study was the contextual analysis suitability of data selections. It consisted of entities such as soil type, elevation, vegetation, land use, production tonnage, rainfall and temperature. In some respects, customizing the information to suit the demands of constructing knowledge for a specific domain represented a lateral expansion as opposed to the vertical reduction that took place in the preparatory phases of analysis involving data cleansing.

Data structuring is concerned with the syntax of domain specific data within a warehouse and is about data architecture. It is within this process that data is administered and monitored and where it determines the internal flow of data (McBrien & Poulovassilis, 2001). Architecturally, data is organized into optimal structures for the subsequent analysis and usage phase that feed the results and reporting applications. For example, in this research, vector data such as shape files were converted into grid-cell raster data with referenced geographic coordinates. This conversion resulted in all attributes being specific to the individual cell coordinates of each grid cell.

Whilst dimensional modelling takes place at a point in the beginning of the DM framework, data structuring occurs after the analytical process has returned other facets of the data and where the administration and monitoring phases of data warehousing occurs (Darmont et al., 2005).

### 4.6.5 Knowledge construction

Useful information is passed through analytical tools such as DM and OLAP to construct practical, domain-based knowledge. The construction of knowledge pertaining to a specific domain is brought about by alternating the views of the information between microscopic and macroscopic visions as well as the important feature of uncovering hidden patterns and statistical supports. The DM framework utilises all three processes of DM, OLAP and statistics in order to provide a multi-facet view of complex data especially when the data is multi-format, multi-source as well as multi-structure and multi-modal (Boussaid et al., 2007). This multi-faceted view of the useful information provides an insight into practical knowledge for the selected domain.

For example, in this research, the DM component involved techniques of classification and clustering, OLAP provided different views of the data in terms of aggregations, slicing and cross tabulations. Lastly, statistics was used in regards to correlations and regression analyses.

### 4.6.6 Data constructs within the knowledge base

A body or digital collection of information may be termed a knowledge domain or repository (Manouselis, Kastrantas, & Tzikopoulos, 2006). One of the ways of representing a knowledge domain is ontology formalization. This is a technique that is used to classify and represent the information and associated knowledge so that it is manageable as a stored entity.

Furthermore, the classification and representation of data entities through the added use of metadata allow the knowledge domain to be managed efficiently. This efficient management in turn allows the information and knowledge to be disseminated through enhanced search and retrieval techniques (Maliappis, 2006), thereby transforming the useful knowledge into actionable and best practices in industry.

### 4.6.7 Formulation of recommended practices

Practical knowledge within a specific domain becomes the basis for formulating the industry's recommended practice of achieving a desired outcome. This is achieved when repetitive and exhaustive analyses of similar information return the same or similar results. Some tools for the implementation of best practices in the West Australian context appear in the form of such constructs as ManageWheat, CropMan and eProfiler. In addition, the recommended industry practices become reinforced and refined when the outcomes of utilisation of the practical knowledge for the specific domain are captured and validated for re-entry into the data cycle through the DM framework.

### 4.6.8 Data flows within the DM Framework

There are basically four data flows occurring within the DM framework. They are the external flow, the internal flow, the reference flow and the maintenance flow. The external flow occurs at the two ends of the framework, at the beginning and end of the cycle. The external flow of data

occurs when data is introduced into the DM framework and when it passes out of the framework or disseminated. The internal flows occur when data is being transformed or gets qualified. The reference flow is when relationships between different parts of the data are established and when the data is classified, categorized and structured. The maintenance flow is when revised and new data re-enters the data system, thereby acquiring new knowledge and revising existing knowledge (Darmont et al., 2005).

### 4.6.9 Use of the DM Framework

The DM framework with reference to both Figure 3-2 and Figure 4-2 was used in a series of steps which could be applied to any dataset or application as follows:

- Survey the database to determine what data is available and how it is stored.
- Determine what analysis can be done with the existing data.
- Determine which data to extract.
- Formulate SQL instructions for data extraction and source the data.
- Make copies of the data in the workstation.
- Structure the data and establish a warehouse.
- Repeat the previous steps as necessary.
- Use external software tools such as ArcMap, ERMapper to prepare the study area if data is location specific.
- Use Revolution R to prepare a data grid.
- Use MSACCESS to update data flows into the data warehouse.
- Merge the data with the data grid to form a composite dataset using a script.
- Extract data (SQL, or Excel slicing/dicing/pivoting) from the database to do specific DM analyses.
- Import data into WEKA
- Analyse and interpret results

- Report and make conclusions.

### 4.6.10 Users

The DM framework also incorporates the stakeholder users and the associated capturing of actual practice and outcome data. This user-driven data is cycled back into the body of useful information by the users. The users may benefit from knowledge and wisdom gained from passing data through a general processing data mining framework. Scientists and policy makers look to utilizing the information for best practices theoretically, whilst farmers and practitioners may depend on practical application and decision making.

## 4.7 Evaluation of the DM Framework

This section is an exercise in utilising the proposed system of metrics as depicted in Table 3-2, to evaluate the DM framework. The evaluation is done as a table wherein all the framework characteristics are matched to the descriptions and ratings of the metric. The evaluation of the current DM framework is displayed in Table 4-1.

The precision and recall measurements relate to the used and unused components of the framework. For example, the OLAP component was not used in its entirety as only SQL selections, cross tabulations pivoting and aggregations and scaling were used, while cubing and slicing and dicing were not used. These components represented 30% of the framework components. In addition, not all aspects such as the organisation and

administration of data warehousing were used. For example, the rated value *specificity* was given as *60%* as it was estimated that 40% of framework components were not used. The *efficiency* was rated as 3 out of a possible 5, which represented the use of the components as a ratio of used to unused components.

**TABLE 4-1. EVALUATION OF THE DM FRAMEWORK**

| FRAMEWORK ATTRIBUTE | DESCRIPTION/RATING | COMMENTS |
|---|---|---|
| Architecture | Processing and analytical framework | Primary purpose was to assist in processing data |
| Abstraction | 3- high | Scale (liekert) |
| Granularity | 3-high | Scale (liekert) |
| Specificity | 60% | Precision |
| Quality | 70% | Precision |
| Reusability | Yes | Recall |
| Task Orientation | 70% | Precision |
| DM Techniques | Clustering, classification | Descriptive |
| Algorithmic | Yes | Indicative |
| Visualisation | Yes | Indicative |
| HCI | 3-good | Liekert scale |
| Efficiency | 3 | Ratio |
| Effectiveness | 4-very good | Scale (liekert) |
| Components | DM, statistics, OLAP, no cubes, database, graphs, GIS | Descriptive |
| Prediction Accuracy | Average to good | Percentage |

The DM framework displays the three main characteristics of abstraction, granularity and specificity. The abstraction was evident from the processes of data modelling and data reduction which formed part of the framework. The granularity was proved with the microscopic view of data through the various DM algorithms that were applied to datasets in order to exploit the specific data characteristics of each. The specificity within the DM framework is evident through the formatting of the data for uptake as input to the presentation software. The main aim however, was for the DM framework to demonstrate its effectiveness in improving the accuracy of rainfall, temperature and land use predictions.

The DM framework may be given relevance in the future evaluation of other datasets, making it possible for it to be regarded as generic and re-usable.

In addition, the new framework, as a construct, is capable of providing more detailed and granular information to the analyst, thereby enabling conclusions to be drawn effectively. The proposed DM framework also possesses the ability to be used in part and not just as a whole, thereby imparting to it a dimension of modularity. As an example, the DM and OLAP section may be used independently in a mutually exclusive way.

## 4.8 Chapter Review

This study was innovative in that it introduced additional dimensions to the evaluation of the industrial effectiveness of the data mining and the interrogation of agricultural production data. The expressed aim was to improve the accuracy of climate, yield predictions and land use. The DM framework seeks to demonstrate that raw contextual data may be transformed into useful information, practical knowledge and effective best practices in a horizontal continuum of information. Each transformation is effected by the different processes operating on the vertical plane of the DM framework. The data may be viewed macroscopically through the analytical processes of OLAP and then microscopically through the knowledge discovery processes of DM. This variability in data focus attributes a character of granularity to the DM framework. The DM framework was therefore shown to provide alternate and complementary views of the data, as well as to uncover the hidden and latent patterns that open the analyses of the data to the processes of visualization. Furthermore, the outcomes of the DM framework extend to business intelligence concepts of dashboards for the purposes of immediate and summary information monitoring.

The DM framework incorporates formatting the DM results for uptake as input to presentation software from where farmers and other agricultural practitioners are able to examine the results as best practices in industry. This part of the DM framework lends to it the characteristic of specificity.

In addition, the DM framework takes into account the various methods of information dissemination through portable devices when implemented. These enhancements to the current analyses as provided by the DM framework will serve to augment the effectiveness of climate, crop yield prediction and effective land use in the WA agricultural growing region.

Following a successful adoption of this method of data analysis, the model could be applied to other agricultural areas both nationally and internationally. Furthermore, the method could be extrapolated to other domain specific datasets, thereby granting the DM framework the attribute of re-usability, in addition to the other framework attributes of abstraction and specificity.

# Chapter 5

# VISUAL AND CLUSTER ANALYSIS OF DICHOTOMOUS DATA

This chapter deals with the processing of the dichotomous data from the two input streams of gradual variation data and continuous variation data. The gradual variation data consisted of the GIS vector and raster file types for the land use, soil, vegetation and elevation profiles of the study area of the agricultural region of South Western Australia. The continuous variation data were the climate data (rainfall and temperature) in comma, separated value (csv) format. The aim of this analysis was to explore the techniques and methodologies that would result in providing the ability to predict crop yields at a shire level in the selected study area. This activity was driven and focused by the aim of finding a relationship between the predominant soil type and rainfall.

In terms of the overall framework outlined in Figure 4-2, the analyses covered two of the three aspects of data warehousing where data was subjected to the ETL phase as well as the analysis and usage phase. In addition, due to the GIS nature of the gradual variation data, GIS software was employed at both these phases.

In respect of the research methodology as depicted in Figure 3-4, this activity constituted producing the portable network graphics images and overlaying individual images into a composite so that the process of visual data mining could be facilitated. The focus of the activity was to show the effect of rainfall on soil type through the use of the data mining techniques of clustering and classification.

## 5.1 Introduction

In this research study, the methodology of empirical research dynamics and a case study was employed in constructing a visual data mining

framework for the general processing and analysis of geographic land-use data in an agricultural context. The geographic data was made up of a digital elevation model (DEM), soil and land use profiles that were juxtaposed with previously captured climatic data from fixed weather stations in Australia. In this pilot study, monthly rainfall profiles for a selected study area were used to identify areas of soil variability. The rainfall was sampled for the beginning (April) of the rainy season for the known 'drought' year 2002 for the South West of Western Australia. The components of the processing framework were a set of software tools such as ArcGIS, QuantumGIS and the Microsoft Access database as part of the pre-processing layer. In addition, the GRASS software package was used for producing the map overlays.

Although this processing framework was used to analyse soil and rainfall climate data pertaining to agriculture in Western Australia, it may be easily applicable to other datasets of a similar attribution in different areas. A GIS framework such as this which utilises software for a land information system is dependent upon many factors. The most important of these factors is the input data, followed by the set of data processing functionality (Tomlinson, 2007), especially when decisions are based upon information provided by them. The framework thus represented an algorithmic way of progress towards the achievement of the outcomes.

Evaluation was carried out using techniques of visual data mining to detect the patterns of soil types found for the cropping land use. This was supported by analysis using WEKA and Microsoft Excel for validation. The results suggested that agriculture in these areas of high soil variability need to be managed differently to the more consistent cropping areas.

## 5.2 Framework application

The aim of this chapter was to apply the data mining framework developed previously. The framework was used for the processing and modelling of agricultural geo-spatial data by using DEM, climate, soil and land use data profiles of the South West agricultural region of Western Australia.

Modelling the data statistically, performing clustering and finding association rules are some of the ways to approach data mining problems. Nevertheless, it is important to find inter-relationships between data entities involving location when dealing with data that have geographic attributes (Keim et al., 2004). This chapter deals with these geographic inter-relationships in the context of agricultural land-use. It follows the three-step process typically employed by visual data exploration. These three steps are overview first, zoom and filter, and details on demand. This trio is sometimes referred to as Shneiderman's visual information seeking mantra (Shneiderman, 1996).

## 5.3 Data extraction and relevancy

The data used in this study had different attributions and were made up of five separate but related entities. All of the datasets were in relation to the South West region of Western Australia.

As mentioned in Section 3.4, the separate datasets were made up of climate, soils, land use and DEM features. The data was extracted at the Department of Agriculture and Food of Western Australia (DAFWA) in Kensington, Perth. The datasets were extracted using ArcGIS from where the data was first projected into the UTM GDA94 zone 50 coordinate reference system measured in centimetres. The study area was a rectangular grid of 104328 cells of 1000m each, stretching from Busselton in the west to Esperance in the east. The ERMapper software was used to create the study area and the DEM extent as a raster (.ers) file. All the datasets were fitted specifically to the extraction region of the selected study area.

The data extraction and relevancy involved experimentation with a number of software tools that represented an admixture of extraction, pre-processing, analysis, data mining and visualization of GIS and climate data sourced from DAFWA. The whole process that pertains specifically to the

subject of this chapter is depicted graphically in Figure 5-1. The process was divided into five basic phases. These included *sourcing the data* and *test iteration, pre-processing, processing, data analysis* and *image evaluation.* Each of the phases denoted in Figure 5-1 consisted of a number of steps.

The first step (stage 1) was to source the data from DAFWA. This was achieved by extracting shape files of the GIS data for the whole South West agricultural growing region and then doing a subset for the selected study area. A detailed description of all the datasets has been provided in Table 5-2. The raw data extraction was followed by a preliminary processing phase (stage 2). This was where the datasets were adapted for software input. The software in this stage included ArcMap, QuantumGIS, ERMapper and the MS ACCESS database. The processing phase (stage 3) involved the QuantumGIS, and the Revolution R statistical packages. The data analysis phase (stage 5) and image evaluation phase (stage 4) dealt mainly with software functionality from GRASS, WEKA, MS ACCESS and MS EXCEL. Details of these phases are provided in the following sections.

## 5.4 Complex data processing

Pre-processing was the second step (stage 2) as denoted in Figure 5-1. In this stage, the study area (extent) was loaded into QuantumGIS and then re-projected into the UTM geographic reference system (GRS) resulting in coordinates with eastings and northings. The DEM, soil and land use layers were also similarly re-projected and then intersected with the study area for relevance. The rainfall data was imported into QuantumGIS as delimited text and then linearly interpolated to cover the study area in grid cell sizes of 1000m.

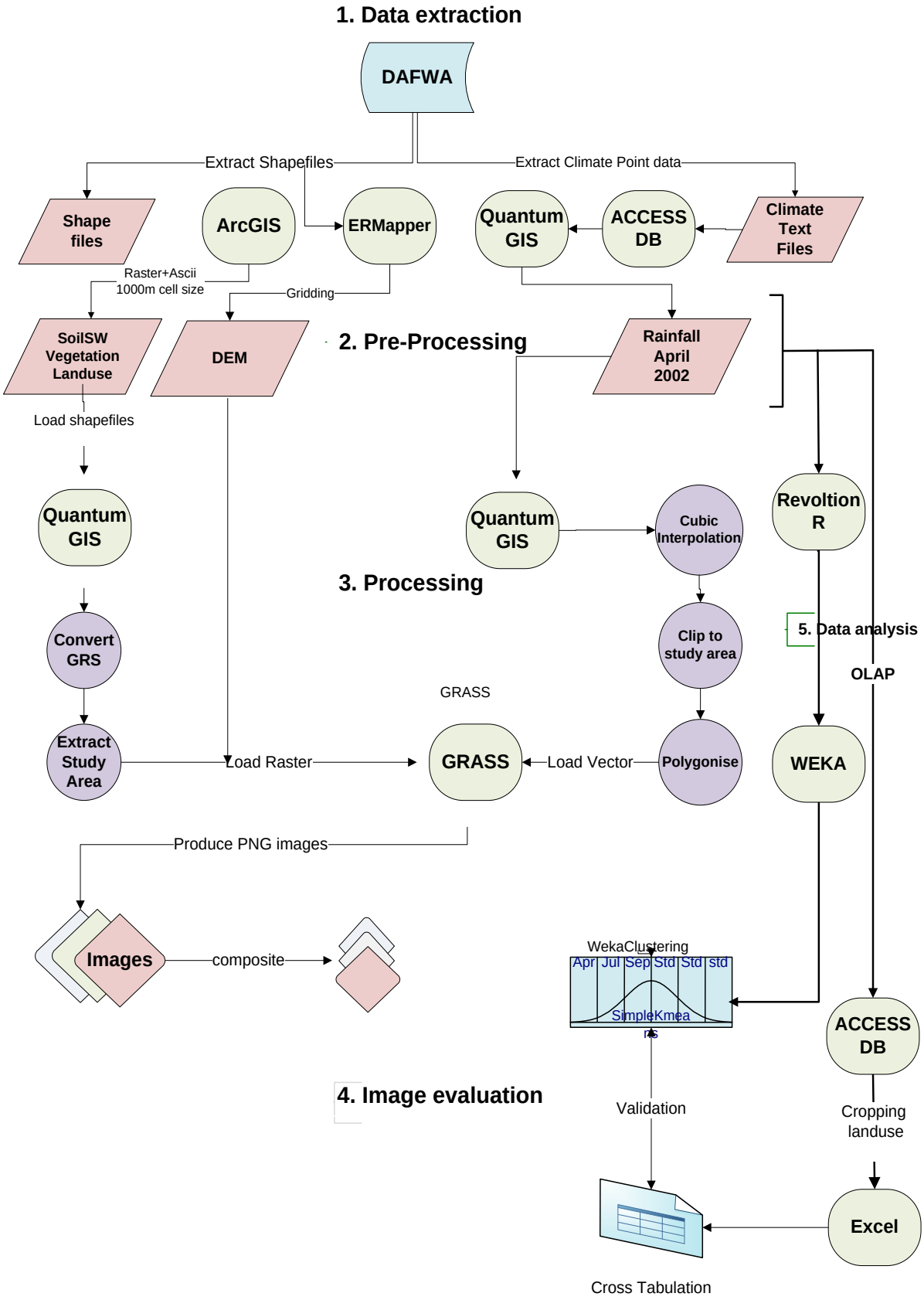



**Figure 5-1. Process methodology of the VDM of climate and geographical data**

The next step (stage 3) involved the processing of the data layers within the GRASS environment. The series of steps for the third stage (3) was performed in the GRASS software package in order to produce an output that would have some visual semantics and from which a discussion could ensue and subsequent conclusions could be drawn.

Subsequent steps were data analysis (stage 4) and image evaluation (stage 5). The first part of the processing stage was to load the raster and vector datasets. Apart from the rainfall data which was in vector format, all of the other datasets were in raster format. Table 5-1 illustrates the exact nature of the different datasets. The spatial type in Table 5-2 is an indication of the GIS data which occurs in two digital formats namely, raster and vector data. Raster data are basically numerically coded grid cells or pixel data, while vector data are comprised of coded points, lines or polygons (Congalton, 1997).

**TABLE 5-1. THE COMPONENTS AND STRUCTURE OF THE COMPOSITE MAP FOR FIGURES 5-2 TO 5-5**

| LAYER NO | DATASET TYPE | RESOLUTION | FEATURE | MAP OPACITY | PROFILE ORIGIN | SPATIAL TYPE |
|---|---|---|---|---|---|---|
| 1 | DEM | 1000m | elevation | 80 | ERMapper | raster |
| 2 | Soils | 1000m | mapping unit | 80 | ArcMap | raster |
| 3 | Land use | 1000m | tertiary land use | 80 | ERMapper | raster |
| 4 | Rainfall Apr 2002 | 1000m | average monthly rainfall | 25 | QuantumGIS | vector polygons |

The work done in producing the composite images of Figures 5-2 to 5-5 represented two aspects of analysis that were not merely a function of the software suite. Instead, they were an exercise in the discovery of the best way to analyse the relationships between the data through effective presentation, modeling of the data, and evaluation and interpretation of the images. In this pilot study, only the rainfall for April 2002 was overlaid to determine the correlations.

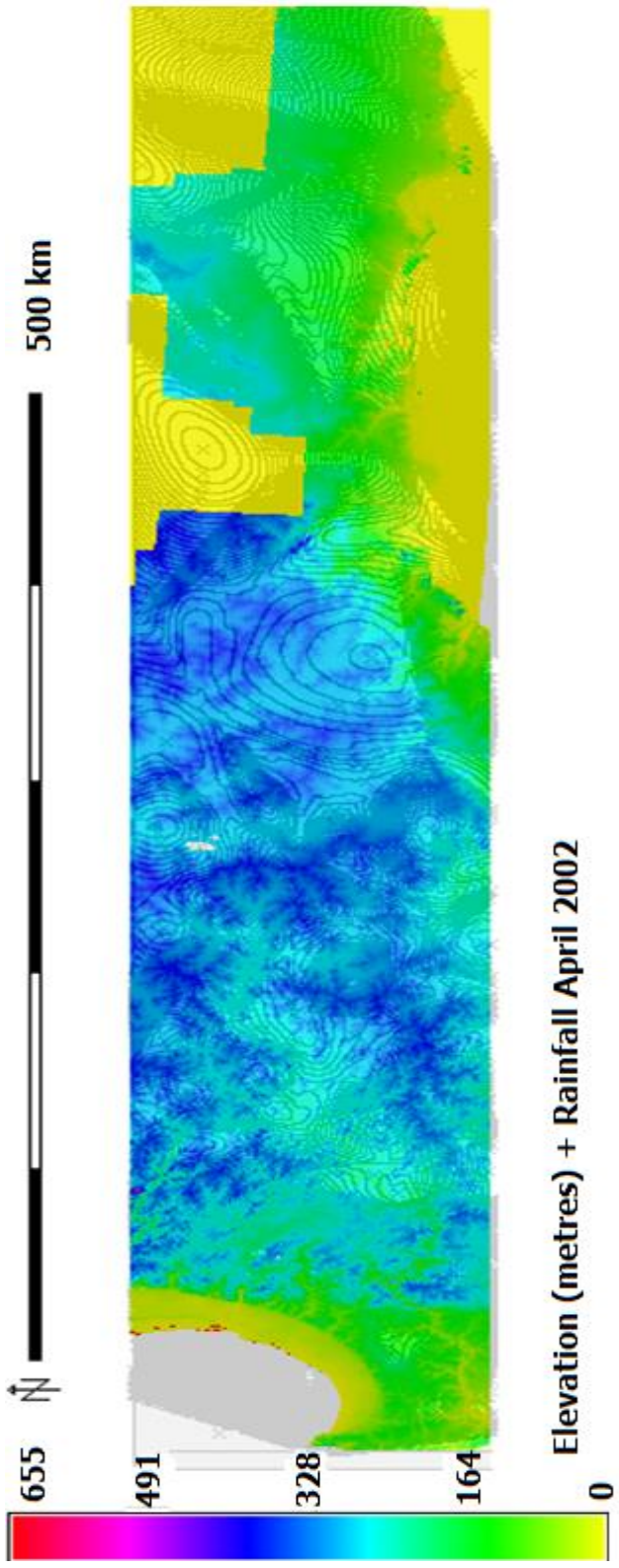



**Figure 5-2. The visual correlation of the underlying DEM (elevation) and the average monthly rainfall for April 2002**

The first dataset that was loaded in the GRASS workspace was the DEM for the elevation characteristic of the study area. This was overlaid with the rainfall data for April 2002 as depicted in Figure 5-2. Rainfall for April was for the start of the 'rainy' season and the year 2002 was a known 'drought' year. With reference to Figure 5-2, the colour-coded scale on the left is a measure of the height above sea-level. For example, yellow represents sea-level with a height of zero metres.

The values for the elevation model corresponded well with the Stirling ranges where the elevation peaks at around 350m. Most of the areas within the western coastal and southern coastal regions had zero elevation due to the ocean. The two anomalous north eastern regions within the study area that were coloured in yellow indicating a zero elevation, were the areas that had not been digitized for elevation within the DEM map. The green areas corresponded to land with elevation between 0 – 200 metres. The few red dots along the west coast were considered to be anomalous data as they were also part of the ocean with zero elevation.

The second dataset loaded into the display area was the soil type followed by the average monthly rainfall for April 2002 as depicted in Figure 5-3. The colour-code scale on the left of Figure 5-3 represents the different soil types in number codes. The predominant soil type over the agricultural region was light blue with some patches of light green and yellow. The soil types that corresponded to these colour codes were not easily identifiable from the GRASS image and could only be recognized with a more detailed analysis after the mapping unit codes were translated.

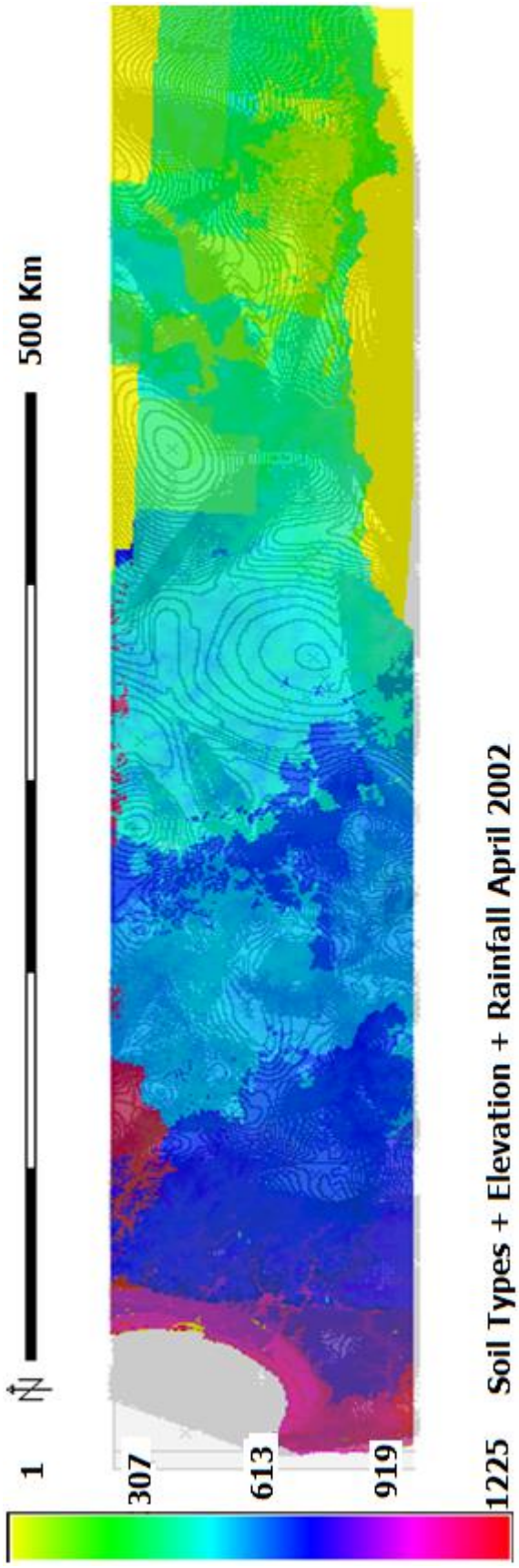



**Figure 5-3. The visual correlation of the underlying soil type and the average monthly rainfall for April 2002**

The third dataset examined was the land use mapping which was again overlaid by the average monthly rainfall for April 2002 as depicted in Figure 5-4. The predominant land use across the agricultural growing region was for cropping (code 340) denoted by the dark blue colourations on the map. The other shade of light blue had feature values between codes 341 to 360, and represented other farming activities such as cereals, grazing, horticulture and fruit tree plantations. The bright green areas represented mainly the natural parks and reserves.

The map in Figure 5-5 shows the distribution of the different rainfall bands over the cropping areas. Generally the coastal areas received higher rainfall as represented by the green (101-150mm) and brown (51-100mm) patches, whilst the majority of the agricultural area received rainfall in the 2 bands of 0-25mm and 26-50mm rainfall as represented by the pink and grey patches.

Each of the separate rainfall bands featured a predominant soil type associated with the amount of rainfall received. This breakdown of soil types to rainfall band is shown in Table 5-2.

**TABLE 5-2. PREDOMINANT SOIL TYPES IN THE DIFFERENT RAINFALL BANDS**

| RAINFALL BAND | PREDOMINANT SOIL TYPE |
|---|---|
| 0-25mm | Grey deep sandy duplex, Yellow/brown deep sandy duplex & Duplex sandy gravel |
| 26-50mm | Grey deep sandy duplex |
| 51-100mm | Grey deep sandy duplex |
| 101-150mm | Brown loamy earth, Brown deep loamy duplex & Friable red/brown loamy earth |
| 151-250mm | Wet soil & Semi-wet soil |

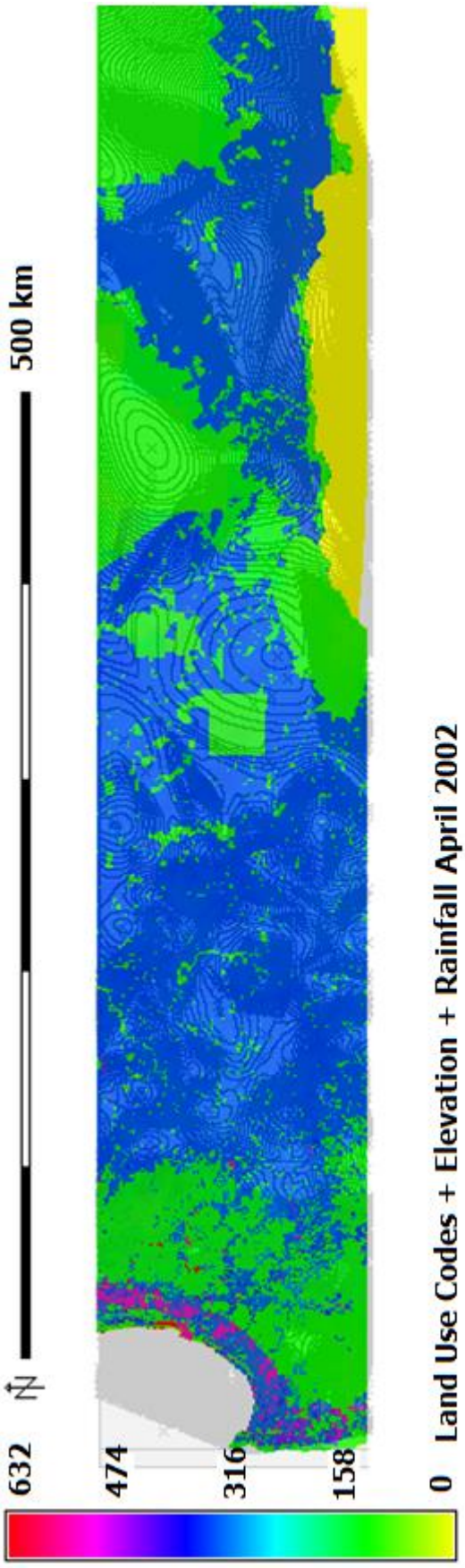



**Figure 5-4. The visual correlation of the underlying land use and the average monthly rainfall for April 2002**

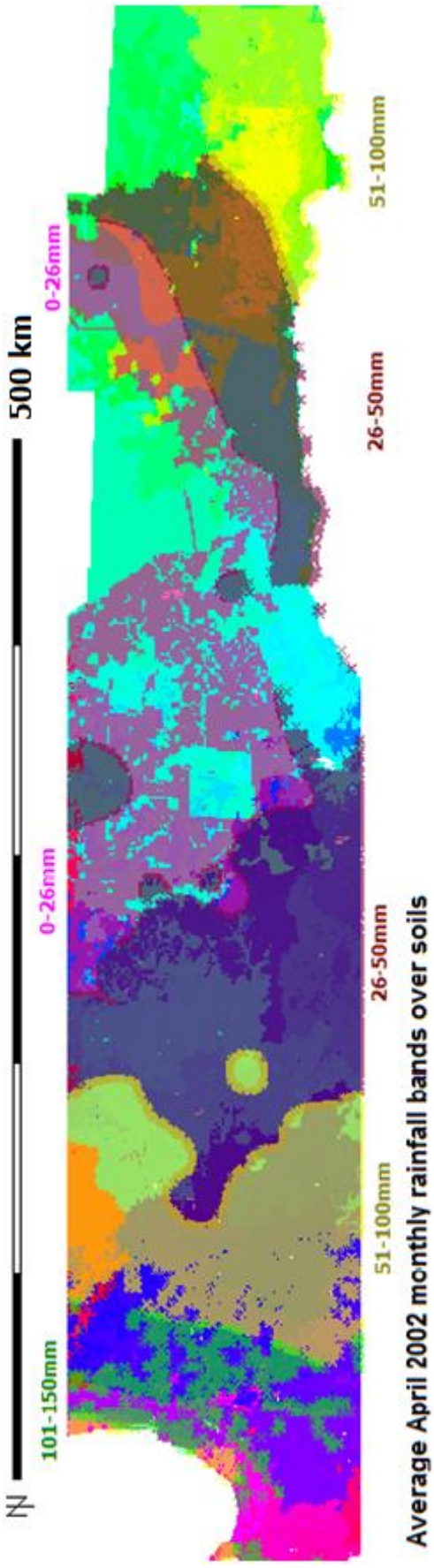



**Figure 5-5. The visual correlation of the underlying soil types and the bands of average monthly rainfall for April 2002**

## 5.5 Visual and cluster analysis

In essence, the visual and cluster analysis spanned the two phases of data analysis and image evaluation as denoted in Figure 5-1. In terms of the visual analysis, the underlying raster image layers formed the dependant variables for the attributes of soil type and the rainfall data was the independent variable. The PNG images displayed in Figures 5-3, 5-4 and 5-5 visually denoted the visible correlation between the rainfall and soil type.

Rainfall during the start of the 'rainy' season in the south-west region clearly falls at a greater concentration along the coast and then tapers off as it progresses inland. There were some large pockets of high rainfall which matched well with the underlying soil type layers (green and light green colouring) that corresponded to the actual agricultural growing region in Figures 5-3, 5-4 and 5-5. The results were collated into two tables for the analysis of stage 5. Table 5-3 represents the observations and visual analysis from the GRASS images of Figures 5-3, 5-4 and 5-5.

**TABLE 5-3. THE VISUAL ANALYSIS OF THE COMPOSITE MAPS**

| MAP | PREDOMINANT COLOUR | SALIENT FEATURE EX WEKA |
|---|---|---|
| DEM | Light blue | Height of 100-200 metres |
| Soil type | Light blue, light green, yellow | Loamy, deep sandy, wet |
| Land use | Blue, light blue, green | Cropping, cereals, hay & silage, seasonal horticulture, irrigated tree fruits |

Table 5-4 on the other hand, represents the WEKA cluster analysis of the average rainfall for the month of April 2002. The rainfall figures obtained from the cubic interpolation and prediction done in the Revolution R statistical package were validated by cluster analysis done in WEKA. This was basically the data analysis section. The cluster method used was simple K-means clustering with a selection of 10 clusters.

**TABLE 5-4. WEKA SIMPLEKMEANS CLUSTER RESULT OF THE APRIL 2002 RAINFALL**

| WEKA CLUSTER NO | CENTROID EASTINGS | CENTROID NORTHINGS | NO OF CLUSTER INSTANCES | % OF TOTAL CLUSTER INSTANCES | ACTUAL APRIL 02 RAINFALL (CENTROID) |
|---|---|---|---|---|---|
| 0 | 999397.68 | 6259963.22 | 10238 | 10 | 56.44 |
| 1 | 521305.50 | 6281938.90 | 8072 | 8 | 139.17 |
| 2 | 882156.89 | 6275105.18 | 11737 | 11 | 21.67 |
| 3 | 384458.62 | 6268890.72 | 9394 | 9 | 44.26 |
| 4 | 495964.35 | 6330428.90 | 11150 | 11 | 32.19 |
| 5 | 682647.66 | 6259607.85 | 11044 | 11 | 89.68 |
| 6 | 553535.22 | 6235644.05 | 11539 | 11 | 27.50 |
| 7 | 923405.75 | 6320754.49 | 11913 | 11 | 27.32 |
| 8 | 661851.52 | 6338114.87 | 8838 | 8 | 64.03 |
| 9 | 682961.36 | 6303379.62 | 10403 | 10 | 51.97 |

The clusters 1, 5 and 8 recorded the highest rainfall figures for April 2002 and these were concentrated on the coastal region to the west of the study region. The centroids for the moderate and high rainfall areas were all located in the agricultural growing region in the middle to right of the study area in Figures 5-3, 5-4 and 5-5.

## 5.6 Correlation of soil types and rainfall

The numbers 1 to 27 represented the 27 different soil types in the agricultural cropping region after the data reduction process from the original 682 soil types. The graph in Figure 5-6 showed that most of the rainfall falling in the agricultural cropping region was mainly within the first three bands of 0-25mm (9745 instances), 26-50mm (13695 instances) and 51-100mm (7398 instances) for the average monthly rainfall of April 2002. These were the blue, red and green bars in Figure 5-6 respectively. The soil types which corresponded to the highest rainfall concentrations of over 600 instances in the agricultural cropping region are shown in Table 5-5.

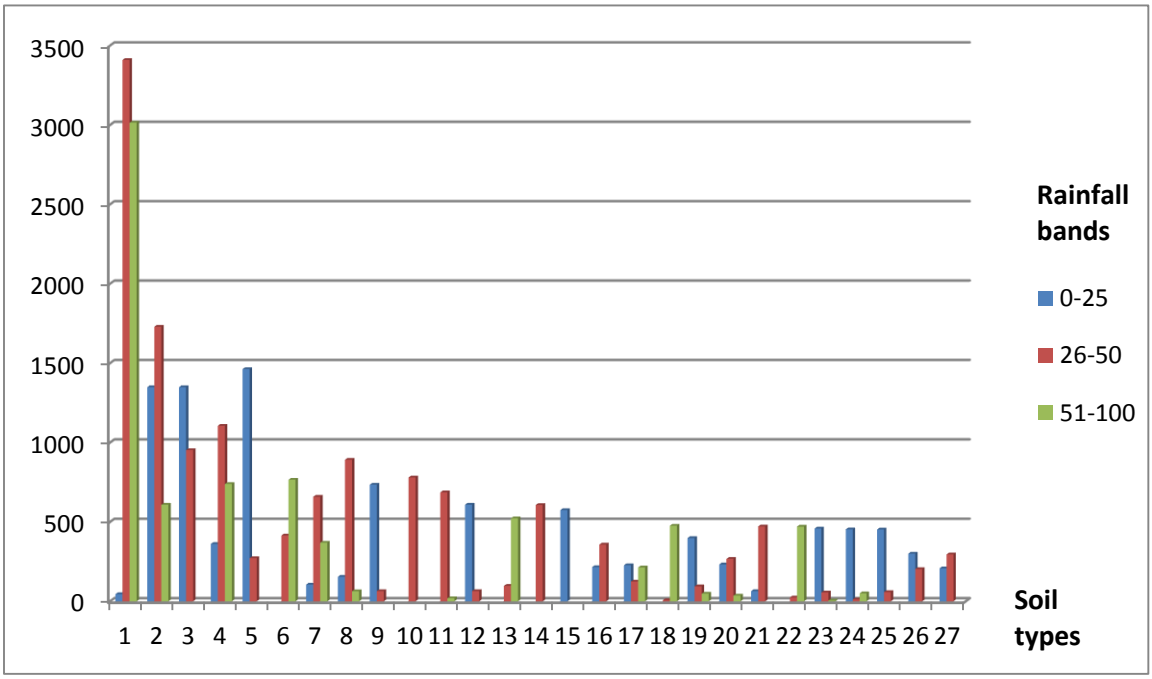



**Figure 5-6. The average monthly rainfall (April 2002) versus soil types**

It is evident from the table that all soil types except numbers 4 and 6 were sandy soils. In addition, the graph in **Figure** 5-6 indicates that different soil compositions pertain to each rainfall band. For example, the dominant soil types which receive the lowest rainfall of between 0-25mm were the numbers 2, 3, 5, 9, 12 and 15 and, the dominant soil types receiving rainfall between 26-50mm were the numbers 1, 2, 3, 4, 7, 8 10, 11 and 14. The soil types receiving the high rainfall averages were the numbers 1, 2, 4 and 6. The combination of the three bands therefore showed a predominance of the first 5 soil type numbers. The results also suggested that higher than 100mm of average monthly rainfall did occur within the agricultural cropping region for some of the 27 soil types such as soil type number 6, 18 and 22. However, these instances only represented 3.2% of the total for the other three selected rainfall bands. Nevertheless, these areas with high rainfall variability may indicate that different soil management techniques may be needed for them due to the rainfall received.

In addition, for the purposes of validation and support, the data mining technique of classification using the J48 algorithm was performed on the dataset containing only the 27 soil types of interest. This was done in order

to determine a relationship between the results obtained from the two separate methods. Although the accuracy of the classification was only 18.86%, an examination of the area under the ROC curve showed some correlation between the two results. For example the soil types designated in blue at the bottom of Table 5-5 showed a high ROC area value.

**TABLE 5-5. THE SOIL TYPES WITH THE HIGHEST RAINFALL INSTANCES FROM THE CROSS TABULATION AND WEKA CLASSIFICATION**

| SOIL COMPLEX DESCRIPTION | RAINFALL INSTANCES IN BANDS | TOTAL PERCENT | ROC AREA EX WEKA |
|---|---|---|---|
| Grey deep sandy duplex | 6468 | 20.97 | 0.70 |
| Alkaline grey shallow sandy duplex | 3685 | 11.95 | 0.70 |
| Saline wet soil | 2205 | 7.15 | 0.62 |
| Pale deep sand | 1179 | 3.82 | 0.75 |
| Grey shallow sandy duplex | 1132 | 3.67 | 0.69 |
| Grey deep sandy duplex, Bare rock & Duplex sandy gravel | 620 | 2.01 | 0.86 |
| Duplex sandy gravel | 483 | 1.57 | 0.84 |
| Duplex sandy gravel, Loamy gravel & Deep sandy gravel | 494 | 1.60 | 0.84 |

## 5.7 Conclusion

The use of a single case-study of geographical data of soil profiles combined with average monthly rainfall data for the month of April and the year 2002 was complex. This was due to the non-existence of a processing and analytical precedent in this context. Accordingly, it was a demonstration of the challenges of working with diverse and multi-source datasets that effectively covered the various dimensions of raster, vector and delimited text data.

Furthermore, the study was an exercise in both displaying a solution visually as well as visually finding a solution as espoused by (Kovalerchuk, 2004). The focus was on the three main aims of visualisation namely; presentation, confirmatory analysis and exploratory analysis (Garcia et al.,

2004). In this way, the information was visually represented, allowing for direct interaction whereby insights, conclusions and decisions were gained, drawn and made respectively (Keim et al., 2008). This work began with the intention of finding a relationship between rainfall, land use and soil substrate and to thereby demonstrate that there was some justification in utilising land for agriculture given the climatic conditions and to also determine future land use depending on the soil type.

The results did indicate a relationship between the soil type and the average monthly rainfall as was evident from Figure 5-6. However, this relationship needed to be investigated further using non-categorical data. Nevertheless, this pilot study helped to uncover a foundation methodology comprised of the different levels of software, data and analytics that was useful in devising and testing a framework for the analysis of complex data with a geospatial dimension.

## 5.8 Chapter review

The validation methods used in this study involved the production of an innovative portable network graphics image that served as a composite picture for visual inspection and analysis. Components of human intuition and background knowledge were used as part of the visual analytics to find solutions and to display them graphically for meaningful interpretation.

The results indicated that particular soil types were associated with consistent rainfall. Farmers, agricultural experts and consultants could use these initial results as a foundation upon which to develop further strategies for soil management in agricultural areas. Farmers could make use of this method to obtain an exact soil composition of their own growing regions given the specific geographic coordinates or a general shire name.

This investigation formed the basis for the next three experiments. Firstly, this study would be extended to examine the effect of rainfall on vegetation and food crops. Secondly, it would be extended to portray a greater

emphasis on the climatic data in terms of expanding the analysis to time-series rainfall data covering different years and months. Thirdly, the analysis would be expanded to cover temperature in order to uncover or confirm further relationships.

# Chapter 6

# THE EFFECTS OF RAINFALL ON CROP YIELD

This chapter deals with the development of the crop model with rainfall and wheat crop yield as the interacting variables. The rainfall profile was generated in the Revolution R statistical software package through the use of a script for interpolation. The wheat crop yields within the crop growing shires for the selected study area were analysed to examine how variations in rainfall impacted upon the wheat crop output.

## 6.1 Introduction

The aim of this chapter was to find a relationship between rainfall, land location (shire) and crop production. This was done firstly to justify agricultural land-use, and then to predict the crop yield at certain locations within the agricultural region, given the rainfall. The geographic data was made up of land use profiles. The land use profiles were then linked through coordinates to previously captured rainfall data from fixed weather stations in Western Australia. Interpolation of the rainfall data was necessary. This was due to the sparseness of the weather stations that recorded the climate data. The rainfall data was therefore interpolated using ordinary krigeing in order to generate stochastic measurements. These estimates were then fitted onto the grid surface of the selected study area.

The resultant estimations for the annual rainfall profiles for the selected study area within the South West Agricultural region of Western Australia were used to identify areas of high crop production. The areas within the study area were spatially scaled to individual shires. The rainfall was sampled for a distributive mix of low and high rainfall as well as high crop yield characteristics. The wheat crop yields could then be examined in

instances where the rainfall was the independent factor. As a corollary, instances of high wheat yield were examined against rainfall. The years chosen were therefore 2002 (low rainfall), 2003 (high crop yield) and 2005 (high rainfall) as per the DAFWA classification. These years were chosen to match these two streams of enquiry.

## 6.2 Historical context

As far back as in 1979, Anderson singled out low rainfall as the factor behind most of the adversities of the agricultural sector (Andersen, 1979). In fact, in the past for example, low rainfall has been attributed to be the cause of large decreases in production yield, as much as 18% in 1983 (Campbell, Crowley, & Demura, 1983). According to Olesin and Bindi (2002) the factors of radiation, temperature and rainfall all affect yield to some degree with rainfall especially affecting the growth and production of the plant (Cantelaub & Terres, 2004). Variability in rainfall from year to year is closely intertwined with crop yields (Lobell & Field, 2007). Although a great deal of the variability of Australian rainfall may be related to the either the El Nino – Southern Oscillation (Duckworth & Stephenson) or Sea Surface Temperature (SST) phenomena (Rasmassun & Wallace, 1983), previous studies have shown the predictions not to be altogether accurate as pointed out by McBride and Nicholls (1983), Pittock (1984) (Pittock, 1984; Rasmassun & Wallace, 1983) as well as by Nicholls (1985) (Nicholls, 1985).

Down-up scaling which is the assignment of links to differently scaled variables (Habersack, 2000) was not carried out in this research as the yield was not correlated with an oceanic index such as ENSO. This study however, considered the relationship between crop yield and rainfall similar to studies such as Challinor et.al (2003) (Challinor, Slingo, Wheeler, Craufurd, & Grimes, 2003) where there was an established physical basis such as a spatial scale upon which the variables operated. Spatial scales are important in that the scale related results are specific to the related

group or agency (Chase & Leibold, 2002). For example, the national scale may be used by governments to determine their economic strategy from food reserves (Atwood, 1991), while results from smaller scale relationships, being used to detect food shortages and associated mitigation possibilities (M. B. Smith, and S. Davies, 1995), and for seasonal forecasting by farmers at the farm level (Eakin, 1999).

This study differs from Challinor's in that it will investigate the correlations between these factors on a spatial scale that is somewhere between the small farm and the large agricultural region levels. Consequently, we have chosen the shire level which is designated as medium and related to local rural government. This was due mainly to the availability of crop yield data on this level as well as the way in which the Western Australian system operates. In order to limit the analysis for the purposes of focus and scope as well as to highlight the relationships under rainfall and yield variability, three specific years were chosen. The production years of 2002 with low rainfall, 2003 with high crop yield, and 2005 with high rainfall were sampled as these were the years showing a considerable variation in both the attributes of rainfall and wheat yield. Furthermore, the selection was designed to avoid the problem of excessive instances of rainfall within each cell of the study area grid.

## 6.3 Related work

The use of observed relationships such as rainfall to predict crop yield has been undertaken by Parthasarathy through an empirical model (Parthasarathy et al., 1992). Interactions between input variables such as rainfall and output variables such as crop yield have been shown to be important (Mearns et al., 1997; Semanov & Porter, 1995). In particular, their relevance was emphasized at critical phenological growth stages (Wheeler et al., 2000).

A relational analysis such as this work needs to be based on a model. There are two approaches for developing models. The first approach is the

process based Crop Model (CM) for the establishment of non-linear relationships between weather variables and crop yield. The second is the General Circular Model (GCM) for the coupling of the ocean and the atmosphere. The use of GCMs for prediction at a seasonal lead-time have been shown to suffer from the problem of simulation of too many low intensity rainfall instances within each grid cell (Ines & Hansen 2006).

Previous researchers have concluded that the complexity of a model can be based on the level of detail of the analysis (Brooks, Semanov, & Jamieson, 2001) or it can be less detailed with only estimations of moisture content (Martin, Washington, & Downing, 2000). Other approaches have been the use of the Normalized Difference Vegetative Index (NDVI) for grouping homogeneous regions to establish the scale (Basso, Richie, Pierce, Braga, & Jones, 2001). There have also been approaches to crop modelling using the derivation of a Probability Distribution Function (PDF) for the assessment of quantifying the risks and benefits of making weather based decisions (Cantelaub & Terres, 2004).

Despite the considerable improvements in understanding and predicting climate variability, the need to further develop understanding and refine tools is ever increasing (Sivakumar, 2006). This, according to Sivakumar (2006), is especially because the atmosphere is intrinsically a chaotic system. As short term weather forecasting rather than long-term climate forecasting is important (Challinor et al., 2003; Murray, 2012), this study is therefore more focused on and suited to the CM approach.

## 6.4 Research design

In order to develop a crop model approach to the research, various lines of investigation were attempted. These included sourcing GIS data, GIS software and analysis tools. After the initial extraction of the GIS data from DAFWA, it was found that GIS data was rather complex and disconnected as each of the land-use, vegetation, soil, elevation, climate and crop production profiles had different formats. The first task was to test different software for hosting the data. Among the software tested were the open

source QuantumGIS, GRASS, PostgresGIS, Revolution R as well as the proprietary ArcMap, ERMapper, GeoMedia and Matlab. During the data-software feasibility stage it was found that the processing of GIS datasets required supercomputing facilities. Further details of these and other related issues have already been covered in Section 3.6.3.

One of the lines of investigation was to join the separate images into a composite one and analyse the resultant composite image for correlation. The use of composite images is a commonly used principle for analysis (Durieux, Lagabrielle, & Nelson, 2008). This approach did not work well as problems were evident in the construction of the composite image. These may have been related to the *black box* effect of packaged functions. Therefore, analysis of the composite image could not be proceeded with. Consequently, the alternate line of investigation of combining the separate data profiles into a composite dataset rather than a composite image was used. Analysis of composite datasets is a well-known method of investigation (Allum, Sturgis, Tabourazi, & Brunton-Smith, 2008).

The GIS datasets covered the whole agricultural region of Western Australia. As this proved to be a formidable task in terms of processing power, time and scope of the analysis, it was decided to limit the research to a specified study area.

## 6.5 The study area

The original raw data sourced from DAFWA was made up of separate geographical land-use profiles, rainfall data and crop production data. The ERMapper, ArcGIS and GeoMedia software packages were used to create the study area, and the various GIS profiles. The *study area* was a rectangular grid of 104328 cells each of which was 100ha in size (1000m by 1000m cell). This study area formed the basis of the analysis for this chapter as well as for chapters seven and eight. The study area extent was from Busselton in the west, to Esperance in the east for the South Western

Agricultural region covering a total area of 104328 square kilometres as noted Figure 1-1.

Each of the shires within the study area was made up of a number of the 100ha grid cells whose coordinates were matched up with the interpolated rainfall data. All the individual datasets were fitted specifically to the extraction region of the selected study area.

## 6.6 Rainfall data interpolation

Historical rainfall and climate data existed only at sparsely located weather stations within the study area. There were only 278 weather stations with recorded climate data within the coordinates of the study area. These weather stations and the associated climate data points were scattered over the total study area. This shortfall of data points was due to the fact that it was not possible for complete recordings due to the infinitesimal nature of the landscape and the limited number of weather stations (DA Roshier, 2001). In order to overcome the limitation of a sparse rainfall dataset, a process of interpolation was carried out resulting in stochastic rainfall data points at each cell of the study area.

Interpolation and approximation of missing data from a sample of the dataset is done by estimating the values of the unmeasured points at variously specified locations and times (Apaydin, Sonmez, & Yildirim, 2004). The approximation functions may not preserve the original measured values due to generality, whereas interpolation on the other hand, does maintain the original measurements (Revesz, 2010).

There are many methods of interpolation. These include the Thiessen polygon technique, the Lagrange approach using least squares, as well as the techniques of inverse distance, multi-quadratic, optimal and krigeing (Tabios & Salas, 1985). Spatial interpolation is interpolation of data when the surface structure is taken into account .The common forms of spatial interpolation are ordinary krigeing, lognormal ordinary krigeing, inverse

distance weighting and splines (Robinson & Metternicht, 2006). The method used in this study was ordinary krigeing.

The interpolation of the rainfall data for the selected study area was carried out in the Revolution R script (Appendix A1) for each of the 12 months for the years 2001 to 2010. The customized script used eliminated the black-box effect of using built-in algorithms. The interpolated data was then projected onto the study area for a grid surface fit by matching the coordinates of interpolated to each cell of the surface gird. The interpolation was done at a high resolution of 1000m by 1000m (i.e. 100ha cell). Uncertainty due to spatial elevation bias was minimal due to the weather stations being a homogenous group. The procedure for the interpolation is given in the following paragraph.

- Obtain the measured climate values for example rainfall, recorded at each weather station from the DAFWA.
- Obtain the list of weather station numbers and the associated coordinates.
- Connect the measured climate values to the weather station coordinates within the ACCESS database using a join.
- Set the grid cell size to 1000m.
- Perform an ordinary krigeing procedure in the Revolution R software package, on the measured rainfall values to obtain the interpolated dataset. (Refer to Appendix A1 for details of the krigeing method used).

## 6.7 Dataset compilation - rainfall

After the generation of the interpolated data was completed, a special purpose database was built that contained the full dataset with all the features of land-use, vegetation, soil type and shires. Shire locations for each point on the surface grid were matched to the rainfall dataset using joins within the database. As analysis was to be done only for the low rainfall year of 2002, the high yield year of 2003 and the high rainfall year

of 2005, only the rainfall data for these years were subsequently extracted. The coordinates of the extracted interpolated rainfall data were matched up with the production data for the same three selected years using joins of the grid numbers of each cell within the database.

The production data only existed at an annual level for each shire within the study area. Consequently, the rainfall data had to be aggregated to an annual level to match the annual crop yield data for the years 2002, 2003 and 2005 only. The sum of the annual rainfall at each crop producing shire was then divided by the number of cells within each shire to obtain the average annual rainfall in each shire. This aggregation of the individual shire, 12 monthly rainfall data was automated using an R Script (Appendix A2). In addition to the annual rainfall, values for the seasonal monthly rainfall and the seasonal average rainfall were also used.

A process of data reduction was performed on the full study area dataset by only selecting land uses reserved for cropping and cereals. The dataset was sorted into ascending shire name order. The rainfall for each of the cells within each shire was scaled by the number of cells within each shire to produce a single line for the shire, average annual rainfall, and annual wheat crop yield for each shire. This dataset was then used for the subsequent analyses using rainfall and wheat yield. The pseudo-code for this process is shown below:

- Set the directory to the source file location.
- Initialise the rainfall variables for the selected years.
- Initialise the loop counter variables and first shire name variable.
- Set up the output matrix with the total number of shires (34 rows) and total output variables (53 columns).
- Transform the matrix into a data-frame for processing.
- Name all the columns in the data-frame.
- Read in the full dataset as a comma separated value file.
- Sort the input file into alphabetical shire order.
- Set up the progress bar for tracking completion.
- For each shire data,

- Display the shire name and the counter.
- Calculate the totals and average rainfall for each selected year.
- Total the crop production values for each crop.
- Read the next shire and test if it was the same as the current one.
- If different, end the loop and write the shire output file with the number of datapoints for each shire, rainfall and crop data and reset all the counters.

- Write out the last output file and do test read of the output.

A snapshot of the final composition of the full composite dataset used for the analyses in this chapter as well as the two succeeding chapters of seven and eight is depicted in Figure 6-1.

## 6.8 Experiments and analysis

There were a number of aspects to the data handling and analyses. They included the pre-processing and metrics, the analysis of the rainfall and the analysis of the wheat yields which formed the macroscopic phase. Conversely, the DM analysis of the wheat crop yields formed the microscopic phase. In order to focus the analysis and to aid the recognition of the patterns, pre-processing and a system of metrics was required.

### 6.8.1 Shire categorisation

The final analysis was carried out on a shire level for the two attributes of stochastic average annual rainfall and actual crop yield within each shire. The annual crop production data was calculated in tonnes per hectare, in order to account for the different shire sizes. The visual inspection of the graphs required a uniform method of evaluation. This was in the form of a baseline metric of classifying the shires into rainfall categories of very high

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Shire | rnApr | rnMay | rnJun | rnJul | rnAu | rnSep | mxApr | mxMa | mxJun | mxJul | mxAug | mxSep | mnApr | mnMa | mnJun | mnJul | mnAu | mnSep | vrApr | vrMay | vrJun | vrJul | vrAug | vrSep | WhtY |
| 2 | Boddington | 39.6 | 23.5 | 43 | 57 | 39.3 | 28.5 | 23.57 | 21.6 | 16.78 | 16.7 | 16.88 | 18.28 | 10.95 | 8.03 | 6.03 | 5.77 | 4.06 | 5.85 | 12.62 | 13.56 | 10.75 | 11 | 12.82 | 12.4 | 2.30 |
| 3 | Boyup Brook | 53.9 | 35.8 | 55.7 | 76 | 58.3 | 43.9 | 22.53 | 21 | 15.89 | 15.7 | 15.95 | 17.4 | 10.46 | 8.12 | 6.31 | 5.63 | 5.17 | 6.67 | 12.07 | 12.9 | 9.58 | 10.1 | 10.78 | 10.7 | 3.31 |
| 4 | Broomehill | 57.1 | 36.2 | 56.4 | 76 | 60 | 44.9 | 22.67 | 22 | 16.54 | 16.3 | 16.71 | 18.46 | 11.23 | 9.24 | 6.88 | 6.08 | 5.53 | 6.51 | 11.44 | 12.78 | 9.66 | 10.2 | 11.18 | 12 | 2.02 |
| 5 | Cranbrook | 59.1 | 33.5 | 58.9 | 78 | 64.5 | 49.8 | 22.47 | 21.8 | 16.5 | 16.2 | 16.57 | 17.89 | 11.09 | 9.35 | 7.18 | 6.27 | 5.8 | 6.89 | 11.38 | 12.41 | 9.32 | 9.94 | 10.77 | 11 | 2.91 |
| 6 | Dumbleyung | 39.3 | 27.2 | 50.7 | 66 | 47.5 | 30.4 | 23.35 | 22.2 | 16.74 | 16.5 | 16.98 | 19.17 | 11.96 | 9.39 | 6.49 | 5.65 | 4.99 | 6.14 | 11.39 | 12.85 | 10.25 | 10.8 | 11.99 | 13 | 1.21 |
| 7 | Esperance | 49 | 35.1 | 55.7 | 73 | 55.3 | 39.9 | 23.09 | 23.2 | 18.48 | 18.4 | 19.13 | 19.82 | 12.95 | 11.01 | 7.97 | 7.18 | 6.35 | 7.12 | 10.14 | 12.15 | 10.51 | 11.3 | 12.78 | 12.7 | 1.11 |
| 8 | Gnowangerup | 56.9 | 33.4 | 56.8 | 75 | 57.8 | 45.2 | 22.54 | 22.8 | 17.32 | 17.2 | 17.63 | 19.14 | 11.68 | 9.85 | 7.3 | 6.11 | 5.84 | 6.69 | 10.86 | 12.93 | 10.02 | 11.1 | 11.79 | 12.5 | 1.70 |
| 9 | Jerramungup | 55.1 | 33.9 | 54.8 | 74 | 57.2 | 43.2 | 22.82 | 24 | 18.74 | 18.8 | 19.04 | 20.27 | 11.44 | 9.94 | 7.21 | 5.89 | 5.51 | 6.51 | 11.38 | 14.1 | 11.53 | 12.9 | 13.53 | 13.8 | 0.71 |
| 10 | Katanning | 49.6 | 36.4 | 57.2 | 76 | 60.1 | 40.2 | 22.87 | 22.2 | 16.64 | 16.4 | 16.91 | 18.83 | 11.37 | 9.12 | 6.62 | 5.91 | 5.27 | 6.36 | 11.5 | 13.05 | 10.02 | 10.5 | 11.64 | 12.5 | 2.08 |
| 11 | Kent | 46.1 | 36.3 | 56.7 | 74 | 56.3 | 37.4 | 22.97 | 23.2 | 17.57 | 17.4 | 17.99 | 19.78 | 11.65 | 9.57 | 6.7 | 5.61 | 5.16 | 6.17 | 11.32 | 13.63 | 10.87 | 11.8 | 12.83 | 13.6 | 1.38 |
| 12 | Kojonup | 57.5 | 35 | 56 | 75 | 59 | 44.2 | 22.53 | 21.4 | 16.11 | 15.8 | 16.18 | 17.74 | 10.79 | 8.68 | 6.59 | 5.85 | 5.23 | 6.58 | 11.74 | 12.75 | 9.52 | 9.98 | 10.95 | 11.2 | 3.17 |
| 13 | Kulin | 41.8 | 32.3 | 59.8 | 68 | 47.9 | 33.1 | 23.67 | 22.4 | 16.82 | 16.6 | 17.13 | 19.49 | 12.23 | 9.3 | 6.31 | 5.48 | 4.72 | 5.94 | 11.44 | 13.09 | 10.51 | 11.1 | 12.41 | 13.6 | 0.68 |
| 14 | Lake Grace | 41.9 | 31.3 | 55.6 | 66 | 48.3 | 32.2 | 23.83 | 23.6 | 17.78 | 17.7 | 18.4 | 20.59 | 11.75 | 9.56 | 6.34 | 5.53 | 4.74 | 5.92 | 12.08 | 14.05 | 11.44 | 12.2 | 13.66 | 14.7 | 0.79 |
| 15 | Narrogin | 38.5 | 27.8 | 55.3 | 62 | 45.5 | 27.5 | 23.19 | 21.6 | 16.38 | 16.3 | 16.32 | 18.37 | 11.54 | 8.73 | 6.03 | 5.41 | 4.53 | 5.74 | 11.65 | 12.88 | 10.35 | 10.8 | 11.79 | 12.6 | 1.97 |
| 16 | Ravensthorpe | 48.3 | 35.9 | 57.6 | 75 | 58 | 40.3 | 24.09 | 24.6 | 19.13 | 19.5 | 20.1 | 21.58 | 12.61 | 11.01 | 7.73 | 6.79 | 6.26 | 7.49 | 11.48 | 13.57 | 11.4 | 12.7 | 13.84 | 14.1 | 0.67 |
| 17 | Tambellup | 61.4 | 38.3 | 60.7 | 79 | 63.5 | 49.1 | 22.54 | 22.1 | 16.69 | 16.4 | 16.82 | 18.37 | 11.34 | 9.61 | 7.26 | 6.31 | 5.9 | 6.78 | 11.2 | 12.48 | 9.43 | 10.1 | 10.92 | 11.6 | 2.85 |
| 18 | Wagin | 43.5 | 23.7 | 47.7 | 61 | 45.6 | 30.2 | 23.17 | 21.7 | 16.45 | 16.2 | 16.46 | 18.46 | 11.49 | 8.86 | 6.19 | 5.68 | 4.79 | 6.16 | 11.68 | 12.84 | 10.26 | 10.6 | 11.67 | 12.3 | 2.13 |
| 19 | West Arthur | 48 | 23.8 | 46.1 | 59 | 46.1 | 35 | 23.06 | 21.5 | 16.33 | 16.2 | 16.35 | 18.13 | 10.75 | 8.17 | 6 | 5.47 | 4.6 | 6.13 | 12.31 | 13.33 | 10.33 | 10.7 | 11.75 | 12 | 2.31 |
| 20 | Wickepin | 41.3 | 31.5 | 58.7 | 71 | 49 | 35.9 | 23.52 | 22.1 | 16.65 | 16.4 | 16.82 | 19.07 | 12.07 | 9.11 | 6.25 | 5.5 | 4.68 | 5.88 | 11.45 | 13 | 10.4 | 10.9 | 12.14 | 13.2 | 0.98 |
| 21 | Williams | 44.7 | 29 | 52.7 | 67 | 48.1 | 31.9 | 23.41 | 21.5 | 16.55 | 16.5 | 16.61 | 18.25 | 10.98 | 8.08 | 5.86 | 5.41 | 4.14 | 5.74 | 12.43 | 13.46 | 10.69 | 11.1 | 12.47 | 12.5 | 3.02 |
| 22 | Woodanilling | 44.8 | 27 | 50 | 66 | 49.2 | 33.3 | 22.84 | 21.8 | 16.3 | 16 | 16.46 | 18.44 | 11.31 | 8.86 | 6.39 | 5.87 | 5.04 | 6.41 | 11.53 | 12.89 | 9.91 | 10.1 | 11.42 | 12 | 2.41 |
| 23 | All Shires | 48.4 | 31.7 | 54.6 | 70 | 53.2 | 37.9 | 23.08 | 22.3 | 16.97 | 16.8 | 17.21 | 18.93 | 11.51 | 9.22 | 6.65 | 5.88 | 5.16 | 6.37 | 11.58 | 13.08 | 10.32 | 10.9 | 12.05 | 12.6 | 1.89 |

**Figure 6-1. A fully featured composite dataset snapshot for the yield/climate relationship**

rainfall (VHR) of over 600mm per annum, high rainfall (HR) of 500-600mm per annum and low rainfall (LR) of less than 500mm per annum.

The shires were also classified as high yield (HY) of over 45,000 tonnes per shire and low yield (LY) of less than 45,000 tonnes per shire per annum based on the wheat yield for 2003. Both the rainfall and yield categorisations were range based from the existing historical data. Accordingly, the HY shires were Dumbleyung, Esperance, Gnowangerup, Jerramungup, Katanning, Kent, Kulin, Lake Grace, Ravensthorpe and Wickepin. The LY shires were Boddington, Boyup Brook, Broomehill, Bridgetown-Greenbushes, Cranbrook, Kojonup, Manjimup, Narrogin, Tambellup, Wagin, West Arthur, Williams and Woodanilling. All the 23 shires are shown in Figure 6-2. The lowest wheat yield in the HY category was the shire of Broomehill (51849 tonnes) and the highest in the LY category was Kojonup (28960 tonnes).

### 6.8.2 Exploratory analysis of the rainfall

The rainfall analyses that followed involved a process of examination from macroscopic inspection to microscopic scrutiny as outlined at the start of Section 6.8. The selected three year rainfall graphs were first plotted in order to establish the rainfall trend across the three selected years. Figure 6-3 showed the stochastic average annual rainfall for the three selected years for each of the 23 shires including the outlier shires of Bridgetown-Greenbushes and Manjimup which were subsequently removed due to non-production of wheat.

The rainfall variation over the three years did reflect and confirm the overall classification of dry and wet years where the rainfall increased progressively across the years 2002, 2003 and 2005. The variation was also clearly visible

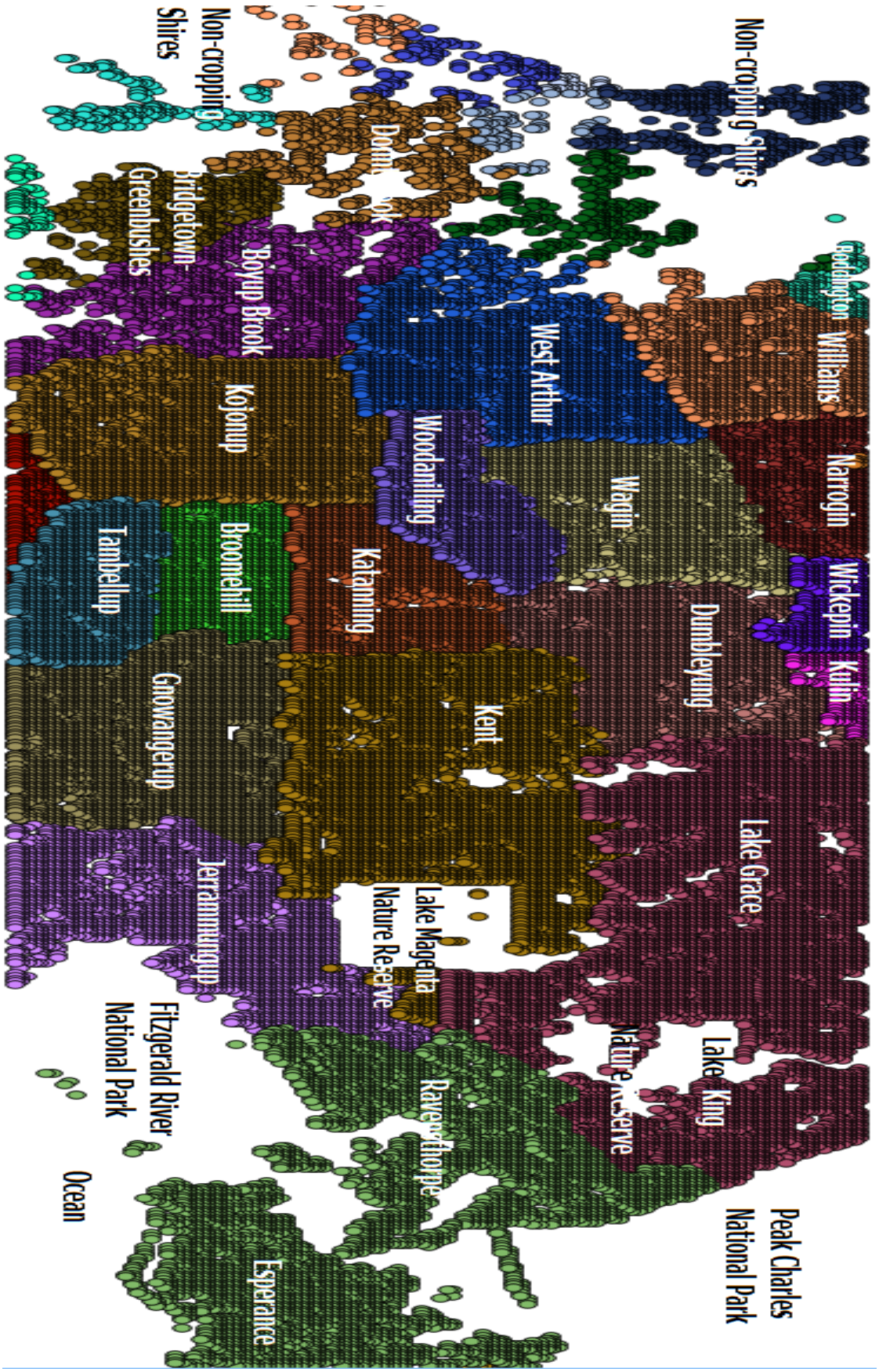



**Figure 6-2. QuantumGIS shires location map**

across the range of the different shires where the VHR shires were Boddington, Boyup Brook, Bridgetown-Greenbushes and Manjimup with the remaining HR shires with an overall high rainfall of 500mm per annum for the three years. The average annual rainfall in the HY shires was generally lower than for the LY shires across the selected years and was recognisable in the peaks of Boddington, Boyup Brook, Bridgetown-Greenbushes and Manjimup as per Figure 6-3. The biggest variation of over 500mm of rainfall occurred over the two shires of Manjimup (coastal) and Lake Grace (inland) when compared to others.

Following the precursory analysis of the rainfall within the different shires for the selected years the pre-cursory analysis of the wheat crops within the same matching shires and years was commenced.

### 6.8.3 Exploratory analysis of the wheat crop yields

The analysis of the individual crop yield across the selected years involved examining the wheat crop yields for the years 2002, 2003 and 2005 over the study area shires as shown in Figure 6-4.

The overall 2003 wheat crop yield was higher than that of the year 2005. The LY shires of Boyup Brook, Kojonup and West Arthur produced the highest crop yields between 4-5 tonnes/hectare for the year 2003. Bridgetown-Greenbushes had a zero yield due to no areas shown, and as a consequence it was eliminated from the ensuing analyses. The overall wheat crop yield trend was higher for the year 2003 and it defied the increased rainfall trend from 2003 to 2005. Only the LY shires of Boddington and Narrogin, and the HY shire of Ravensthorpe matched the trend with higher wheat crop yields in the year 2005.

The overall 2003 wheat crop yield was higher than that of the year 2005. The LY shires of Boyup Brook, Kojonup and West Arthur produced the highest crop yields between 4-5 tonnes/hectare for the year 2003. Bridgetown-Greenbushes had a zero yield due to no areas shown, and as

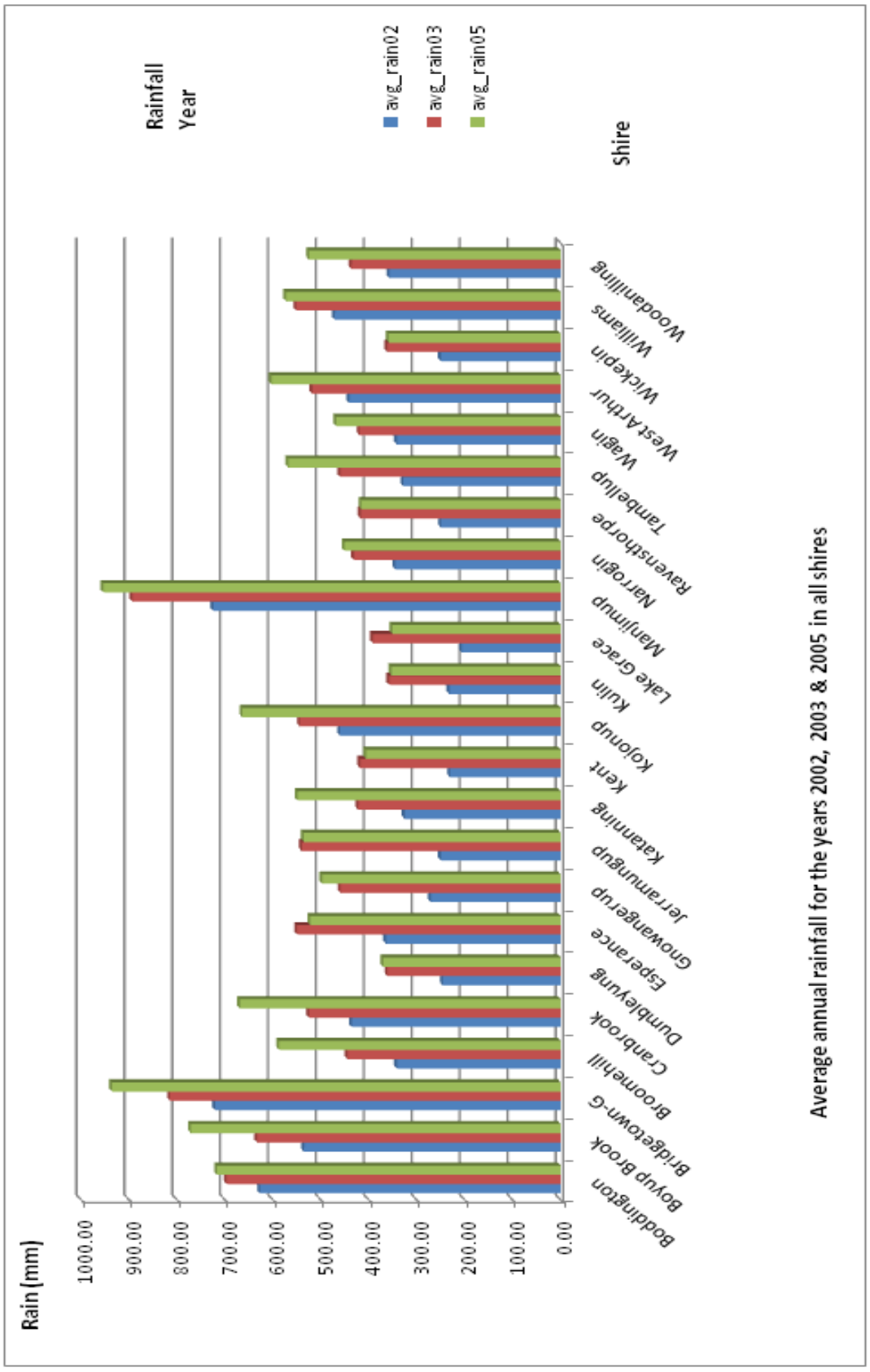



**Figure 6-3. Average annual rainfall for 2002, 2003 & 2005**

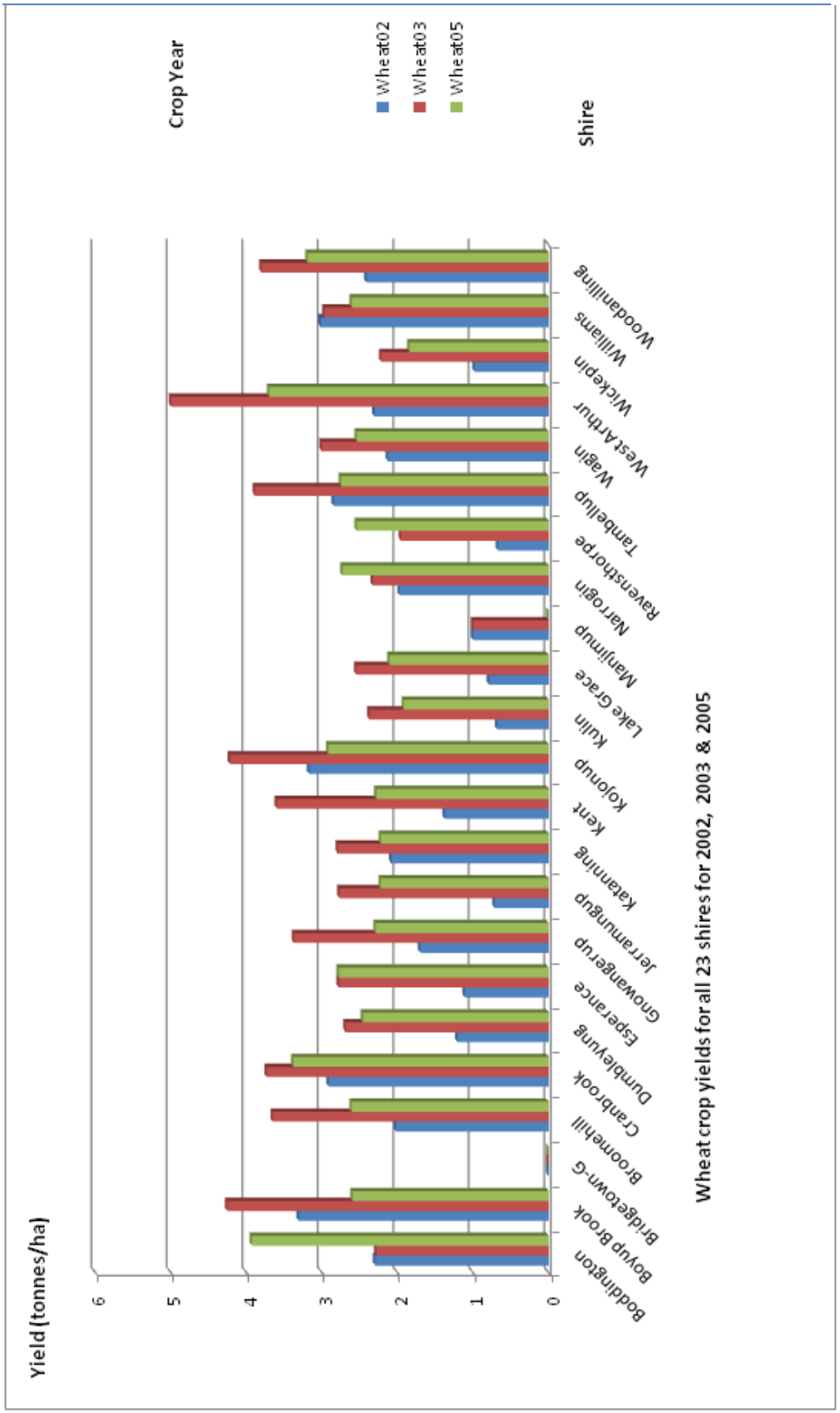



**Figure 6-4. Wheat crop yield for the years 2002, 2003 & 2005**

a consequence it was eliminated from the ensuing analyses. The overall wheat crop yield trend was higher for the year 2003 and it defied the increased rainfall trend from 2003 to 2005. Only the LY shires of Boddington and Narrogin, and the HY shire of Ravensthorpe matched the trend with higher wheat crop yields in the year 2005.

### 6.8.4 Correlation analysis of the wheat crop yields

The next part of the analysis was the in-depth scrutiny of the microscopic phase which involved the examination of the wheat crop yield for patterns and prediction accuracy. The aim was to find a serial correlation of dependance (Gow, Ormazabal, & Taylor, 2010) between rainfall and the wheat yield for the matching areas in both the HY and LY shires. This was done through a process of aggregation of the wheat crop yields for the selected three years in order to determine the time sequence as well as for prediction of crop yields using data mining. Normalisation was performed on the two differing scale attributes of rainfall and wheat in order to facilitate a simultaneous sequence chart plot in SPSS as shown in Figure 6-5. The SPSS software routine does not display all the shires in order to fit the whole sequence. However, as the all the shires are alphabetically ordered, the missing names are easily identifiable.

The visual comparison of the graph enabled a year by year comparative analysis which showed the correlation between the average annual stochastic rainfall and the wheat yield across the three selected years. The overall repetitive cycle for the three years was visible for both lines. It also displayed the increased rainfall trend from 2002 and 2003 to 2005. The crop yield across the three years was also mirrored, showing the increased crop yields over the relevant shires.

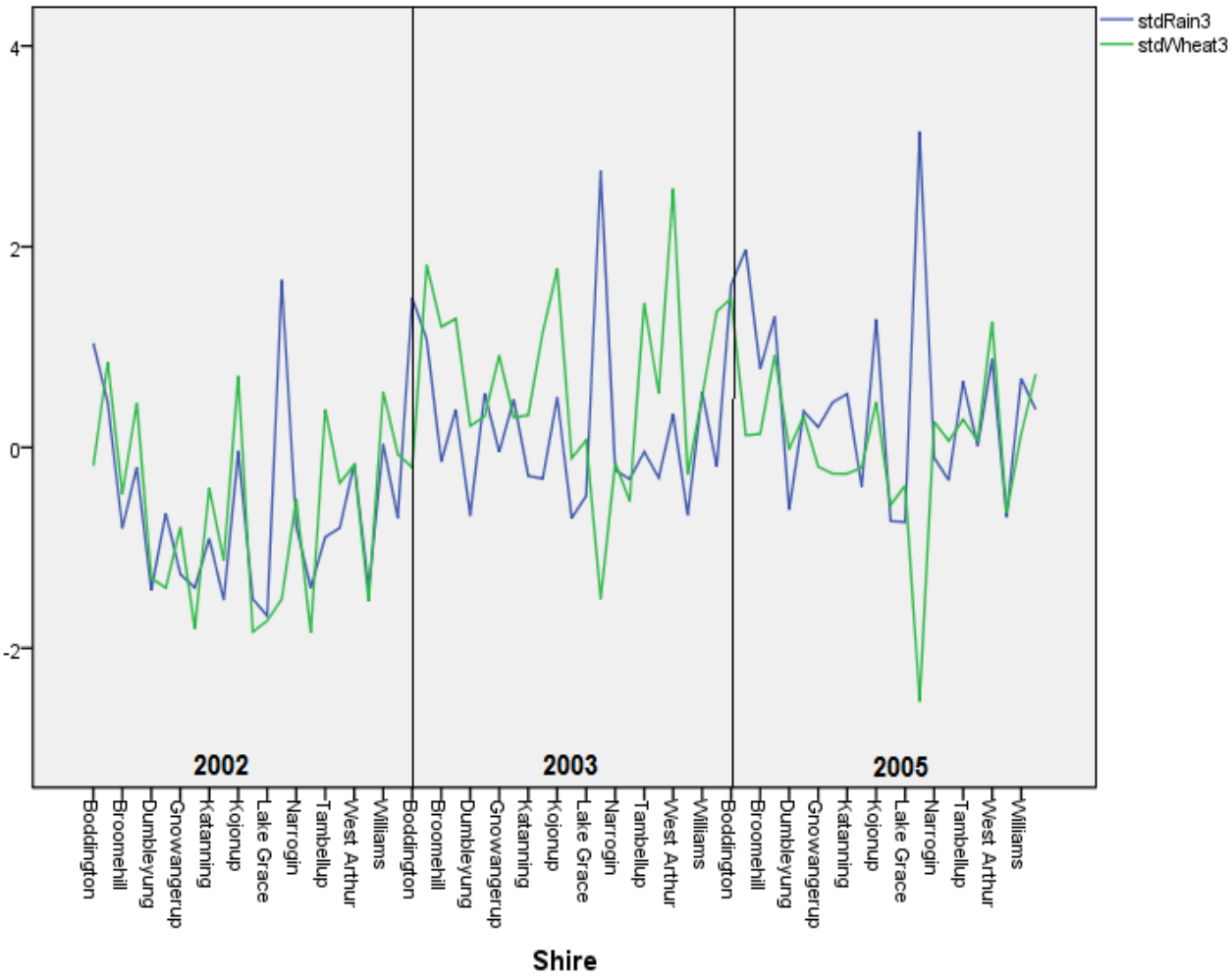



**Figure 6-5. Sequence plot of standardised rainfall and wheat for the years 2002, 2003 & 2005**

The negative impact on wheat crop yields was also demonstrated for the LY shire of Narrogin where the rainfall patterns were evident. The correlation coefficient for wheat yield and annual rainfall in the HY shires for the dry year 2002 was 0.49, 0.32 for the productive year, and 0.45 for the wet year 2005. The correlation was positive for the year 2002 and improved progressively for the years 2003 and 2005. The overall correlation coefficient for the three years taken together was 0.75 for the HY shires. These results are displayed in Figure 6-6.

| A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|
| shire | avg_rain02 | Wheat02 | cor02 | avg_rain03 | Wheat03 | cor03 | avg_rain05 | Wheat05 | cor05 | Cor020305 |
| Dumbleyung | 246.43 | 1.21 | 0.49 | 361.52 | 2.69 | 0.32 | 370.55 | 2.46 | 0.45 | 0.75 |
| Esperance | 364.98 | 1.11 | | 550.37 | 2.78 | | 522.67 | 2.78 | | |
| Gnowangerup | 271.09 | 1.70 | | 459.78 | 3.37 | | 497.87 | 2.29 | | |
| Jerramungup | 250.40 | 0.71 | | 540.23 | 2.77 | | 536.29 | 2.22 | | |
| Katanning | 325.93 | 2.08 | | 422.59 | 2.79 | | 549.24 | 2.22 | | |
| Kent | 231.32 | 1.38 | | 418.47 | 3.60 | | 406.42 | 2.28 | | |
| Kulin | 232.91 | 0.68 | | 358.00 | 2.37 | | 353.25 | 1.92 | | |
| Lake Grace | 206.67 | 0.79 | | 392.03 | 2.55 | | 351.35 | 2.11 | | |
| Ravensthorpe | 250.13 | 0.67 | | 417.69 | 1.96 | | 416.14 | 2.54 | | |
| Wickepin | 250.33 | 0.98 | | 362.10 | 2.21 | | 359.02 | 1.84 | | |

**Figure 6-6. Correlation coefficients between actual rainfall and wheat yield in the HY shires**

On the other hand, the correlation coefficient for the LY shires after the shire of Manjimup was excluded as an outlier, was -0.55 for the dry year 2002, -0.69 for the productive year 2003 and -0.69 for the wet year 2005. The overall correlation coefficient was -0.50 for the three years taken together. This indicted that the overall positive correlation for the HY shires was better than the for the LY shires. It also indicated that the LY shires tended to be negative for the three years. This suggested that an increase in rainfall would invariably result in an increase in crop yield across the selected shires, but the increase in wheat yield diminishes as the rainfall increases in some shires. The correlation results are shown in Figure 6-7.

| A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|
| shire | avg_rain02 | Wheat02 | Cor02 | avg_rain03 | Wheat03 | Cor03 | avg_rain05 | Wheat05 | Cor05 | Cor020305 |
| Boddington | 627.54 | 2.30 | -0.55 | 697.32 | 2.28 | -0.69 | 717.34 | 3.93 | -0.69 | -0.50 |
| Boyup Brook | 535.81 | 3.31 | | 633.52 | 4.26 | | 771.86 | 2.60 | | |
| Bridgetown-G | 721.48 | 0.00 | | 815.15 | 0.00 | | 936.39 | 0.00 | | |
| Broomehill | 341.86 | 2.02 | | 444.70 | 3.65 | | 587.76 | 2.61 | | |
| Cranbrook | 436.11 | 2.91 | | 525.04 | 3.73 | | 669.38 | 3.38 | | |
| Kojonup | 460.62 | 3.17 | | 544.62 | 4.22 | | 664.47 | 2.92 | | |
| Manjimup | 725.23 | 1.00 | | 894.06 | 1.00 | | 954.38 | 0.00 | | |
| Narrogin | 345.29 | 1.97 | | 431.83 | 2.33 | | 450.95 | 2.73 | | |
| Tambellup | 328.50 | 2.85 | | 460.07 | 3.89 | | 568.81 | 2.75 | | |
| Wagin | 342.35 | 2.13 | | 419.89 | 3.00 | | 468.58 | 2.54 | | |
| West Arthur | 441.44 | 2.31 | | 518.29 | 5.00 | | 603.42 | 3.70 | | |
| Williams | 471.67 | 3.02 | | 552.21 | 2.97 | | 573.13 | 2.61 | | |
| Woodanilling | 357.71 | 2.41 | | 436.75 | 3.80 | | 525.34 | 3.19 | | |

**Figure 6-7. Correlation coefficients between actual rainfall and wheat yield in the LY shires**

As an extension of this activity the standardised data of the 23 wheat production shires was separated into the HY shires and the LY shires. These were then plotted on a graph for further clarification of the correlation.

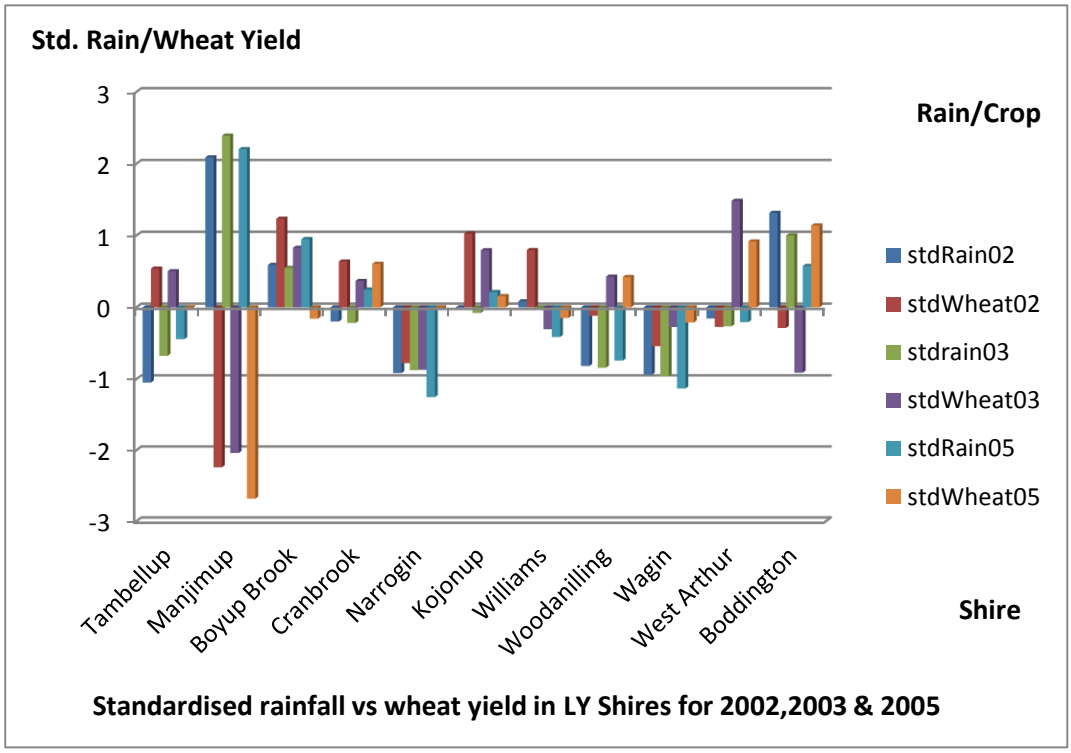



**Figure 6-9. Standardised rainfall/yield graph for the LY shires**

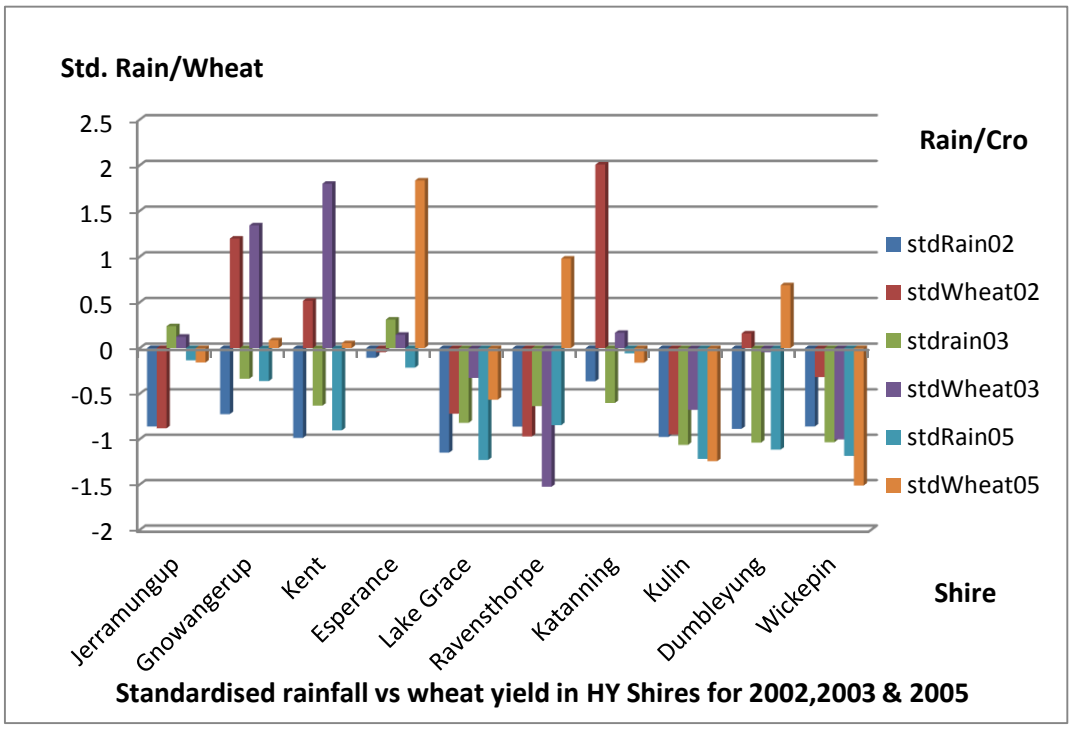



**Figure 6-8. Standardised rainfall/yield graph for the HY shires**

Figure 6-8 and 6-9 showed the standardised wheat crop yield and rainfall graph for the HY and LY shires respectively. Most of the HY shires in Figure 6-8 showed a positive correlation between rainfall and wheat yield, thereby demonstrating that a decrease in rainfall generally resulted in a decrease in wheat yield. There were some exceptions to this trend such as the shires of Gnowangerup and Kent which showed negative correlations for the year 2002 and 2003 and the shire of Katanning showing a negative correlation for the year 2002. The shires of Esperance, Ravensthorpe and Dumbleyung showed negative correlations for the year 2005. All of these exceptions resulted in high wheat yields from low rainfall.

With reference to the LY shires in Figure 6-9, the shire of Manjimup was considered to be an outlier due to unknown delivery area tonnage and wheat yield. The two LY shires of Narrogin and Wagin showed positive correlations with low rainfall matching the low wheat crop yields. The shire of Tambellup and Woodanilling showed a negative correlation but with a high wheat crop yield from low rainfall especially for the years 2002 and 2003. The overall trend for the LY shires was a high yield from low rainfall precipation. These trends were evident in Figure 6-9.

### 6.8.5 DM analysis of the wheat crop yields

The next step in the individual scrutiny of the exercise was the use of regression in order to determine if the relationship established through correlation could be supported by a mechanism of predicting the wheat crop yield through the rainfall. This was carried out using the classification technique of data mining in the Waikato Environment for Knowledge Analysis (WEKA) software. The aggregated data for average annual rainfall and wheat crop yield for the 22 shires were used for this activity. The aggregated wheat crop yield and rainfall dataset was split up into a training set (2001, 2002, 2004, 2006 data) and a test set (2003 & 2005 data). The split was designed to intersperse the 2002 low rainfall data amongst the other non-remarkable production/rainfall years of 2001, 2004

and 2006. This was also done in order that the 2003 (high yield) year and 2005 (high rainfall) years could be matched for predicted and actual values.The exploratory part of the DM activity was to use the training set to determine the best-fit algorithm using a simple model of crop yield as a function of the location class and the average annual rainfall.

### 6.8.6 Classification algorithms and comparisons

All of the classification algorithms within WEKA were tested in this step and a short-list of six algorithms was selected. These algorithms were Gaussian Processes (GP), Multilayer Perceptron (MLP), Radial Based Function Network (RBF), Kstar, Sequential Minimal Optimisation (SMO) and Additive Regresssion (AR). All these algorithms use regression for predicting continous values in response to input values. GP is a form of regression where the distribution is over mean and covariance functions without hyper parameter tuning for the classifier function (Rasmussen, 2004); MLP is a feed forward multi-layer Artificial Neural Network (ANN) function approximator classifier that uses the supervised learning technique of back propogation to classify instances (Nazzal, El-Emary, & Najim, 2008); RBF network is a neural network model function used for pattern classification (Fu & Wang, 2005); SMO uses the support vector machine for its regression by quadratically scaling the number of training patterns (L. J. Cao et al., 2006); the lazy Kstar algorithm is an instance based classifier that classifies a test instance based on its similarity to the training instance and AR is a meta classifier that seeks to enhance the performance of the regression based classifier (Witten et al., 2011). Each of the algorithms trialled had different characteristics, correlation co-efficients and Root Mean Square Errors (RMSE) as shown in Table 6-1.

**TABLE 6-1. WEKA ALGORITHMS RESULTS FROM THE TRAINING DATASET**

| WEKA ALGORITHM | CORREL COEFF TRAINING | RMSE TRAINING SET | RMSE CROSS VALID | RMSE TEST SET |
|---|---|---|---|---|
| Gaussian Processes | 0.998 | 0.411 | 0.718 | 1.057 |
| MLP | 0.999 | 0.074 | 0.628 | 0.952 |
| RBF Network | 0.754 | 0.549 | 0.619 | 0.794 |
| SMOreg | 1.000 | 0.003 | 0.659 | 0.897 |
| Kstar | 0.998 | 0.059 | 0.474 | 0.735 |
| Additive Regression | 0.987 | 0.141 | 0.529 | 0.749 |

GP, MLP and SMO were ruled out in the first instance due to the high RMSEs for the cross validation results. The GP algorithm had the second lowest RMSE for the predictions on the test data of 0.75, and a correlation co-efficient of 0.99. Based on the results from Table 6-2, together with a good cross validation result of 0.53, the GP algorithm was selected and run for the prediction phase of the DM activity.

Although no outlier deletion or fixed adjustment was used, optimisation of the GP algorithm calculations in the WEKA software was done using a cross validation. This was done for each of the years 2003 and 2005 to supplement the test set results. The prediction results for years 2003 and 2005 were displayed side by side in Table 6-2. The shires in both tables were split up into HY (unshaded) and LY (shaded) shires.

Based on a graded scale, good predictions were considered to have a percentage error of less than 20%, average predictions a percentage error of 21-40% and weak predictions a percentage error of over 40%. This was a range based classification formulated from the WEKA predictions**.** Accordingly, the HY shires of Dumbleyung, Esperance, Gnowangerup, Katanning, Kent, Lake Grace and Ravensthorpe all had good predictions. The remaining HY shires of Jerramungup, Kulin, and Wickepin had average predictions. There were no weak predictions in the HY shires for the year 2003. The LY shires of Boddington, Broomehill, Cranbrook, Tambellup**,** Wagin and Williams had good predictions for the year 2003,

**TABLE 6-2. THE WHEAT YIELD GP/MLP RESULTS IN WEKA FOR RAINFALL**

| HIGH YIELD YEAR 2003 | | | | HIGH RAINFALL YEAR 2005 | | |
|---|---|---|---|---|---|---|
| **Rural Shire HY+ LY** | **Actual Yield ton/ha** | **Pred. Yield ton/ha** | **% Error** | **Actual Yield ton/ha** | **Pred. Yield ton/ha** | **% Error** |
| Dumbleyung | 2.69 | 2.98 | 10.86 | 2.46 | 2.61 | 6.18 |
| Esperance | 2.78 | 3.14 | 12.95 | 2.68 | 2.47 | 7.72 |
| Gnowangerup | 3.37 | 3.43 | 1.66 | 2.29 | 2.66 | 16.24 |
| Jerramungup | 2.77 | 3.46 | 24.87 | 2.54 | 2.50 | 1.61 |
| Katanning | 2.79 | 3.31 | 18.46 | 2.22 | 2.62 | 18.15 |
| Kent | 3.60 | 3.21 | 10.75 | 2.28 | 2.49 | 9.39 |
| Kulin | 2.37 | 2.86 | 20.84 | 1.92 | 2.14 | 11.25 |
| Lake Grace | 2.55 | 2.89 | 13.25 | 2.11 | 2.54 | 20.19 |
| Ravensthorpe | 1.96 | 1.89 | 3.52 | 2.54 | 2.62 | 3.11 |
| Wickepin | 2.21 | 2.90 | 31.40 | 1.84 | 2.15 | 17.07 |
| Boddington | 2.28 | 1.89 | 17.02 | 3.93 | 2.55 | 35.14 |
| Boyup Brook | 4.26 | 3.27 | 23.26 | 2.60 | 2.55 | 1.85 |
| Broomehill | 3.65 | 3.31 | 9.26 | 2.61 | 2.54 | 2.57 |
| Cranbrook | 3.73 | 3.36 | 9.89 | 3.38 | 2.57 | 23.93 |
| Kojonup | 4.22 | 3.29 | 22.04 | 2.92 | 2.53 | 13.36 |
| Narrogin | 2.33 | 2.94 | 26.14 | 2.22 | 2.52 | 13.69 |
| Tambellup | 3.89 | 3.39 | 12.83 | 2.75 | 2.63 | 4.47 |
| Wagin | 3.00 | 2.94 | 1.87 | 2.54 | 2.62 | 3.11 |
| West Arthur | 5.00 | 2.78 | 44.32 | 3.70 | 2.51 | 32.19 |
| Williams | 2.97 | 3.00 | 0.98 | 2.61 | 2.55 | 2.38 |
| Woodanilling | 3.80 | 2.95 | 22.47 | 3.19 | 2.49 | 21.82 |
| All Shires | 3.15 | 3.09 | 1.97 | 2.62 | 2.50 | 4.47 |

while Boyup Brook, Kojonup, Narrogin and Woodanilling had average predictions with the only weak prediction for West Arthur. All the shires taken together as the All Shires location had a good yield prediction of 1.97% for the year 2003.

Analysis of the prediction results for 2005 in Table 6-2 were again used to determine the accuracy of the predictions. Overall there were more positive prediction errors for the year 2005. On the basis of the good, average and weak scale as outlined earlier, the HY shires of Dumbleyung, Esperance, Gnowangerup, Jerramungup, Kent, Katanning, Kulin, Ravensthorpe and

Wickepin had good predictions whilst only Lake Grace had an average prediction. On the other hand, the LY shires of Boyup Brook, Broomehill, Kojonup, Narrogin, Tambellup, Wagin and Williams all had good predictions, with average predictions for Boddington, Cranbrook, West Arthur and Woodanilling. There were no weak predictions at either the HY or LY shires for the year 2005. The *All Shires* location which represented the average of all the shires had a good prediction of 4.47 for the year 2005. Overall, the prediction errors were better in 2003 than in 2005 as depicted in Table 6-3. As predictions were better in the HY shires in the high rainfall year of 2005, it indicated a wheat yield link to the rainfall. The actual wheat yield and the corresponding predictions fell and rose respectively as the rainfall increased. However, these rainfall related fluctuations were not pronounced.

## 6.9 Discussion

In the establishment of a relationship between stochastic average annual rainfall and crop yield, a number of considerations had to be made. They included the following:

- Was rainfall a crucial factor in determining the final crop yield?
- Whether rainfall was affected by the physical location in terms of elevation and other climatic conditions such as wind and temperature;
- The geographic and climatic scaling and resolution; and,
- The effect of using interpolated rainfall and the use of the shire crop yield in tonnes/hectare.

Notwithstanding these effects and interactions, a pre-cursory relationship using a simple crop model was used where the classification entity was the rural shire and the average annual rainfall was used as a predictor. The selected production years of 2002, 2003 and 2005 were chosen for their considerable variation in both the attributes of rainfall and wheat yield. The selected study area had a number of intrinsic anomalies, such as shires

within the study area that were not designated for cropping and cereal production land uses as well as outlier crop yield shires which were eliminated from the analysis.

## 6.10 Conclusion

The actual wheat yields showed considerable variation between the 21 selected shires. As a consequence, it was concluded that rainfall may therefore not be such a decisive factor for wheat yield especially for the LY shires. The use of the data mining classification function of GP showed that the correlation between the stochastic average annual rainfall and wheat yield was a strongly positive one and that, as a result, generally wheat yield in the South West agricultural region could be expected to increase with an increase in rainfall, but there could be an increasing under-estimation error in predicting the wheat yields. The uncertainty of the prediction was thought to be related to the influence of other factors. These include the effect of the total seasonal rainfall as opposed to the rainfall for the months separately, as well as to the sparseness related to the intrinsic shire yield measurement of the dataset. This finding is similar to results of the NDVI approach used by Moriondo et.al (2007) and it differs to the spatial variability impact study on Australia wheat in as far as the causal effect of the El Nino effect on rainfall (Potgieter, Hammer, Meinke, Stone, & Goddard, 2005) was concerned. It was found that the performance of the WEKA algorithms could be enhanced by increasing the sample size of the crop yields from the three year selection to the 10 year selection for the years 2001-2010. Furthermore, a 10 year sample would allow for more years to be used as training datasets.

## 6.11 Chapter Review

The analysis carried out within this chapter was based on the earlier established metrics as defined in Section 6.8.1 and on the general two stage of macroscopic and microscopic scrutiny. The macroscopic analysis

involved the inspections of the rainfall and the individual wheat crop over the selected three years, as part of an EDM process. The simple bar graph visualization allowed the initial physical recognition of patterns and trends of wheat production as related to the stochastic average annual rainfall. The microscopic analysis involved scrutinizing the selected wheat crop yield for the selected years using DM and classification algorithms.

The results from the two stage analysis showed that there was a considerable correlation between stochastic average annual rainfall and wheat yield. The sequence chart of the standardized rainfall and wheat yield was useful in showing a gradual shift in rainfall and crop yield across the HY shires. Overall, the results did show a high positive correlation between stochastic rainfall and the wheat yield with some notable exceptions.

# Chapter 7

# THE EFFECTS OF TEMPERATURE ON CROP YIELD

This chapter represents the extension of the crop model from the previous chapter. The aim was to investigate the continuous variation relationships between temperature and the wheat crop yield. It formed a continuation of the series of qualitative and quantitative investigations carried out for the processing and analysis of both the geographic space and attribute space data. The qualitative aspect was the visual analysis of the graphs, while the quantitative aspect was the DM analyses. In this part of the work, the effect of the average stochastic monthly temperature was examined in detail for its impact on the crop yield in the wheat growing shires of the study area within the agricultural region of South Western Australia.

Similar to the case for the rainfall dataset, the temperature profile datasets were interpolated from recordings of fixed weather stations within the study area using ordinary krigeing and fitted onto a grid surface within the Revolution R environment. In this case however, the temperature was multi-faceted in that it was comprised of separate profiles of maximum, minimum and a calculated variation in average monthly temperature.

For this analysis, once again the grid surface area of the study area was similarly scaled down to a shire level in order to conform to the crop yield measurement scale. The monthly measurements were also scaled up to an annual level for the same reason. The evaluation was carried out using graphical, correlation and data mining regression techniques.

## 7.1 Introduction

Although crop production has been linked to a number of factors such as seasonal temperature, temperature variations, radiation, evaporation, soil moisture and crop management practices (Priya & R, 2001), this work

investigated the effect of temperature variation on crop production. It formed part of a continuation of the previous chapter covering the effect of rainfall on crop production (Y Vagh, 2012). The research design was similar to the one used in Chapter Six. The aim was to determine the relationship between the stochastic average monthly maximum and minimum temperatures, temperature variation and actual crop production at the shire level that was scaled to a grid cell with a resolution of 1000m by 1000m and to thereby justify the agricultural land use at these locations.

Consequently, this work was site-specific and also had a grid-cell spatial significance similar to the work done in Chapter Six. The research was carried out to predict the crop production at certain locations within the agricultural region, given the prevalent temperature conditions. It was established in Section 6.2, that factors of radiation, temperature and temperature variation all affect yield to some degree. In particular, the increased temperature variability was found by previous researchers to increase the crop yield variability of the plant (Trnka, Dubrovsky, Semera´dova, & Zalud, 2004). Winter wheat crop yield was found to be especially conducive to cold winters and hot summers (Wheeler et al., 2000). This was especially so when taken in conjunction with other management factors such as sowing time, stage of plant growth, fertilization and harvesting over the growing season of wheat from May to October in the agricultural region of South Western Australia (Priya & R, 2001).

In Chapter Six, spatial scales were established as important in their relation to the specific group or agency. As mentioned in Section 6.2, scales can be at the national level, regional level or farm management level. Similarly, this study investigated the serial correlations between these specific factors at the intermediary shire level. The temperature readings were aggregated from the fine resolution of 1000m by 1000m of a GIS grid cell.

In order to limit the analysis for the purposes of focus and scope as well as to highlight the relationships between high temperature variability and yield variability, the production years 2002, 2003 and 2005 were again selected.

The reasons for the selection of production years were similar to the case in Chapter Six. However, the low rainfall year of 2002 was equivalent to a year of high temperature and the high rainfall year of 2005 was equivalent to a year of low temperature. The selection therefore constituted a mix of varying conditions of temperature and yield as a test permutation for testing the effect of all facets of temperature on wheat crop yield.

## 7.2 Historical context

The effect of observed seasonal climatic conditions such as temperature variability on crop yield prediction has been undertaken by other researchers previously through an empirical crop model (Trnka et al., 2004). Interactions between input variables such as temperature variability and output variables such as crop yield have been shown to be important and have affected the yields statistically (Asseng, Foster, & Turner, 2011) . In particular, the relevance of changing temperatures was emphasized at critical phenological growth stages of a crop such as wheat (Wheeler et al., 2000). Previous research has also found that the yields of winter wheat are reduced when temperatures rise, due to the consequent reduction of the growth phases of the plant (Batts, Morison, Ellis, Hadley, & Wheeler, 1997; R. A. Brown & Rosenberg, 1997). Other studies have shown that temperature affected crops such as corn, soybeans and cotton in the United States where the yield either increased or decreased in response to temperature limits (Schlenker, Roberts, & Smith 2009). This research is similar to these precedents in its temperature-yield effect, but differs from them in that it examines the temperature in all its variability including maximum, minimum and variation between maximum and minimum temperature.

Based on the fundamentals of the two non-linear modelling approaches of Crop Models and General Circular Models previously established in Chapter Six, it was found that this investigation was also more suited to the CM approach. This is because the model does not feature the coupling

of the ocean and temperature effect due to its simulation problem as well as the focus on short-term weather forecasting as explained in Section 6.3. Crop model complexity is variable and dependant on the level of the analysis. They can be detailed (Brooks et al., 2001) or simply based on estimations of temperature (Martin et al., 2000). Models using other approaches such as the NDVI or a Probability Distribution Function (X. C. Zhang, Nearing, Garbrecht, & Steiner, 2004) have also been used. Sivakumar (2006) contends that there is always a need for refinement, even though there have been considerable improvements in understanding and predicting climate variability (Sivakumar, 2006) . This is largely due to the chaotic nature of the atmosphere and climate change phenomenon (Shah & Akhtar, 2011).

## 7.3 Dataset compilation - temperature

The dataset compilation done in this investigation was similar to the rainfall dataset compilation in Section 6.7. The study area was the same as outlined in Section 6.5. All the separate datasets were fitted specifically to the extraction region of the selected study area as was done for the investigation in Chapter Six. The difference was that the climate data was restricted to temperature in this part of the investigation to isolate its specific effect on crop yields.

Historical temperature data, just as for rainfall, existed only at sparsely located weather stations within the study area. In order to overcome the limitation of the associated sparse temperature dataset, a process of interpolation was carried out similar to the interpolation done for rainfall previously as detailed in Section 6.6. The pseudo-code for the generation of the interpolated data was the same except for the change in the climate variable from rainfall to minimum, maximum and variation in temperature. This interpolation resulted in temperature data points at each 1000m by 1000m (100ha) cell of the study area grid. The interpolation was carried out in the Revolution R script (Appendix A2) for temperature (maximum,

minimum, and variation) for each of the 12 months for the selected years of 2002, 2003 and 2005. This multi-faceted temperature profile was fitted onto the high resolution grid surface.

The production data only existed at an annual level for each shire within the study area. Consequently, the temperature data had to be aggregated using an R Script (Appendix A3) to an annual level to match the crop yield data for the years 2002, 2003 and 2005 only. Similar to what was done in Chapter Six, a process of data reduction was performed on the full study area dataset by only selecting land uses reserved for cropping and cereals, which effectively reduced the dataset to 53387 grid cells.

The temperature dataset compilation was similar to Section 6.7 for rainfall in that the structure was the same in terms of the production years and the wheat yield. It was different to the rainfall dataset as rainfall was a single variable whereas temperature was multi-faceted. The dataset was sorted into shire name order and the temperature for each of the cells within each shire was averaged by the number of cells within each shire. This process was carried out using a Revolution R script and resulted in a single line for the shire, maximum temperature, minimum temperature, temperature variation and crop yield. This dataset was then analysed for the wheat crop yield in relation to the temperature across the shires.

## 7.4 Experiments and analysis

There were a number of aspects to the data handling and analyses. They included the pre-processing (Section 7.4.1), the analysis of the maximum, minimum and temperature variation (Section 7.4.2), and the analysis of the individual wheat crop yields (Section 7.4.3), which formed the macroscopic phase. This was followed by the DM analysis of the individual wheat crop yields (Section 7.4.4) which formed the microscopic phase.

### 7.4.1 Pre-processing

The data used for the analysis was scaled down to match the crop yield data at the rural shire level. The QuantumGIS software was used to visualise the sizes and locations of the 21 shires as shown originally in Figure 6-2. Crop production data was normalised to tonnes/hectare for each shire. The shire categorisations used for subsequent analyses was the same as established in Section 6.8.1.

The maximum, minimum and variation in temperatures were also averaged for all the shires as a mean. Accordingly, temperatures over the mean were considered high; temperatures below the mean were considered low; and any temperatures within a degree Celcius of the mean were considered to be medium.

### 7.4.2 Temperature variation analysis

The first stage in the analysis was the general visual inspection of the data which was part of the exploratory data mining process. The temperature variation for the crop growing season months of April through to September for the 21 shires in the agricultural region of the South West of Western Australia was examined for the three selected years of 2002 for high temperature, the high crop yield year of 2003 and the low temperature years of 2005. Production years with high rainfall such as 2005 were instances of low temperature. Conversely, production years with low rainfall such as 2003 were instances with high temperatures.

The highest temperature variations in 2002 were for the month of May with an average across all the shires at 13.0 degrees, whilst the month of June showed the lowest temperature variations with an average of 10.3 degrees. The April temperature variation was moderate at 11.5 degrees and was similar to the season average temperature variation 11.7 degrees. The transition in the temperature variation from low to moderate and to high therefore occur in a time sequence from April at the start of the

growing season progressively to the month of May, prior to the big decreases in temperature in June and July. The July temperature variation was also low at 10.9 degrees. The temperature variations for the months of August and September were high at 12.0 and 12.5 degrees respectively. These observations were summarized in Table 7-1 for the low rainfall year of 2002. The seasonal temperature in Figure 7-1 was the average temperature for the growing season months.

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| Year | shire | April | May | June | July | August | Septembe | Seasonal | Wheat |
| 2002 | All Shires | 11.53 | 13.05 | 10.28 | 10.89 | 12.00 | 12.54 | 11.72 | 1.89 |
| 2003 | All Shires | 11.61 | 11.92 | 10.97 | 10.89 | 10.36 | 11.35 | 11.18 | 3.15 |
| 2005 | All Shires | 12.80 | 9.71 | 8.45 | 11.37 | 11.05 | 11.63 | 10.84 | 2.65 |

**Figure 7-1. Temperature variations for the growing seasons of 2002, 2003 and 2005**

The temperature variations for 2003 were similar to the year 2002 with highest variation for the month of May with an average of 11.9 degrees whilst the month of August, rather than June, with the lowest temperature variation at an average of 10.4 degrees. The month of April had a moderate temperature variation with an average of 11.6 degrees. The seasonal variation was moderate with an average of 11.2 degrees. The temperature variation for July and August were low with an average of 10.9 and 10.4 degrees respectively, whilst the month of September experienced a high temperature variation with an average of 11.3 degrees. These observations were also summarized in Table 7-1 for the high crop yield year of 2003.

The temperature variation for 2005 was unlike both the years 2002 and 2003 with highest variation for the month of April with an average of 12.8 degrees. The month of June was similar to the year 2002 with the lowest temperature variation at an average of 8.5 degrees. The month of May rather than April as for 2002 and 2003, had a moderate temperature variation with average of 9.7 degrees. The seasonal variation was moderate with an average of 10.8 degrees. The temperature variation for

July and August were low to moderate with an average of 11.4 and 11.1 degrees respectively, whilst the month of September experienced a high temperature variation with an average of 11.6 degrees. The notable decreases in temperature variations were likely due to the cooling effect of the higher rainfall for the year 2005. These observations of the temperature variations for the high rainfall year of 2005 were again discretised and shown in Table 7-1.

**TABLE 7-1. THE THREE YEAR GROWING SEASON COMPARISON TEMP VARIATIONS APR – SEP YEAR 2002, 2003 & 2005**

| YEAR | APR | MAY | JUN | JUL | AUG | SEP | 6 MTHS |
|---|---|---|---|---|---|---|---|
| 2002 | Mod | High | Low | Low | High | High | Mod |
| 2003 | Mod | High | L-M | Low | Low | High | Mod |
| 2005 | High | Mod | Low | L-M | L-M | High | Mod |

The discretised temperature rating details in Table 7-1 facilitated the yearly comparison of the temperature variations across the growing season months for the three selected years. The significant observations from the table were that differences occurred across all combinations of the years 2002, 2003 and 2005. The temperature variation switched from high to moderate in 2005 as opposed to moderate to high in 2002 and 2003 for the months of April and May. In addition, the month of August was different in that the temperature variation changed from high in 2002 and a low in 2003 to a low-moderate (L-M) in 2005. This showed that the months with differences in temperature variations were April and May, the start of the growing season and August near the end of the growing season spectrum. This translated to a trend of rising temperatures at the start of the growing season (April and May), low temperatures in the middle (June, July and August) followed by high temperatures at the end of the growing season.

### 7.4.3 Analysis of the wheat crop yield

The next step in the investigation was to examine the individual wheat crop yield over the three selected years of 2002, 2003 and 2005 across all the shires in the study area of the agricultural growing region.

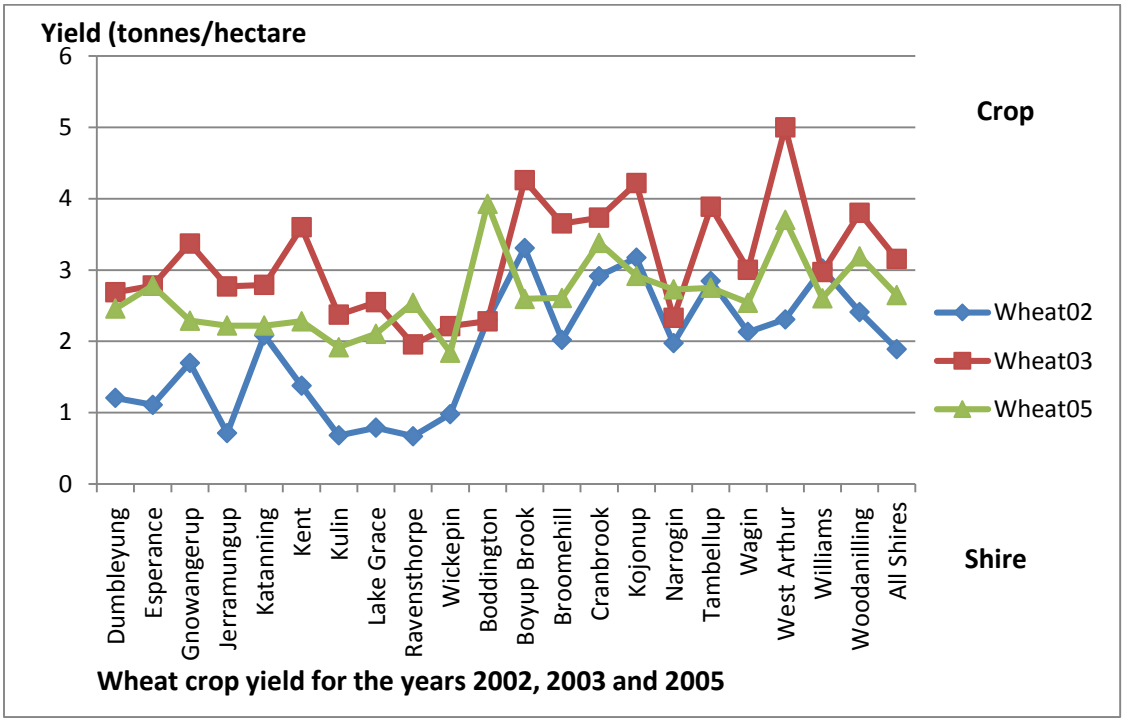



**Figure 7-2. Wheat crop yields for the years 2002, 2003 and 2005**

Figure 7-2 showed the crop yields in tonnes per hectare across all the shires. The highest yield was in the high crop yield (or productive) year of 2003, with the lowest yield in the low rainfall (or dry and hot) year of 2002. The moderate yield was in the high rainfall (or wet and cold) year of 2005. These demarcations were very distinct at the HY shires, but not so marked at the LY shires. The average of all the shires was the *All Shires* location which confirmed the initial observation. The mean wheat crop yield was 1.89 tonnes per hectare for the low rainfall year of 2002, 3.15 tonnes per hectare for the high crop yield year of 2003 and 2.65 tonnes per hectare for the high rainfall year of 2005. Consequently, any shire with higher than the mean was considered as having had a high yield, whereas a shire with

a crop yield of lower than the mean was considered to be low. Accordingly, Table 7-2 reflects these observations as translated from Figure 7-2.

**TABLE 7-2. ANNUAL WHEAT CROP YIELD RATING FOR 2002, 2003 & 2005 H=HIGH, L=LOW & M =MOD**

| HY SHIRE | 2002 | 2003 | 2005 | LY SHIRE | 2002 | 2003 | 2005 |
|---|---|---|---|---|---|---|---|
| Dumbleyung | L | L | M | Boddington | H | L | H |
| Esperance | L | L | H | Boyup Brook | H | H | L |
| Gnowangerup | M | H | M | Broomehill | H | H | L |
| Jerramungup | L | L | M | Cranbrook | H | H | H |
| Katanning | H | L | M | Kojonup | H | H | H |
| Kent | L | H | M | Narrogin | M | L | H |
| Kulin | L | L | L | Tambellup | H | H | H |
| Lake Grace | L | L | M | Wagin | H | M | M |
| Ravensthorpe | L | L | M | West Arthur | H | H | H |
| Wickepin | L | L | L | Williams | H | M | M |
| All Shires | M | H | H | Woodanilling | H | H | H |

It immediately became obvious that the HY shires were in fact producing less wheat per hectare than the LY shires. Normally, HY shires would be expected to have a greater crop yield. However, the LY shires in Table 7-2 had a higher preponderance of high (H) ratings. The main reason for this apparent discrepancy was that the HY shires produced more wheat due to the greater area of cultivation for those shires, thereby resulting in an overall increase in the measured crop yield. The HY rating was based on overall crop yield measurement. Most of the HY shires experienced an increase in crop yield when the temperature variation was high at the beginning of the growing season. These are obvious from the temperature variations in April and May in relation to the corresponding wheat yields for each of the HY shires as shown in Figure 7-3. The exceptions were the shires of Kent which experienced a drop in crop yield as opposed to Wickepin which seemed to be unaffected.

| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| shire | AprVari | MayVari | JunVari | JulVari | AugVari | SepVari | SeasonVari | Wheat |
| Dumbleyung | 11.39 | 12.85 | 10.25 | 10.83 | 11.99 | 13.03 | 11.72 | 1.21 |
| Esperance | 10.14 | 12.15 | 10.51 | 11.26 | 12.78 | 12.7 | 11.59 | 1.11 |
| Gnowangerup | 10.86 | 12.93 | 10.02 | 11.11 | 11.79 | 12.45 | 11.53 | 1.70 |
| Jerramungup | 11.38 | 14.1 | 11.53 | 12.9 | 13.53 | 13.76 | 12.87 | 0.71 |
| Katanning | 11.5 | 13.05 | 10.02 | 10.49 | 11.64 | 12.47 | 11.53 | 2.08 |
| Kent | 11.32 | 13.63 | 10.87 | 11.82 | 12.83 | 13.61 | 12.35 | 1.38 |
| Kulin | 11.44 | 13.09 | 10.51 | 11.07 | 12.41 | 13.55 | 12.01 | 0.68 |
| Lake Grace | 12.08 | 14.05 | 11.44 | 12.17 | 13.66 | 14.67 | 13.01 | 0.79 |
| Ravensthorpe | 11.48 | 13.57 | 11.4 | 12.66 | 13.84 | 14.09 | 12.84 | 0.67 |
| Wickepin | 11.45 | 13 | 10.4 | 10.9 | 12.14 | 13.19 | 11.85 | 0.98 |

**Figure 7-3. Temperature variation in the growing season months at the HY shires**

On the other hand, most of the LY shires experienced an increase in crop yield. The exceptions were the shires of Boddington, Narrogin, Boyup Brook and Broomehill, which experienced a drop in crop yield. These are apparent in Figure 7-4. These trends were relative to all the shires in each of the selected years of 2002, 2003 and 2005. Nevertheless, the analysis thus far was largely user driven visual inspection.

| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| shire | AprVari | MayVari | JunVari | JulVari | AugVari | SepVari | SeasonVari | Wheat |
| Boddington | 12.62 | 13.56 | 10.75 | 10.95 | 12.82 | 12.43 | 12.19 | 2.30 |
| Boyup Brook | 12.07 | 12.90 | 9.58 | 10.05 | 10.78 | 10.73 | 11.02 | 3.31 |
| Broomehill | 11.44 | 12.78 | 9.66 | 10.18 | 11.18 | 11.95 | 11.20 | 2.02 |
| Cranbrook | 11.38 | 12.41 | 9.32 | 9.94 | 10.77 | 11.00 | 10.80 | 2.91 |
| Kojonup | 11.74 | 12.75 | 9.52 | 9.98 | 10.95 | 11.16 | 11.02 | 3.17 |
| Narrogin | 11.65 | 12.88 | 10.35 | 10.84 | 11.79 | 12.63 | 11.69 | 1.97 |
| Tambellup | 11.20 | 12.48 | 9.43 | 10.11 | 10.92 | 11.59 | 10.96 | 2.85 |
| Wagin | 11.68 | 12.84 | 10.26 | 10.55 | 11.67 | 12.30 | 11.55 | 2.13 |
| West Arthur | 12.31 | 13.33 | 10.33 | 10.68 | 11.75 | 12.00 | 11.73 | 2.31 |
| Williams | 11.53 | 12.89 | 9.91 | 10.13 | 11.42 | 12.03 | 11.32 | 3.02 |
| Woodanilling | 11.53 | 12.89 | 9.91 | 10.13 | 11.42 | 12.03 | 11.32 | 2.41 |

**Figure 7-4. Temperature variation in the growing season months at the LY shires**

In order to bring the variables of temperature in all its facets, as well as crop yield together in a single analysis, more sophisticated and automated means of analysing the data was required. Consequently, DM was explicitly chosen for this purpose in order to reveal potentially hidden patterns and relationships that would otherwise be obfuscated by the multiplicity as well as the opposing nature of the variables. This transition in the analyses from coarse grain to fine grain scrutiny became part of the microscopic investigation which was a departure from the hitherto coarse grain macroscopic and visual inspection. This transition also included the facility for prediction from regression analyses as well as a correlation between the opposing variables. This was what was termed the process of refining the exploration where the analyses shifted in emphasis from observable trends in graphs and tables to predictive DM analysis.

### 7.4.4 DM analysis of the wheat yield

The next step in the individual scrutiny of the exercise was the use of regression in order to determine if the relationship established through simple observed dependence correlation could be supported by a mechanism of predicting the wheat crop yield. This was specifically for the variation in temperature across the growing months from April through September. This step was carried out using the classification technique of DM in the WEKA software.

The aggregated data for average monthly maximum, minimum and variation in temperature together with the wheat crop yields for the 23 shires was used for this activity. However, the shires of Bridgetown-Greenbushes and Manjimup were eliminated from the ensuing crop analyses as there were no recorded wheat crop yields for the three selected years. This reduction effectively brought the number of shires investigated to 21. In addition, an average of all the shires was added for the mean crop yield and temperature variables.

The first part of the DM was the analysis of the average maximum monthly temperature in relation to wheat yield. The average maximum monthly temperatures were taken across the growing season from the planting month of April to harvesting month of September. The shires were sorted into HY and LY groups. The data containing the aggregated wheat crop yield and the average monthly maximum temperature was split up into a training set and a test set. This splitting of the dataset into a training and test set was necessary for training the classification algorithms in the WEKA software suite. The trained set could then be used to classify a test set using the learning process derived from the training process. The training set comprised the data for the dry year 2002 and the test set was made up of the data for both the productive year 2003 and the wet year 2005. This training-test combination would ensure that the selection bias (Phillips, 2008) would be negative rather than positive and the corresponding results would be more conservative. The exploratory part of the DM activity was to use the training set to determine the best-fit algorithm using a simple model of crop yield as a function of the location class and the average maximum monthly temperatures. All of the classification algorithms within WEKA were tested in this step and a short-list of 10 algorithms was selected. These algorithms were Gaussian Processes (GP), Multilayer Perceptron (MP), Lazy LWL, RBF Network (RBF), M5 rules (M5R), Decision Stump (DS), M5P, Kstar, SMO and Additive Regresssion. All these algorithms use regression for predicting continuous values in response to input values. These results are summarised in Table 7-3.

Each of the algorithms trialled had different characteristics, correlation co-efficients and Root Mean Square Errors as shown in Table 7-3. The two main factors of correlation coefficient and the RMSEs were established as initial criteria for selection of the final algorithm upon which the test set would run.

**TABLE 7-3. THE WEKA ALGORITHMS RESULTS FROM THE 2002 TRAINING DATASET MAX TEMP APR – SEP**

| WEKA ALGORITHM | CORREL COEFF TRAINING | RMSE TRAINING SET | RMSE CROSS VALID | RMSE TEST SET |
|---|---|---|---|---|
| Gaussian Processes | 0.9984 | 0.4255 | 0.8402 | 0.8827 |
| MLP | 1.0000 | 0.0056 | 0.7620 | 1.1700 |
| RBF Network | 0.2762 | 0.8035 | 0.9023 | 0.9938 |
| SMOreg | 1.0000 | 0.0022 | 0.7730 | 1.4708 |
| Kstar | 1.0000 | 0.0002 | 0.9471 | 0.963 |
| Lazy LWL | 0.7818 | 0.5353 | 0.7453 | 0.8633 |
| Additive Regression | 0.9564 | 0.2444 | 0.8707 | 1.1316 |
| M5 Rules | 0.9969 | 0.0662 | 0.9159 | 1.2669 |
| Decision Stump | 0.6482 | 0.6366 | 0.7231 | 0.8711 |
| M5P | 0.9969 | 0.0662 | 0.9159 | 1.2669 |

The third criterion was the RMSE of the cross validation. The algorithm with the best performance in these three criteria turned out to be GP with a correlation of 0.9984, a RMSE for the training set of 0.4255 and a RMSE of 0.8402 for the cross validation run. The RMSE for the test set was 0.8827.

Consequently, the GP algorithm was run on the test set containing the productive year 2003 and the wet year 2005 data. No outlier deletion or adjustment was carried out as none of the outliers were considered to be any of the additive, innovational, temporary or level-shift types (L. M. Liu, Bhattacharyya, Sclove, Chen, & Lattyak, 2001). Instead optimisation was done by separating the test data into each year and cross-validation with 22 folds (Hall et al., 2009) was carried out. The results were displayed in Table 7-4. The shires in Table 7-4 were grouped into HY and LY (shaded blue) shires. The average of all the shires was denoted as the *All Shires* location and was unshaded.

On a discerning grade scale metric, good predictions were considered to have a percentage error of less than 20%, with average predictions having a percentage error of 21-40% and weak predictions with a percentage error of over 40%. Accordingly for 2003, the HY shires of Esperance,

**TABLE 7-4. THE WHEAT YIELD GP/MLP RESULTS IN WEKA MAXIMUM TEMPERATURE**

| HIGH YIELD YEAR 2003 | | | | HIGH RAINFALL YEAR 2005 | | |
|---|---|---|---|---|---|---|
| RURAL SHIRE HY+ LY | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR |
| Dumbleyung | 2.69 | 3.40 | 26.21 | 2.46 | 2.65 | 7.56 |
| Esperance | 2.78 | 2.72 | 2.23 | 2.78 | 2.54 | 5.30 |
| Gnowangerup | 3.37 | 3.35 | 0.53 | 2.29 | 2.75 | 20.09 |
| Jerramungup | 2.77 | 2.63 | 5.23 | 2.22 | 2.71 | 22.25 |
| Katanning | 2.79 | 3.29 | 18.06 | 2.22 | 2.66 | 19.62 |
| Kent | 3.60 | 2.94 | 18.36 | 2.28 | 2.62 | 14.91 |
| Kulin | 2.37 | 3.03 | 27.68 | 1.92 | 2.121 | 10.47 |
| Lake Grace | 2.55 | 2.92 | 14.35 | 2.11 | 2.177 | 3.18 |
| Ravensthorpe | 1.96 | 2.33 | 18.72 | 2.54 | 2.393 | 5.79 |
| Wickepin | 2.21 | 2.107 | 4.66 | 1.84 | 2.134 | 15.98 |
| Boddington | 2.28 | 2.293 | 0.57 | 3.93 | 2.52 | 35.80 |
| Boyup Brook | 4.26 | 3.54 | 17.00 | 2.60 | 2.91 | 11.77 |
| Broomehill | 3.65 | 3.36 | 7.95 | 2.61 | 2.85 | 9.20 |
| Cranbrook | 3.73 | 3.47 | 6.89 | 3.38 | 2.84 | 15.92 |
| Kojonup | 4.22 | 3.37 | 20.26 | 2.92 | 2.83 | 2.98 |
| Narrogin | 2.33 | 2.159 | 7.34 | 2.73 | 2.82 | 3.26 |
| Tambellup | 3.89 | 3.21 | 17.38 | 2.75 | 2.82 | 2.36 |
| Wagin | 3.00 | 3.04 | 1.27 | 2.54 | 2.19 | 13.78 |
| West Arthur | 5.00 | 3.01 | 39.84 | 3.70 | 2.70 | 27.05 |
| Williams | 2.97 | 3.32 | 11.85 | 2.61 | 2.80 | 7.36 |
| Woodanilling | 3.80 | 3.36 | 11.58 | 3.19 | 2.52 | 21.13 |
| All Shires | 3.15 | 3.07 | 2.41 | 2.65 | 2.70 | 2.00 |

Gnowangerup, Jerramungup, Katanning, Kent, Lake Grace, Ravensthorpe and Wickepin had good predictions. The remaining shires of Dumbleyung and Kulin had average predictions. No weak predictions were evident in the HY shires. Conversely, the LY shires of Boddington, Boyup Brook, Broomehill, Cranbrook, Narrogin, Tambellup, Wagin, Williams and Woodanilling had good predictions. The shires of Kojonup and West Arthur had average predictions. There were no shires with weak

predictions. Overall both HY and LY together had an average prediction of 2.41% for the year 2003.

Analysis of the results in Table 7-4 for 2005 were again used to determine the accuracy of the predictions. On the basis of the good, average and weak scale, the HY shires of Dumbleyung, Esperance, Katanning, Kent, Kulin, Lake Grace, Ravensthorpe and Wickepin had good predictions. Only the shires of Gnowangerup and Jerramungup had average predictions. On the other hand, the LY shires of Boyup Brook, Broomehill, Cranbrook, Kojonup, Narrogin, Tambellup, Wagin and Williams had good predictions while the shires of Boddington, West Arthur and Woodanilling had average predictions. There were no weak predictions for either the HY or the LY shires, with the overall prediction of 2% for the *All Shires* location.

Overall, the prediction errors relative to the average monthly maximum temperatures were better in 2005 than in 2003. The overall maximum monthly temperature decrease across all shires for the whole season was just 0.5 degrees. This situation indicated a link to the average maximum monthly temperature where the wheat yield rose as the temperature decreased. This was verified by the fact that the reduced temperature of 0.5 degrees in 2005 resulted in significant lower errors in prediction of wheat yield generally with an overall shift in prediction from a good prediction in 2003 (2.41%) to a better prediction in 2005 (2%).

This DM analysis method was repeated for the average monthly minimum temperatures for the months from April through to September. The results for both 2003 and 2005 were tabulated in Table 7-5.

Analysis of the results in Table 7-5 were used to determine the accuracy of the predictions for the productive year 2003 in relation to minimum temperature. On the good, average and low scale, the HY shires of Dumbleyung, Esperance, Gnowangerup, Jerramungup, Kent, Katanning, Kulin, Lake Grace and Ravensthorpe, while only Wickepin had an average prediction. On the other hand the LY shires of Boddington, Boyup Brook, Broomehill, Cranbrook, Kojonup, Tambellup, Wagin, Williams and Woodanilling had good predictions.

Only the LY shires of Narrogin and West Arthur had average predictions. There were no weak predictions for the LY shires. Overall though, all the shires together had good predictions of 3.46%.

The prediction results for the year 2005 in relation to minimum temperature were also depicted in Table 7-5. On the basis of the good, average and low scale, the HY shires of Dumbleyung, Esperance, Gnowangerup, Jerramungup, Katanning Kent, Lake Grace, and Ravensthorpe all had good predictions. Only the shires of Kulin and Wickepin had average

**TABLE 7-5. THE WHEAT YIELD GP/MLP RESULTS IN WEKA MINIMUM TEMPERATURE**

| HIGH YIELD YEAR 2003 | | | | HIGH RAINFALL YEAR 2005 | | |
|---|---|---|---|---|---|---|
| RURAL SHIRE HY+ LY | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR |
| Dumbleyung | 2.69 | 2.70 | 0.41 | 2.46 | 2.48 | 0.89 |
| Esperance | 2.78 | 2.97 | 6.80 | 2.78 | 2.38 | 14.28 |
| Gnowangerup | 3.37 | 3.20 | 5.16 | 2.29 | 2.43 | 6.20 |
| Jerramungup | 2.77 | 3.18 | 14.87 | 2.22 | 2.56 | 15.39 |
| Katanning | 2.79 | 3.23 | 15.54 | 2.22 | 2.66 | 19.86 |
| Kent | 3.60 | 2.98 | 17.29 | 2.28 | 2.43 | 6.66 |
| Kulin | 2.37 | 2.81 | 18.41 | 1.92 | 2.46 | 28.33 |
| Lake Grace | 2.55 | 3.04 | 19.23 | 2.11 | 2.35 | 11.39 |
| Ravensthorpe | 1.96 | 2.27 | 15.82 | 2.54 | 2.33 | 8.27 |
| Wickepin | 2.21 | 2.84 | 28.23 | 1.84 | 2.338 | 27.07 |
| Boddington | 2.28 | 2.503 | 9.78 | 3.93 | 2.87 | 27.10 |
| Boyup Brook | 4.26 | 3.66 | 14.14 | 2.6 | 3.176 | 22.15 |
| Broomehill | 3.65 | 3.08 | 15.66 | 2.61 | 2.70 | 3.56 |
| Cranbrook | 3.73 | 3.23 | 13.42 | 3.38 | 2.70 | 20.08 |
| Kojonup | 4.22 | 3.31 | 21.57 | 2.92 | 2.92 | 0.03 |
| Narrogin | 2.33 | 3.14 | 34.92 | 2.75 | 2.51 | 8.76 |
| Tambellup | 3.89 | 3.24 | 16.54 | 2.22 | 2.424 | 9.19 |
| Wagin | 3.00 | 3.08 | 2.63 | 2.54 | 2.76 | 8.50 |
| West Arthur | 5.00 | 3.33 | 33.41 | 3.70 | 2.90 | 21.56 |
| Williams | 2.97 | 3.04 | 2.42 | 2.61 | 2.75 | 5.53 |
| Woodanilling | 3.80 | 3.22 | 15.41 | 3.19 | 2.87 | 9.96 |
| All Shires | 3.15 | 3.26 | 3.46 | 2.62 | 2.476 | 5.50 |

predictions. On the other hand the LY shires of Broomehill, Kojonup, Tambellup, Narrogin, Wagin, Williams and Woodanilling all had good predictions whilst Boddington, Boyup Brook, Cranbrook and West Arthur had average predictions. There were no weak predictions for 2005 in any either the HY or the LY shires.

Taken together both the HY and LY shires had good predictions of 5.5%. Overall, the prediction errors relative to the average monthly maximum temperatures were better in 2003 than in 2005 as depicted in Table 7-5.

The overall drop in minimum temperature across all shires for the whole growing season was just 0.13 degrees between the years 2003 and 2005. This effectively translated to a slightly better prediction in 2003 by around 2%. This situation indicated a link to the average minimum monthly temperature where the predicted wheat yield rose marginally from a 3.46% error in prediction for the year 2003 to a 5.5 % error in prediction for the year 2005, as the minimum temperature decreased marginally.

The DM analysis method done previously for the maximum and minimum temperature was similarly applied for the average monthly temperature. The analysis once again featured the months from April through to September. The results for both the years 2003 and 2005 were tabulated in Table 7-6.

With reference to Table 7-6 and the year 2003, on the basis of the previously established good, average and weak scale, the HY shires of Dumbleyung, Gnowangerup, Jerramungup, Katanning, Lake Grace all had good predictions. The shires of Esperance, Kulin, Kent, Ravensthorpe and Wickepin had average predictions. On the other hand, only the LY shires of Boddington, Narrogin and West Arthur had a average predictions. The rest of the shires of Boyup Brook, Broomehill, Cranbrook, Kojonup, Wagin and Williams and Woodanilling had good predictions. There were no weak predictions for both the HY and the LY shires in 2003. Overall all the shires as a group of both HY and LY had good predictions with a prediction error of 4.2% for the high crop yield year of 2003.

On the basis of the good, average and weak prediction scale for the year 2005, the HY shires of Dumbleyung, Esperance, Gnowangerup, Jerramungup, Katanning, Kent, Lake Grace and Ravensthorpe all had good predictions, while only the shires of Kulin and Wickepin had average predictions. Conversely, the LY shires of Boyup Brook, Broomehill, Kojonup, Narrogin, Tambellup, Wagin, Williams and Woodanilling all had good predictions. The shires of Boddington, Cranbrook, and West Arthur had average predictions. There were no weak predictions in either the HY or the LY shires for both the years 2003 and 2005.

**TABLE 7-6. THE WHEAT YIELD GP/MLP RESULTS IN WEKA TEMPERATURE VARIATION**

| HIGH YIELD YEAR 2003 | | | | HIGH RAINFALL YEAR 2005 | | |
|---|---|---|---|---|---|---|
| RURAL SHIRE HY+ LY 2003 | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR | ACTUAL YIELD TON/HA | PRED. YIELD TON/HA | % ERROR |
| Dumbleyung | 2.69 | 3.16 | 17.47 | 2.46 | 2.48 | 0.81 |
| Esperance | 2.78 | 3.49 | 25.58 | 2.69 | 2.61 | 2.72 |
| Gnowangerup | 3.37 | 2.81 | 16.87 | 2.29 | 2.64 | 15.24 |
| Jerramungup | 2.77 | 2.42 | 12.56 | 2.22 | 2.64 | 19.05 |
| Katanning | 2.79 | 3.13 | 12.04 | 2.22 | 2.59 | 16.80 |
| Kent | 3.60 | 2.64 | 26.67 | 2.28 | 2.60 | 13.90 |
| Kulin | 2.37 | 2.85 | 20.04 | 1.92 | 2.47 | 28.80 |
| Lake Grace | 2.55 | 2.49 | 2.39 | 2.11 | 2.36 | 11.80 |
| Ravensthorpe | 1.96 | 2.63 | 34.13 | 2.54 | 2.56 | 0.67 |
| Wickepin | 2.21 | 2.69 | 21.63 | 1.84 | 2.51 | 36.14 |
| Boddington | 2.29 | 2.91 | 27.25 | 3.93 | 2.57 | 34.58 |
| Boyup Brook | 4.26 | 3.60 | 15.59 | 2.60 | 2.72 | 4.66 |
| Broomehill | 3.65 | 3.24 | 11.15 | 2.61 | 2.81 | 7.78 |
| Cranbrook | 3.73 | 3.66 | 1.90 | 3.38 | 2.59 | 23.46 |
| Kojonup | 4.22 | 3.41 | 19.10 | 2.92 | 2.83 | 2.98 |
| Narrogin | 2.33 | 3.02 | 29.90 | 2.22 | 2.64 | 19.05 |
| Tambellup | 3.89 | 3.51 | 9.87 | 2.75 | 2.65 | 3.71 |
| Wagin | 3.00 | 3.27 | 9.00 | 2.54 | 2.77 | 9.17 |
| West Arthur | 5.00 | 3.00 | 39.95 | 3.70 | 2.71 | 26.84 |
| Williams | 2.97 | 3.43 | 15.42 | 2.61 | 2.75 | 5.36 |
| Woodanilling | 3.80 | 3.43 | 9.68 | 3.19 | 2.61 | 18.34 |
| All Shires | 3.15 | 3.28 | 4.16 | 2.61 | 2.75 | 5.36 |

## 7.5 Discussion

The DM results did show a high positive dependence correlation between stochastic temperature and the wheat yield, but the situation was rather complex. For example, the crop yield prediction improved from a good prediction (2.4%) in 2003 to a *better* prediction (2.0%) in 2005 for the overall maximum temperature average decrease of just 0.5 degrees. This trend was in contrast with predictions of 3.3% in 2003 and 5.5% in 2005 in response to an overall decrease in minimum temperature of just 0.13 degrees. The contrast with the maximum temperature effect on yield was duplicated with predictions of 4.2% in 2003 and 5.4% in 2005 with a temperature variation decrease of just 0.34 degrees between the two years.

Taken together it was apparent that a marginal increase in temperature at the higher end (15-20 degrees) and middle (10-12 degrees) of the scale resulted in significant improvement in the wheat crop yield. This was in opposition to the low end (6-9 degrees) of the scale where a marginal increase in temperature resulted in a marginal improvement to the crop yield. The actual wheat yield predictions between the years 2003 and 2005 were significant in that they showed a considerable variation as noted for the year 2003 (2.4%, 3.5% and 4.2%) and 2005 (2.0%, 5.5% and 5.4%) for the maximum, minimum and variation in temperature.

In addition, the HY shires had better crop yields as well as better predictions overall due to the slightly higher temperatures prevalent in them. This amounted to a complex relationship between temperature and the wheat crop yield prediction where the crop yield could be expected to be better at higher temperatures and worse at lower temperatures generally, with exceptions for slight decreases at the higher and lower ends of the temperature scale.

## 7.6 Conclusion

The use of the data mining classification functions GP as well as MLP showed that the correlation between the stochastic average monthly temperature and wheat yield was a strongly positive one and that as a result generally, wheat yield in the South West agricultural region could be expected to increase with an increase in temperature. Lower temperatures tended to lower the prediction accuracy albeit marginally. Very high temperatures also impacted negatively on the wheat yields.

The findings of this research conform to the results of the effect of the variability of temperature on annual crop yields (Wheeler et al., 2000). It was also noted that there was a difference in how temperature variability and seasonal temperature affected the wheat yields differently. The use of the maximum temperature effect on the wheat yield was particularly important in that it helped fill the void in the research context (Wheeler et al., 2000). The results of this study was similar in nature to findings of Schlenker and Roberst, (2009) who also concluded that temperature affected certain crops (corn, soybeans, cotton) by both increasing and decreasing the yield depending on the temperature limit. It was different in that this study was done in Australia instead of the United States and the crop yield measured, was wheat instead of cotton, soybeans or corn. The predictions could be refined it the temperature effect was related to the influence of other factors, including the combined effect of rainfall and soil moisture retention. Therefore, a further extension to this study was to examine the dual effect of rainfall and temperature (climate) on wheat yield.

## 7.7 Chapter review

This chapter covered the examination of the isolated effect of all facets of temperature on the wheat crop yields at the shires within the study area.

This was in terms of the overall correlation and the temperature based predictability of the wheat yield.

The temperature was found to affect the wheat crop yield in a complex way where small increases in temperature resulted in gains in wheat yield, but large increases in temperature caused a reduction in the wheat yield. Wheat yield predictions were found to be better for the HY shires than the LY shires. They were also found to be more accurate when the temperature was lower due to higher rainfall in the high rainfall and low temperature year of 2005.

# Chapter 8

# THE EFFECTS OF RAINFALL AND TEMPERATURE ON CROP YIELD

This chapter examines the effect of both the continuous variation variables (rainfall and temperature) effect on the wheat crop yield at the crop growing shires within the selected study area. In essence, the chapter has the gestalt effect of bringing all the components to a climactic end, where the total effect is greater than the sum of the individual effects (Gold, Mundy, & Tjan, 2012).

## 8.1 Introduction

In this chapter the final crop model investigation within the series of qualitative and quantitative investigations is carried out for the processing and analysis of geographic land-use data in an agricultural context. The geographic data was made up of crop and cereal production land use profiles similar to the preceding investigations in Chapter Six and Chapter Seven. These were similarly linked to the climatic data from fixed weather stations in Australia interpolated using ordinary krigeing to fit a grid surface. Temperature and rainfall together constituted the climatic effect on the wheat crop yield.

There are a couple of main differences between this chapter and chapters six and seven. Firstly, the effect of the temperature and rainfall variables on wheat yield was examined simultaneously instead of separately. This constituted a lateral expansion to the data dimensions. Secondly, the temperature and rainfall were sampled for a selected decade of wheat crop production for the years from 2001 to 2010. The increase in sample years constituted the vertical expansion to the data dimensions. Effectively, the extension to the previous studies was the 10 year time-series investigation to the DM analysis. This was done in order to augment the previous work

done on the three selected years of 2002 (low rainfall/high-temperature), 2003 (high production) and 2005 (high rainfall/low-temperature). Therefore, this exercise expanded upon the original mix of varying conditions of rainfall, temperature and yield as a test permutation.

The evaluation was carried out using graphical, correlation and data mining regression techniques in order to detect the patterns of crop production in response to the climatic effect across the wheat cropping shires of the agricultural region of South Western Australia.

## 8.2 Historical context

Although crop production has been linked to a number of factors such as seasonal temperature, temperature variations, radiation, evaporation, soil moisture and crop management practices (Priya & R, 2001), this analysis investigated the effect of rainfall in conjunction with temperature variability on crop production. It formed part of a continuation of the previous work of the effect of rainfall (Chapter Six) and temperature (Chapter Seven) separately on crop production. The purpose was to determine the relationship between climate variables such as rainfall and temperature, taken together, on the actual wheat crop production at the shire level that was scaled to a grid cell resolution of 1000m. The research was carried out in order to further verify the agricultural land use at these locations as well as to provide the agricultural industries an insight for crop decision making with consideration to the climatic effect. Similar to the investigations done in Chapter Six and Chapter Seven, this work was site-specific together with having a grid-cell spatial significance.

According to Olesin and Bindi (2002) the factors of rainfall, radiation, temperature and temperature variation all affect yield to some degree with increased temperature variability especially increasing the crop yield variability of the plant (Trnka et al., 2004). The winter wheat crop yield is especially conducive to cold winters and hot summers when taken in

conjunction with other management factors such as rainfall, sowing time, stage of plant growth, fertilization and harvesting over the growing season of wheat from May to September in the agricultural region of South Western Australia (Priya & R, 2001). As mentioned previously in Section 6.2, adjustments and considerations relating to spatial scales were taken into account for the local government rural shire locations.

The rest of this chapter is organized as follows. Section 8.3 deals with *Related work* of a similar nature. Section 8.4 covers the *Data compilation* for the research. Section 8.5 deals with the *Experiments and analysis* which is further divided into four sub-sections. Sub-section 8.5.1 details the *Pre-processing.* Sub-section 8.5.2 deals with the *Analysis of climate variables*. Sub-section 8.5.3 relates to *the Analysis of the wheat crop yield.* Sub-section 8.5.4 presents the final *Data mining analysis*. Lastly, the chapter is rounded off with a *Conclusion* in Section 8.6.

## 8.3 Related work

The effect of seasonal climatic conditions such as observations of rainfall and temperature variability on crop yield prediction has been undertaken by researchers (Trnka et al., 2004) before as indicated in Section 7.2. The empirical model in that study showed that yields were affected statistically by the important interactions between temperature variability and crop yield. In particular, the relevance of changing temperatures was emphasized at critical phenological growth stages of a crop such as wheat (Wheeler et al., 2000). Previous research has found that the yields of winter wheat were reduced when temperatures rise, due to the consequent reduction of the growth phases of the plant (Batts et al., 1997; R. A. Brown & Rosenberg, 1997). Details of the two approaches for the development of

crop yield models, namely the Crop Model and the General Circular Model have been established in Section 6.3 previously.

Apart from the problem of the simulation of too many low intensity temperature instances within each grid cell (Ines & Hansen 2006) for GCMs, there were other potential problems. They included the low spatial resolution and the low precision of the simulation of the local current climate (Foulkes et al., 2011). Nevertheless, this study was therefore more suited to the CM approach due to the current emphasis on short term weather forecasting as outlined previously Section 6.3.

Furthermore, there are two approaches to investigating the impact of climate change on crop production which include the crop suitability approach and the production function approach (Foulkes et al., 2011). The crop suitability approach, also known as the Agro-Ecological Zoning (AEZ) approach, uses climate to determine crop suitability at an agricultural location using simulated rather than measured crop yields (Deressa, 2007). On the other hand, the production function approach uses either and empirical or experimental production function to measure the relationship between crop production and climate change (Bakewell & Garbutt, 2005).

The complexity issues of any crop model in relation to level of detail of the analysis as well as the other factors featuring the modeling approach have already been discussed in Section 7.2. Furthermore, previous studies that link the yield quality to temperature variability in as far as bread making for example, have been well documented (Gooding, 2003; R. Kimball & Ross,

2002; Simoff et al., 2008). It is within the backdrop of this research spectrum that the current study is situated.

## 8.4 Dataset compilation – rainfall and temperature

The study area was the same as outlined in Section 6.4. The difference was that the climate data included both rainfall and temperature in this part of the investigation. This was done in order to determine the main climatic effect on crop yields without obfuscation from other factors such as soil moisture retention and evaporation. Historical temperature and rainfall data were interpolated for the reasons and manner as mentioned in Section 6.5. The data selection and extraction procedure was therefore similar to those done separately for Chapters 6 and 7 respectively except for the two extensions to length (10 year selection) and breadth (rainfall and temperature). The annual production data was also taken for the decade from 2001 to 2010. Both the temperature and rainfall data were growing season monthly interpolations carried out using an R Script (Appendix A2) for the years 2001 - 2010.

The interpolation was carried out in the Revolution R script (Appendix A1) for both rainfall and temperature for each of the 12 months for the years 2001 to 2010. The temperature and rainfall profiles were fitted onto the high resolution grid surface. A second Revolution R script (Appendix A2) was used to extract the interpolated rainfall and temperature data from the 120 separate files for consolidation into a single file. This extraction, interpolation and aggregation procedure was done in four stages, for maximum temperature, minimum temperature, temperature variation and rainfall for each of the years in the selected decade. The consolidated temperature/rainfall data file was then trimmed to only have the data relevant to the agricultural land-uses of cropping and cereals associated

with yield data as was done previously in Chapters Six and Seven. This data reduction effectively scaled the climate dataset down in size.

The dataset was organized into a composite to produce a single line for the shire, rainfall, maximum temperature, minimum temperature, temperature variation and crop yield. The whole process is depicted in Figure 3-4 (not repeated here) and the spatial scaling progression is shown in Figure 8-1.

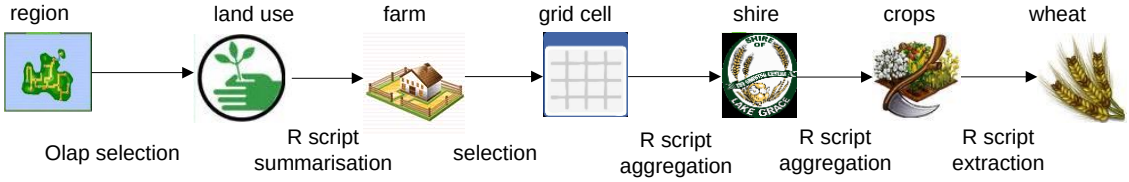



**Figure 8-1. Spatial scaling of the data to shire-grid-cell level**

## 8.5 Experiments and analysis

There were a number of aspects to the data handling and analyses. They included the pre-processing and associated metrics, the analysis of the rainfall, temperature and the individual wheat crop yields. This first aspect formed the macroscopic phase. The second aspect was the DM analysis of the individual wheat crop yields which formed the microscopic phase.

### 8.5.1 Pre-processing

The data used for the analysis was scaled down to match the crop yield data at the rural shire level. The top-down scaling was from the agricultural region, cropping areas and shires to the individual crop as shown in Figure 8-1. The final analysis was carried out on a shire level for the multiple attributes of stochastic average monthly rainfall, maximum, minimum and

temperature variation together with the annual wheat crop yield in each shire.

The study area within the South West Australian agricultural zone encompassed a number of shires as detailed in Section 6.4 previously. Crop production data existed for each shire on a total annual basis. As the shires within the selected study area were of different sizes, ranging from 1,409,900 hectares for Esperance to 5,800 hectares for Manjimup, a process of normalisation was required. This was calculated by dividing the tonnes delivered in a shire by the delivery area of the shire in order to determine the crop yields in tonnes/hectare. The rainfall, maximum, minimum and variation in temperature for the years 2001 to 2010 and the months from April to September were averaged for each shire using an R script (Appendix A3) individually. These were then collated with the wheat crop yield data for the same 10 year period in the database for each of the shires to produce one aggregated file in preparation for the DM analyses in WEKA.

The visual inspection of the graphs was again based on the uniform method of evaluation established in Section 6.8.1. All the shires were therefore classified in the same manner as was done for the previous two chapters.

### 8.5.2 Analysis of climate variables

The first stage in the analysis was the general visual inspection of the data which was part of the EDM process for the rainfall and temperature for the crop growing season months of April through to September for the 21 crop yielding shires in the agricultural region of the South West of Western Australia. The variations in average monthly rainfall were plotted for the selected years in the decade from 2001 – 2010 as data point time series charts which were depicted in Figure 8-2 (HY shires) and Figure 8-3 (LY shires). The fact that the rainfall was taken for the whole growing season

from April to September made it difficult to determine the different shires on the graph.

It was evident from the graphs that the rainfall patterns were similar for both the HY and LY shires generally. The exceptions were that the LY shires experienced a marginally higher rainfall as evidenced by the peak of Cranbrook of 160mm in June 2005. The HY shires of Jerramungup and Katanning were consistently high for most years. These shires were outperformed by Kent and Gnowangerup in 2001, 2004 and 2005. The HY shire of Gnowangerup received a lower rainfall of 138.53 mm (Figure 8-2) as opposed to the LY shire of Tambellup which received a higher rainfall of 160.05 mm (Figure 8-3) for June

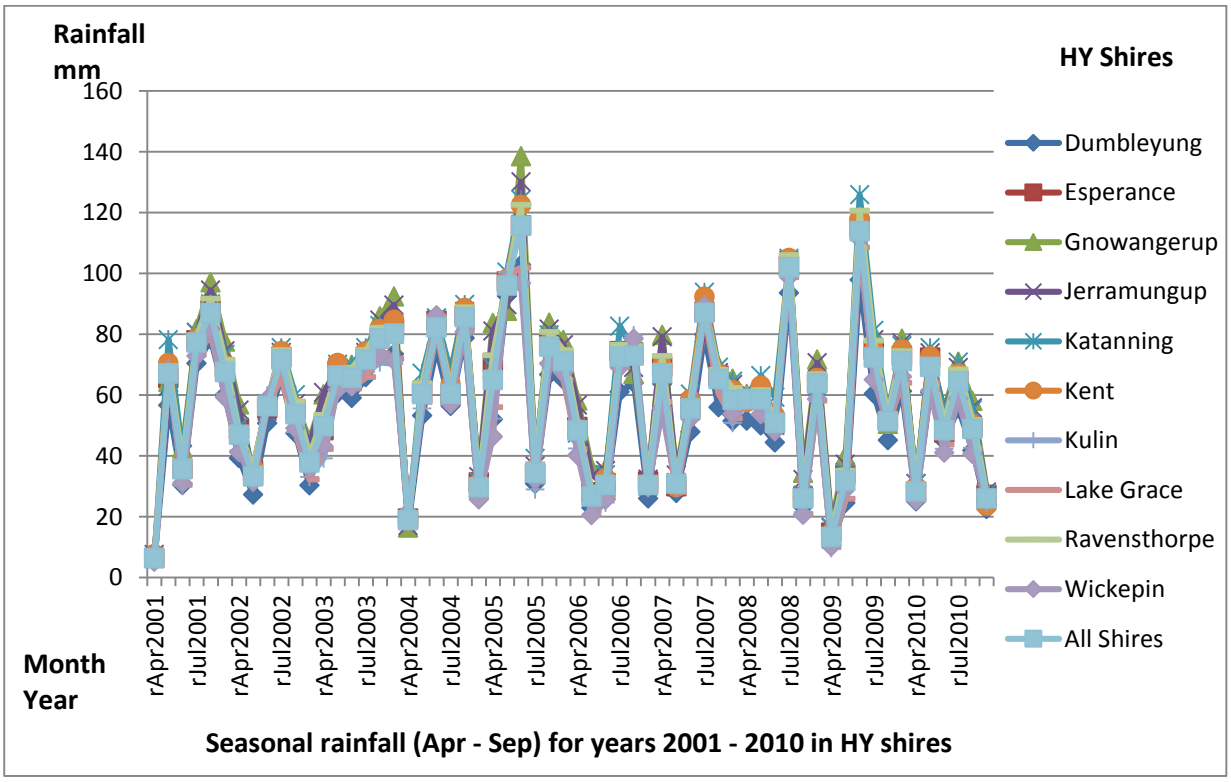



**Figure 8-2. Average monthly rainfall for the HY shires in the 2001-2010 decade**

2005. In addition, the steep rise to high rainfalls in June and July and sharp declines to September were very pronounced for 2001, 2002, 2005, 2008 and 2009.

The rainfall patterns of 2001, 2003 and 2005 were prominent especially with the high rainfall year of 2003 where rainfall did not feature in the August and September months. The peaks were observed at the shire of Cranbrook for June 2005. The rainfall trend for the following five years from 2006 to 2010 was similar in that both the HY and LY shires matched each other more closely except for the peaks of 2005, 2008 and 2009 again at the shires of Cranbrook, Kojonup and Tambellup respectively. On the other hand, the HY shire of Katanning received a rainfall of 114.76 mm similar to the LY shire of Tambellup which also received a high rainfall of 119.16 mm for June 2008.

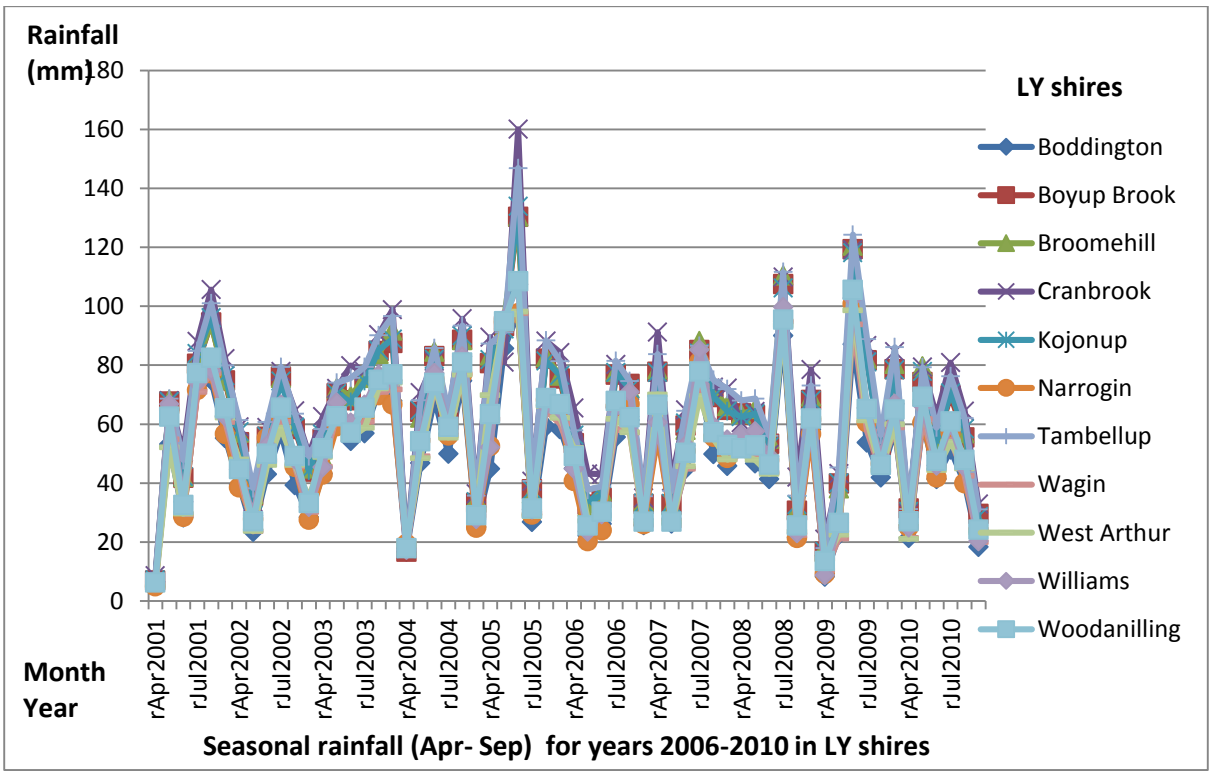



**Figure 8-3. Average monthly rainfall for the LY shires in the 2001-2010 decade**

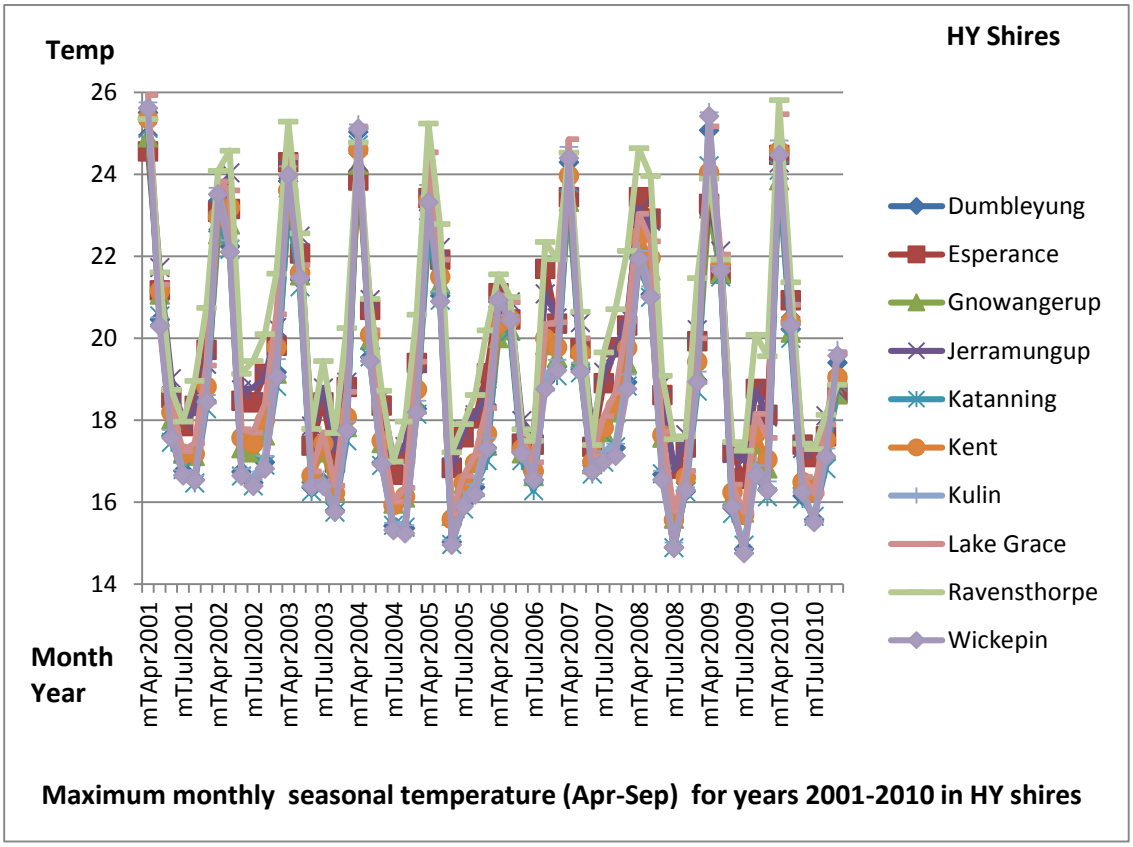



**Figure 8-4. Ave maximum monthly temperature for the HY shires over 2001-2010**

The maximum temperature for all the shires in the study area for the decade from 2001 – 2010 was plotted. The graph for the HY and LY shires are shown in Figure 8-4 and Figure 8-5 respectively.

The visual inspection of the data point plot for the maximum temperature in Figure 8-4 showed some interesting features at the HY shires. The highest maximum temperature of 25.8 degrees occurred at Ravensthorpe in April 2010 and the lowest maximum temperature of 14.8 degrees was at Wickepin in July 2009. The highest mean maximum temperature for the 10 years in HY shires was 19.95 degrees in 2002 and lowest mean maximum temperature was 18.87 in 2005.

Conversely, the maximum temperature data point plot at the LY shires showed further characteristics. The highest maximum temperature of 25.6

degrees Celcius was recorded at the LY shire of Williams in April 2009 and the lowest temperature of 14.6 degrees occurred at Woodanilling in July 2008. The highest mean maximum temperature for the 10 years in the LY shires was 18.91 degrees in 2001 and lowest mean maximum temperature was 17.80 in 2005. The maximum temperature ranged from 15 to 26

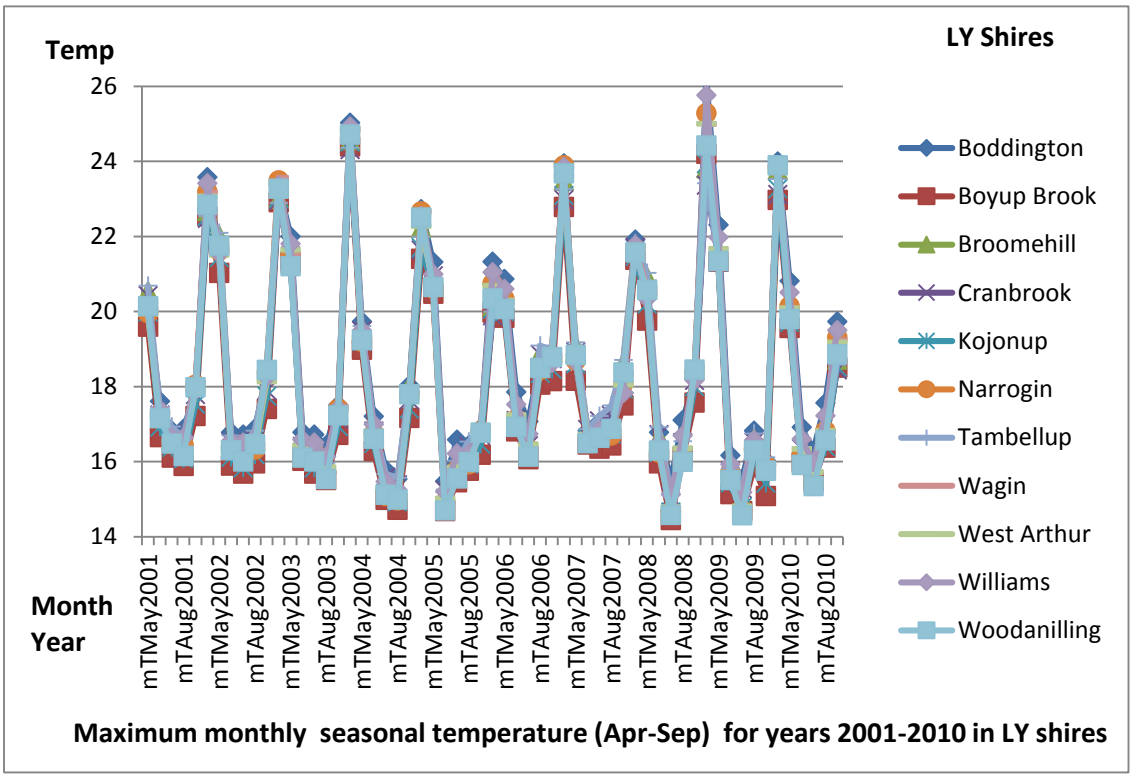



**Figure 8-5. Ave maximum monthly temperature for the LY shires over 2001-2005**

degrees Celcius for both the HY and LY shires. The maximum temperature for the year 2005 was lower for the LY shires than the for the HY shires with a season maximum of 21 degrees for the HY shires and 17 degrees for the LY shires.

The preceding rainfall and temperature visual inspections were followed by an examination of the wheat yield across all the shires in the study area

for the decade from 2001-2010. This process marked the beginning of the exploratory data mining  of the wheat yield.

### 8.5.3 Analysis of the wheat crop yield

The EDM included the analysis of annual wheat crop yields over the the ten years from 2001 to 2010 across all the shires in the study area of the agricultural growing region. This analysis was a two stage process. In the first stage the wheat yields were examined individually. In the second stage they were examined in conjunction with the rainfall and temperature

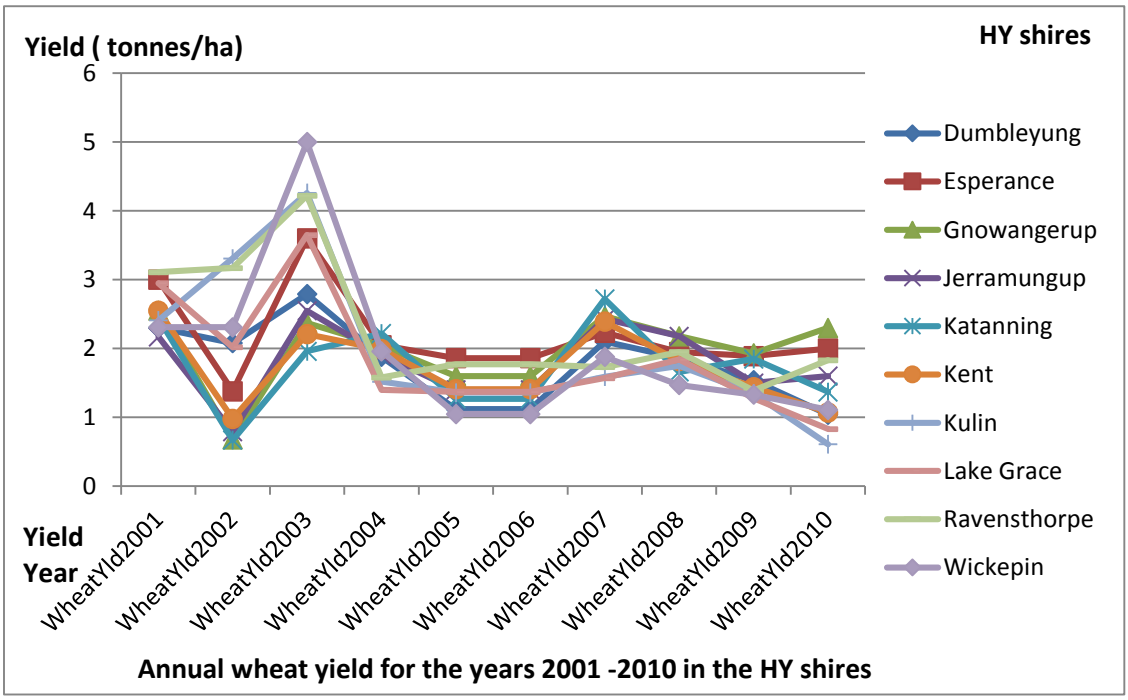



**Figure 8-6. Wheat yield for the years 2001-2010 in the HY shires**

variables. The individual wheat yields for the HY and LY shires were shown in Figure 8-6 and Figure 8-7 respectively.

From Figures 8-6 and 8-7, it can be seen that the wheat yields for 2003 were the highest for the year amongst most of the HY shires especially for the shire of Wickepin, with low yields at some of the HY shires such as Gnowangerup for the year 2002.  This corresponded well with the rainfall for those two years. On the other hand, the wheat yields for the LY shires revealed that the shire of Boddington performed well in 2001, 2004 and

2007 with the shire of Woodanilling producing a low yield in 2002 and high in 2003. Overall the LY shires had higher annual wheat yields than the HY shires.

The second stage of the EDM of the wheat yield was the examination of the combined effect of rainfall and temperature on the wheat yield. The dataset included the April to September snapshots for the attributes for rainfall, temperature (maximum, minimum, variation) and the annual wheat yield. The results were plotted simultaneously as a time-series chart in SPSS as shown in Figure 8-8.

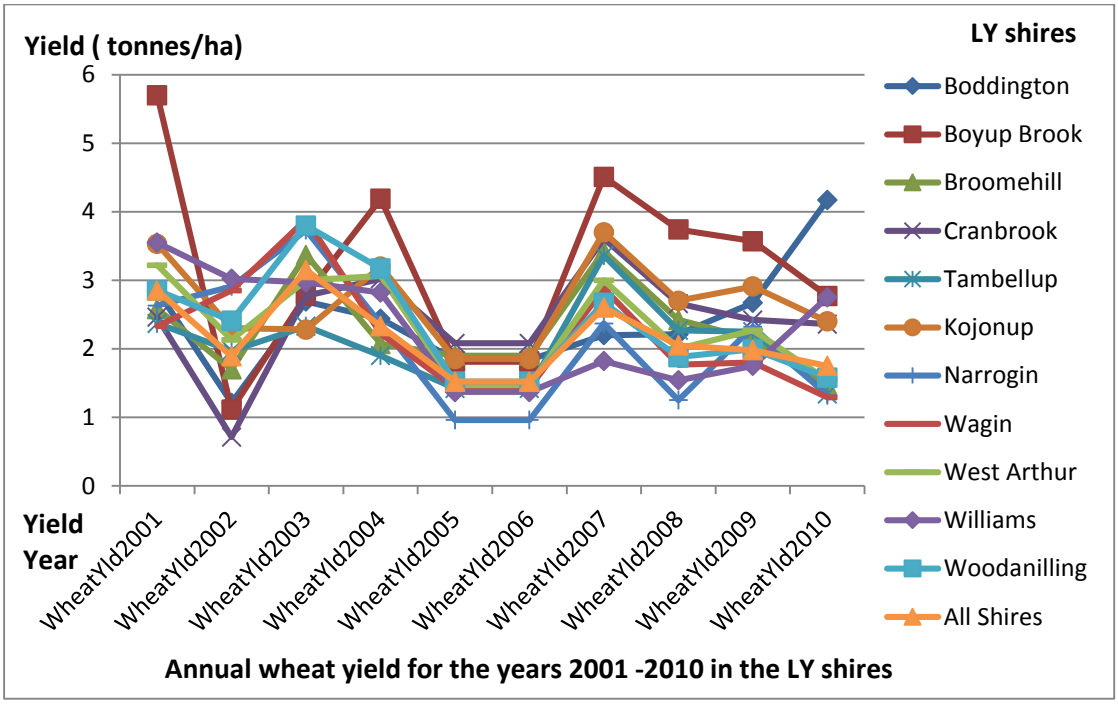



**Figure 8-7. Wheat yield for the years 2001-2010 in the LY shires**

The visual inspection of the rainfall graph highlighted the year 2005 as a particularly wet year, with some shires receiving rainfall around 150mm in June and July. However, the wheat yields were not correspondingly high for 2005. The dry years were also visible as 2002, 2006 and 2010. The wheat yields for these years were correspondingly very low in general and for 2006 in particular, and extremely low in some shires for 2002 and 2010. The wheat yield in 2001 was very high in most shires, and was exceptionally high for the LY shires with a wheat yield of 5.7 tonnes per hectare for the shire of Boyup Brook for example, as per Figure 8-7. The

only outstanding characteristic was the higher maximum temperatures in 2001 at the LY shires as opposed to the other years in the selected decade.

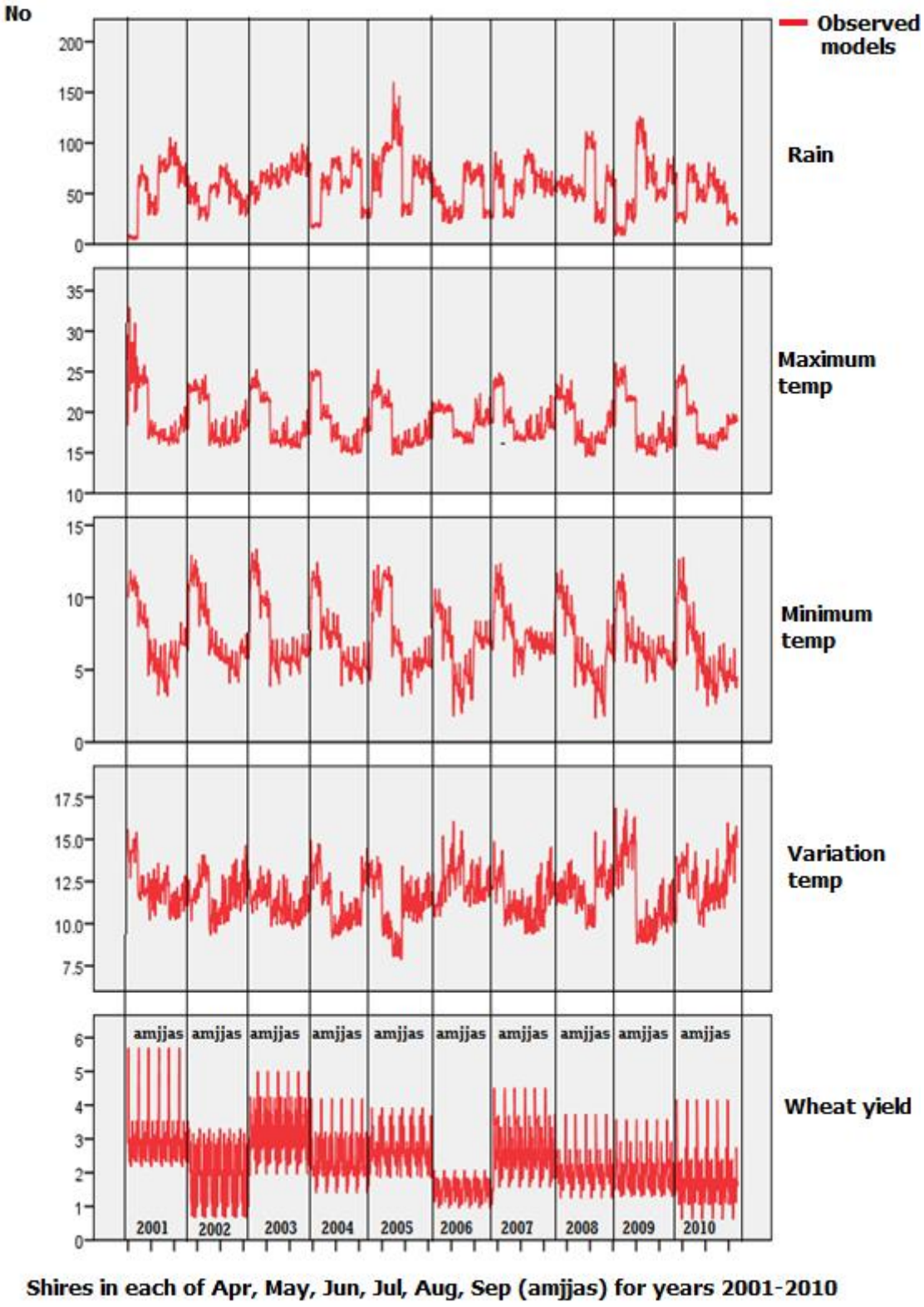



**Figure 8-8. SPSS Time Series Sequence Chart for the 10 years**

Although the year 2003 was a low rainfall year, the rainfall steadily increased over the growing season months.

In addition the maximum and minimum temperatures were fairly uniform over the growing season. This steady rainfall increase and temperature consistency produced a reasonable wheat yield for 2003. The wheat yield for 2007 was similarly characterized for rainfall and temperature except the variation in temperature was higher. The wheat yield for 2007 was therefore similar to the year 2003. The wheat yield for 2009 was slightly lower than for 2005 and the year 2009 had higher rainfall for the growing season months.

For the purposes of clarification, a mean seasonal rainfall below 50mm was rated as dry, 51-60mm as average and over 60 as wet. Mean seasonal temperatures between 18- 18.5 degrees were rated as low, 18.51 to 19.00 degrees as medium and over 19.00 degrees as high. Wheat yields between 1.50 to 2.00 tonnes/hectare were rated as low, 2.01 to 2.50 tonnes/hectare as medium, and 2.51 to 3.00 tonnes/hectare as high and over 3.01 tonnes/hectare as very high. These measurements were discretised and summarized in Table 8-1.

**TABLE 8-1. THE GENERAL WHEAT YIELD RESPONSE TO MEAN SEASONAL RAINFALL AND MAXIMUM TEMPERATURE**

| YEAR | MEAN RAINFALL | RAINFALL RATING | MEAN TEMP | TEMP RATING | ANNUAL YIELD | YIELD RATING |
|---|---|---|---|---|---|---|
| 2001 | 56.84 | Ave | 19.93 | High | 2.85 | High |
| 2002 | 49.32 | Dry | 19.22 | High | 1.89 | Low |
| 2003 | 68.62 | Wet | 18.75 | Med | 3.15 | Vhigh |
| 2004 | 55.41 | Ave | 18.53 | Med | 2.33 | Med |
| 2005 | 76.45 | Wet | 18.27 | Low | 2.62 | High |
| 2006 | 46.82 | Dry | 18.91 | Med | 1.52 | Low |
| 2007 | 60.12 | Ave | 18.97 | Med | 2.60 | High |
| 2008 | 59.92 | Ave | 18.52 | Med | 2.05 | Med |
| 2009 | 58.23 | Ave | 18.50 | Low | 1.98 | Low |
| 2010 | 47.75 | Dry | 18.83 | Med | 1.75 | Low |

Accordingly, one could expect the wheat yield to be high when the rainfall was above 56mm and the seasonal temperature was over 18.51 degrees Celcius. However, it was uncovered that the yield could still be high with lower temperatures, provided that the average seasonal rainfall was very high such as for 76.45mm for the year 2005.

A second time-series model was created using rainfall as a predictor for both temperature and wheat yield. The rainfall and temperature variables were annualized. Each model within the series maps the predictor variable with the dependant variable. The series of models is shown in Figure 8-9. From the simultaneous plot, it became apparent that rainfall affected the maximum temperature, as was evident for the years 2003 and 2005. Generally, the wheat yield was high when the rainfall was high and the maximum temperature low. Under these conditions, the LY shires produced higher yields as opposed to the HY shires. High temperature variations also decreased the wheat yield. The conclusions from these observations were as a result of visual correlations constructed from the data.

In order to augment the correlation analyses, a facility for prediction through the use of regression was necessary. Consequently, DM was explicitly chosen for this purpose in order to reveal potentially hidden patterns and relationships that would otherwise be obfuscated by the multiplicity as well as the opposing nature of the variables. This transition to microscopic exploration included the facility for prediction from regression analyses as well as a correlation between the opposing variables.

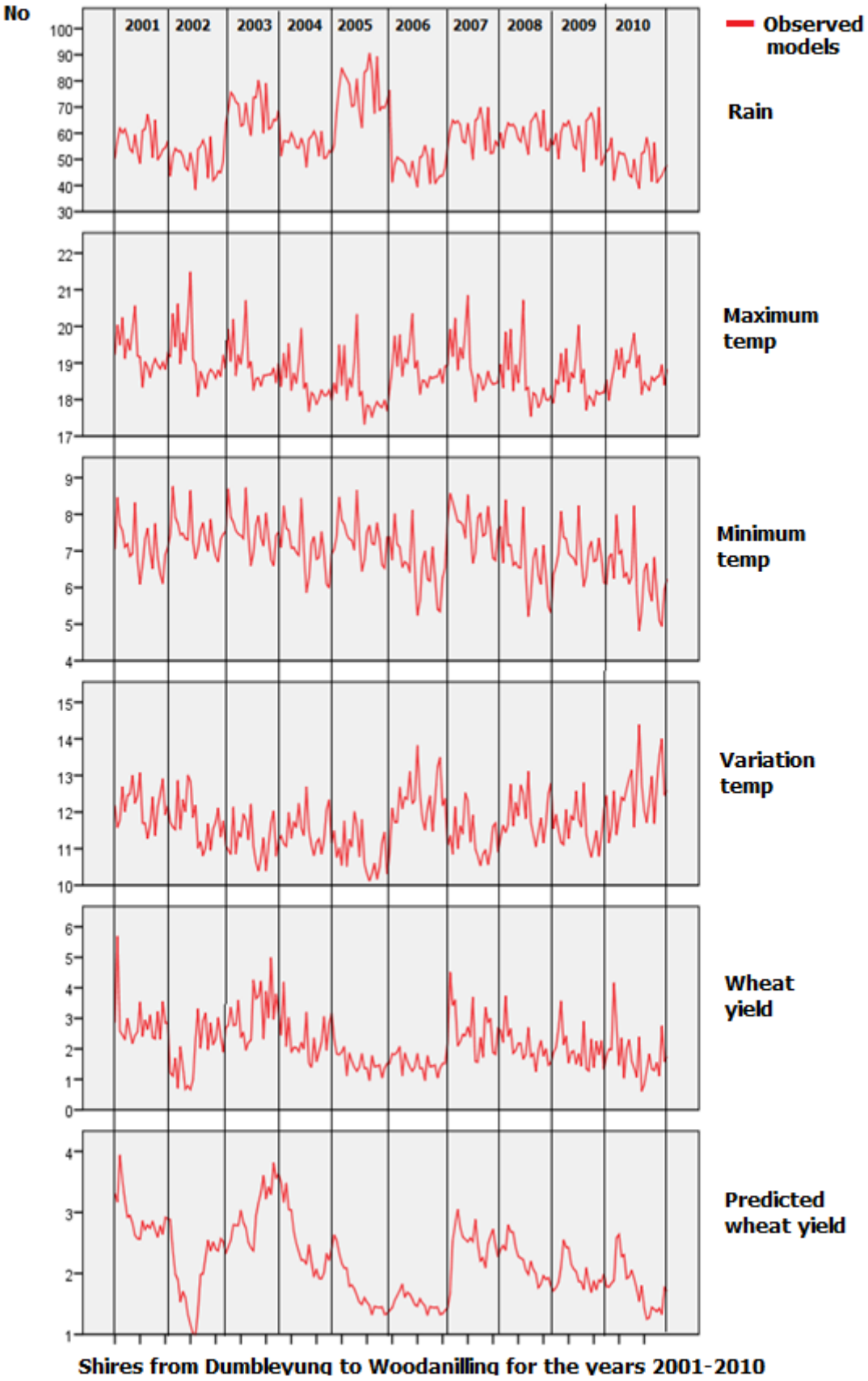



**Figure 8-9. SPSS Prediction model using annualised climate data for 10 years**

### 8.5.4 Data mining analysis

The next step in the individual scrutiny of the exercise was the use of regression. This was done in order to determine if the relationship established through simple observed correlation could be supported by a mechanism of predicting the wheat crop yield. The analysis would involve investigating the variation in temperature and rainfall across the growing months from April through September. This was carried out using the classification technique of DM in the WEKA software. The aggregated data for average monthly rainfall, maximum, minimum and variation in temperature together with the wheat crop yields for the 21 shires was used for this activity. In addition, an average of all the shires was added for the mean crop yield and temperature variables.

The first part of the investigation was the analysis of the average maximum monthly temperature. The average maximum monthly temperatures were taken across the growing season from planting month of April to harvesting month of September. The shires were sorted into HY and LY groups. The aggregated wheat crop yield and the average monthly maximum temperatures and rainfall dataset was split up into a training set and a test set. The training set comprised the data for the years 2001, 2004 and 2006 and the test set was made up of the data for years 2002, 2003, 2005, 2007, 2008, 2009 and 2010. The specific choices for the test set were made in order to provide a mix of the varying conditions high and low rainfall, temperature and yield. The exploratory part of the DM activity was to use the training set to determine the best-fit algorithm using a simple model of crop yield as a function of the location class and the average maximum monthly temperatures for the growing season months as the attributes.

All of the classification algorithms were tested in this step and a short-list of 10 algorithms was selected. These algorithms were Gaussian Processes (GP), Multilayer Perceptron (MP), Lazy LWL, RBF Network (RBF), M5 rules (M5R), Decision Stump (DS), M5P, Kstar, Sequential, SMO and Additive Regresssion. All these algorithms use regression for

predicting continous values in response to input values. These results are summarised in Table 8-2.

**TABLE 8-2. WEKA ALGORITHMS RESULTS FROM THE TRAINING DATASET TEMP AND RAINFALL APR – SEP**

| WEKA ALGORITHM | CORREL COEFF TRAINING | RMSE TRAINING SET | RMSE CROSS VALID | RMSE TEST SET |
|---|---|---|---|---|
| Gaussian Processes | 0.9874 | 0.3336 | 0.6724 | 0.8168 |
| MLP | 0.9982 | 0.0826 | 0.4331 | 1.1879 |
| RBF Network | 0.6297 | 0.6250 | 0.6997 | 0.8879 |
| SMOreg | 0.9454 | 0.2626 | 0.4046 | 1.0664 |
| Kstar | 1.0000 | 0.0848 | 0.5798 | 0.7997 |
| Lazy LWL | 0.9157 | 0.3243 | 0.6731 | 0.9381 |
| Additive Regression | 0.9566 | 0.2369 | 0.6446 | 1.5812 |
| M5 Rules | 0.9453 | 0.2625 | 0.9167 | 1.0560 |
| Decision Stump | 0.6847 | 0.5863 | 0.4551 | 1.1359 |
| M5P | 0.9414 | 0.2736 | 0.9167 | 1.0882 |

In addition to the classification algorithms used in Chapters Six and Seven as described in Section 6.8.5, a few other algorthms were tested. Decision Stump is a classifier regression scheme that is a one node decision tree that limits the predictions to a single input feature (Matthews & McCaffery, 2012); M5 Rules are a way of using a model decision tree built from a cluster divided attribute space to from rules (van Gool, 2011); M5P is a version of M5 called M5 Prime (van Gool, 2011); and LWL is a local weighted learning classifer that is categorised as lazy and which uses original data and its neighbours to learn (Howden et al., 2007). Each of the algorithms trialled had different characteristics, correlation co-efficients and Root Mean Square Errors as shown in Table 8-2. The two main factors of correlation coefficient and the RMSEs were established as initial criteria for selection of the final algorithm upon which the test set would run. The third criterion was the RMSE of the cross validation. These considerations were similar to what was described for Chapter Six and Chapter Seven.

The algorithm with the best performance in these three criteria turned out to be GP with a correlation of 0.9874 and a RMSE for the training set of

0.3336 as well as a RMSE of 0.6724 for the cross validation run and RMSE of 0.8168 for the test set. Consequently, the results for the GP algorithm were used for the analyses and split into two tables of actual and predicted wheat crop yields, one for the years 2003 and 2005 and another for the years 2007 and 2009 as Table 8-3 and Table 8-4 respectively. The results for the years 2002 and 2008 were similar and were therefore omitted.

Similar to the analysis of temperature effects in Chapter Seven, no form of outlier deleltion or adjustment was made on the results. Rather, the results were optimised using the MLP cross validation of 22 folds for each of years 2003, 2005, 2007 and 2009. The predicted wheat results for the year 2002 was markedly lower than the actual wheat yield. It was therefore omitted as an outlier. On the other hand, the preditions for the year 2008 very closely matched those for the year 2007. It was therefore removed from the comparisons for the reason of avoiding similarity. The shires in both tables were grouped into HY and LY shires, where the LY shires are denoted by blue shading. The average of all the shires was denoted as the *All Shires* location and was unshaded.

## 8.6 Prediction Results

The prediction results for the high wheat yield of 2003 and the high rainfall year of 2005 are shown in Table 8-3.

With reference to Table 8-3, there were both underestimated and overestimated predictions for the shires for the high crop yield year of 2003. None of the HY shires had positive prediction errors. On the other hand, only the LY shires of Boddingtion and Williams had positive prediction errors for the year 2003. The remaining LY shires had negative prediction errors for the year 2003. Based on the range-based, prediction errors for the year 2003. The remaining LY shires had negative prediction errors for the year 2003.

**TABLE 8-3. THE WHEAT YIELD GP/MLP RESULTS IN WEKA ALL FEATURES 2003/2005**

| HIGH YIELD YEAR 2003 | | | | HIGH RAINFALL YEAR 2005 | | |
|---|---|---|---|---|---|---|
| Rural Shire HY+ LY | Actual Yield ton/ha | Pred. Yield ton/ha | % Error | Actual Yield ton/ha | Pred. Yield ton/ha | % Error |
| Dumbleyung | 2.69 | 2.96 | 10.04 | 2.46 | 2.43 | 1.42 |
| Esperance | 2.78 | 3.14 | 12.88 | 2.68 | 2.60 | 3.10 |
| Gnowangerup | 3.37 | 3.30 | 2.11 | 2.29 | 2.23 | 2.49 |
| Jerramungup | 2.77 | 3.15 | 13.61 | 2.22 | 2.23 | 0.54 |
| Katanning | 2.79 | 3.42 | 22.62 | 2.22 | 2.24 | 0.68 |
| Kent | 3.60 | 3.07 | 14.61 | 2.28 | 2.23 | 2.06 |
| Kulin | 2.37 | 2.26 | 5.95 | 1.92 | 2.23 | 16.25 |
| Lake Grace | 2.55 | 2.23 | 11.29 | 2.11 | 2.23 | 5.69 |
| Ravensthorpe | 1.96 | 2.26 | 16.17 | 2.54 | 2.23 | 12.20 |
| Wickepin | 2.21 | 2.28 | 1.31 | 1.84 | 2.23 | 21.41 |
| Boddington | 2.28 | 4.47 | 0.96 | 3.93 | 2.59 | 34.15 |
| Boyup Brook | 4.26 | 3.28 | 23.00 | 2.6 | 2.24 | 13.88 |
| Broomehill | 3.65 | 3.34 | 8.38 | 2.61 | 2.23 | 14.41 |
| Cranbrook | 3.73 | 3.32 | 10.88 | 3.38 | 2.60 | 23.20 |
| Kojonup | 4.22 | 3.36 | 20.38 | 2.92 | 2.64 | 9.73 |
| Narrogin | 2.33 | 3.54 | 3.26 | 2.22 | 2.24 | 0.72 |
| Tambellup | 3.89 | 3.30 | 15.12 | 2.75 | 2.68 | 2.65 |
| Wagin | 3.00 | 3.04 | 1.20 | 2.54 | 2.62 | 3.03 |
| West Arthur | 5.00 | 2.97 | 40.62 | 3.70 | 2.58 | 30.16 |
| Williams | 2.97 | 3.25 | 9.46 | 2.61 | 2.79 | 6.80 |
| Woodanilling | 3.80 | 3.09 | 18.58 | 3.19 | 2.50 | 21.60 |
| All Shires | 3.15 | 3.16 | 0.19 | 2.62 | 2.48 | 5.31 |

Based on the range-based, discerning grade scale previously established, good predictions were considered to have a percentage error of less than 20%, with average predictions a percentage error of 21-40% and weak predictions a percentage error of over 40%. This was the metric established for classifying the wheat crop yield in relation to rainfall and temperature prediction performances.

Accordingly, the HY shires of Dumbelyung (10.0%), Esperance (12.9%), Gnowangerup (2.1% ), Jerramungup (13.6%), Kulin (5.9%), Kent (14.6%), Lake Grace (11.3%), Ravensthorpe (16.2%) and Wickepin (1.31%) had good predictions. Only the HY shire of Katanning (22.6%) had an average prediction for the year 2003. There were no weak predictions for the HY shires for the year 2003. Conversely, the LY shires of Boddington (1.0%) Broomehill (8.4%), Cranbrook (10.9%), Narrogin (3.2%), Tambellup (15.1%), Wagin (1.2%), Williams (9.5%) and Woodanilling (18.6%) had good predictions for the year 2003. The LY shires of Boyup Brook (23%), and Kojonup (20.3%) had average predictions. Only the shire of West Arthur (40.6%) had a weak prediction. Overall both HY and LY together and Kojonup (20.3%) had average predictions. Only the shire of West Arthur (40.6%) had a weak prediction. Overall both HY and LY together had an average prediction with a prediction error of 0.19%.

The prediction results for the wet year 2005 are also shown in Table 8-3. Overall there were more underestimations of the wheat yield predictions as opposed to overestimations for the year 2005. On the basis of the good, average and low scale, all of the HY shires had good predictions. The shires of of Dumbleyung (1.4%), Esperance (3.1%), Gnowangerup (2.5%) , Jerramungup (0.5%), Katanning (0.7%), Kent (2.1%), Kulin (16.2%), Lake Grace (5.7%) and Ravensthorpe (12.2%) all had good predictions with the only average prediction for Wickepin (21.4%). On the other hand the LY shires of Boyup Brook (13.9%), Broomehill (14.4%), Kojonup (9.73%), Narrogin (0.7%), Tambellup (2.6%), Wagin (3.0%), and Williams (6.8%), had good predictions. The LY shires of Boddington (34.1%). Cranbrook (23.2%), West Arthur (30.1%) and Woodanilling (21.6%). There were no weak predictions for either the HY or the LY shires for the year 2005. All the shires taken as the *All Shires* location had a good prediction of 5.3%. Overall, the prediction errors relative to the combined effect of rainfall and temperature were better in 2005 than in 2003.

The DM results using the GP algorithm and the MLP individual cross validation run for the years 2007 and 2009 were shown in Table 8-4.

**TABLE 8-4. THE WHEAT YIELD GP/MLP RESULTS IN WEKA ALL FEATURES 2007/2009**

| HIGH YIELD YEAR **2007** | | | | HIGH RAINFALL YEAR **2009** | | |
|---|---|---|---|---|---|---|
| Rural Shire HY+ LY | Actual Yield ton/ha | Pred. Yield ton/ha | % Error | Actual Yield ton/ha | Pred. Yield ton/ha | % Error |
| Dumbleyung | 2.10 | 2.45 | 16.81 | 1.54 | 1.86 | 21.04 |
| Esperance | 2.23 | 2.57 | 15.07 | 1.89 | 1.93 | 2.33 |
| Gnowangerup | 2.47 | 2.75 | 11.42 | 1.93 | 2.00 | 3.52 |
| Jerramungup | 2.43 | 2.53 | 4.20 | 1.50 | 1.94 | 29.47 |
| Katanning | 2.72 | 2.71 | 0.44 | 1.85 | 2.12 | 14.81 |
| Kent | 2.39 | 2.47 | 3.35 | 1.44 | 1.86 | 29.44 |
| Kulin | 1.59 | 2.19 | 37.86 | 1.35 | 1.80 | 33.63 |
| Lake Grace | 1.57 | 2.19 | 39.62 | 1.28 | 1.90 | 48.59 |
| Ravensthorpe | 1.73 | 2.22 | 28.21 | 1.39 | 2.00 | 43.60 |
| Wickepin | 1.88 | 1.88 | 0.05 | 1.33 | 1.85 | 39.40 |
| Boddington | 2.2 | 2.22 | 1.09 | 2.67 | 2.24 | 16.25 |
| Boyup Brook | 4.51 | 2.64 | 41.42 | 3.57 | 2.25 | 37.11 |
| Broomehill | 3.42 | 2.78 | 18.65 | 2.14 | 2.17 | 1.26 |
| Cranbrook | 3.61 | 2.73 | 24.49 | 2.42 | 2.11 | 12.81 |
| Kojonup | 3.70 | 2.89 | 22.03 | 2.91 | 2.25 | 22.85 |
| Narrogin | 2.37 | 2.51 | 6.08 | 2.32 | 2.24 | 3.41 |
| Tambellup | 3.36 | 2.80 | 16.64 | 2.25 | 2.09 | 7.02 |
| Wagin | 2.88 | 2.53 | 12.15 | 1.80 | 1.96 | 8.72 |
| West Arthur | 3.00 | 2.53 | 15.53 | 2.27 | 2.24 | 1.19 |
| Williams | 1.82 | 2.25 | 23.52 | 1.75 | 2.03 | 15.71 |
| Woodanilling | 2.67 | 2.63 | 1.50 | 1.99 | 2.00 | 0.65 |
| All Shires | 2.60 | 2.58 | 0.96 | 1.98 | 1.87 | 5.71 |

The HY shires of Dumbelyung (16.8%), Esperance (15.0%), Gnowangerup (11.4%), Jerramungup (4.2%), Katanning (0.4%), Kent (3.3%) and Wickepin (0.1%), had good predictions. The HY shires of Kulin (37.9%), Lake Grace (39.6%) and Ravensthorpe (28.2%) had average predictions for 2007. Conversely, the LY shires of Boddington (1.1%), Broomehill (18.7%), Narrogin (6.1%), Tambellup (16.6%), Wagin (12.1%), West Arthur (15.5%) and Woodanilling (16.1%) had good predictions. The shires of Cranbrook (24.5%), Kojonup (22.0%), and Williams (23.5%) had average predictions. Only the shire of Boyup Brook (41.4%) had a weak prediction. Overall both HY and LY shires taken together as the *All Shires* location had an good prediction of 1.0%.

The prediction results for the wet year 2009 were analysed with reference to Table 8-4. Overall there were more positive prediction estimations for the year 2009 which meant that there were over estimations in predictions rather under estimations. On the basis of the good, average and low scale, the HY shires of Esperance (2.3%), Gnowangerup (3.5%), Katanning (14.8%) had good predictions with an average prediction for the shires of Dumbleyung (21.0%), Jerramungup (29.5%), Kent (29.4%), Kulin (33.6%), Wickepin (39.4%). The shires of Lake Grace (48.6%) and Ravensthorpe (43.6%) had weak predictions.

On the other hand the LY shires performed much better with the shires of Boddington (16.2%), Broomehill (1.3%), Cranbrook (12.8%), Narrogin (3.4%), Tambellup (7.0%), Wagin (8.7%), West Arthur (1.2%), Williams (15.7%) and Woodanilling (0.6%) had good predictions. The shires of Boyup Brook (37.1%) and Kojonup (22.9%) had average predictions. There were no weak predictions in the LY shires for 2009. All the shires taken together as the *All Shires* location had a good prediction of 5.7%. Overall, the prediction errors relative to the combined effect of rainfall and temperature were better in 2007 than in 2009.

## 8.7 Discussion

The establishment of the crop model for climatic effects involved taking into account a number of additional factors as previously noted in Section 6.9.

Despite the effects and interactions of these contributory yet extraneous factors, a pre-cursory relationship using a simple crop model was used where the classification entity was the rural shire. The average monthly rainfall, maximum, minimum and variation in temperature were used as predictors of the wheat crop yield individually. The analysis was based on the earlier established metrics and on the general two stage method of macroscopic and microscopic scrutiny. The macroscopic analysis involved the visual graphical inspection, and the individual wheat crop over the selected decade of examination, as part of an exploratory data mining

process. The simple bar graph visualizations allowed the initial physical recognition of patterns and trends of specific wheat production as related to the stochastic average monthly rainfall and temperature. The microscopic analysis involved scrutinizing the selected wheat crop yield for the selected years of 2003, 2005, 2007 and 2009.

## 8.8 Conclusion

The wheat yield predictions across the years 2003 (3.6%), 2005 (5.31%), 2007 (0.9%), 2009 (5.7%) showed a significant trend. The years 2003 and 2007 indicated a gradual improvement of 2.5%. Conversely, the trend between 2005 and 2007 showed a marginal decrease in yield predcition of 0.4%. This indicates that the accuracy of the predictions improved as the actual yields dropped. It could be concluded that higher rainfall and lower temperatures caused a decrease in wheat yield, but allowed for more accurate predictions. In addition, the LY shires had better crop yields as well as better predictions overall due to the slightly lower temperatures prevalent in them. This amounted to a complex relationship between temperature and the wheat crop yield prediction where the crop yield could be expected to be better at higher temperatures and worse at lower temperatures. However slight decreases at the higher end of the temperature scale caused significant improvement to the crop yield, whilst slight decreases at the lower end of the scale caused marginal improvement to the crop yield.

The use of the data mining classification functions GP together with the use of individual yield year, MLP cross validation showed that the correlation between the stochastic average monthly temperature, rainfall and wheat yield was a strongly positive one and that as a result generally, wheat yield in the South West agricultural region can be expected to increase marginally with an increase in temperature associated with a decrease in rainfall. It also showed that wheat yield could increase when the temperature fell, as rainfall increased. These variations in wheat yield

were more prevalent in most of the LY shires and only some of the HY shires. However, there could be under-estimation errors in predicting the wheat yields. It is thought that the predictions could be further refined. This can be achieved if the influence of other factors as well as to the sparseness related to the intrinsic shire yield measurement of the dataset. The effect of other factors such as soil type and prior season soil-water retention respective to the shire areas under scrutiny could be investigated in future extensions to the research.

## 8.9 Chapter Review

The work done in this chapter covered the establishment of a relationship between climate represented as stochastic average monthly rainfall, temperature and the wheat crop yield. Similar to the analytical foundations of Chapter Six and Chapter Seven, the investigation featured both a macroscopic (visual and times-series) and a microscopic (DM) analysis.

Data mining was utilized in order to supplement the simple correlations that were established between rainfall, temperature and crop yield with a facility for crop prediction. The data mining algorithms that were tested were all based on the classification technique. The results were extracted from a combination of the GP and the MLP classification algorithms. The results from the two stage analysis showed that there was a good correlation between stochastic average monthly climate and wheat yield.

# Chapter 9

# RESEARCH SUMMARY

This chapter briefly revisits the major themes and research questions and conclusions, outlines the limitations and explores potential for new directions. As such, the chapter is organized into four sections: Section 9.1 briefly reviews the present research; Section 9.2 outlines the main contributions made by the research; Section 9.3 discusses the limitations of the research and Section 9.4 charts a pathway for future study and investigations.

## 9.1 Overview

The focus of this research was on improving the quality of wheat crop yield through a reliable predictive process within a study area in the South Western Agricultural region of WA. This prediction process was dependent upon the dichotomous influences of both gradual and continuous variation parameters. The continuous variation of the climate (rainfall and temperature) variables and gradual variation of predominant soil type were examined with reference to shires in the study area for the decade 2001 to 2010.

## 9.2 Conclusions

Pattern establishment within complex data is a difficult task. The researcher sought to find ways and methods to efficiently and effectively process such raw data into meaningful information, then practical knowledge and finally into insightful industrial best practices for the agricultural industry in the selected study area. Part of the process was to cycle feedback into the system to maintain the continuum of information in a dynamic manner. However, the main outcomes were the results of the

experiments and definition of the relationships between the variables. The research activities outlined in the preceding chapters, and subsequent results led to the following conclusions for the specified study area:

1. Rainfall affects the wheat crop yield in a directly proportional relationship, i.e., wheat yield increases with rainfall. As the rainfall varied between the years within the 10 year span, the predictions for wheat yield are similarly affected.
2. Temperature affects the wheat crop yield in a directly proportional relationship, i.e., wheat yield increases with an increase in temperature. However, only marginal increases in temperature resulted in yield increases, as large increases in temperature caused lower yields.
3. Moreover, when the rainfall increase caused a decrease in temperature, the wheat crop yield increase is also arrested marginally.
4. Rainfall affects the soil moisture content and, by extension, the soil type (due to water retention) of particular shires with the study area.
5. The over-arching influential factor was rainfall.

Consequently, the contributions of this research, in light of the conclusions drawn above, were both of a direct and indirect nature. The conclusions themselves constituted major contributions of a direct nature. However, the detailed contributions of this research have already been covered in Section 1.5. Instead, only a summarised version is presented here.

The major contributions of an indirect nature related to the contextual setting and the locus of the attached variables and parameters. Firstly, farmers and agricultural practitioners within the agricultural region of South Western Australia are now able to seek from DAFWA, improved guidelines for crop planning and estimation. These can be tailored to the gradual and continuous variation parameters prevalent in their particular growing environment. This is because each 100 hectare area within a shire, within

the study area, can now be attached to a detailed rainfall, temperature and soil composition profile. It is now possible for farmers to know the predominant type of natural vegetation within their growing areas to assist in crop variety selection. This will complement the actual and predicted wheat yields for the whole shire. They will also be able to supplement their locally kept climate records and official records held at DAFWA with the stochastic measurements from this research. This means that current data records and information will be greatly augmented thereby facilitating more informed crop recommendation and selection choices.

Secondly, this research has resulted in a guideline for processing raw and complex data with both continuous and gradual variation characteristics. The guideline is the DM framework, which was produced from generalisation and abstraction of the methodology used in this research. Future researchers can be guided into effective and methodical data extraction, organisation and storage, interrogation, analysis and presentation. Effectively, this means that research activities may be made in relation to the various processes outlined in the framework. Selection of framework components, through addition or deletion, will allow other researchers to refine their own frameworks.

A third consequence of this research approach is the metrics of framework evaluation that was devised for the purpose of establishing a standard. This standard could be used as guideline upon which to gauge the measure of effectiveness, usefulness and validity of similar frameworks through the established scale. Researchers could use other characteristics of frameworks defined in the metrics of evaluation such as predictability, DM techniques and reusability for design purposes. The metrics could also serve as a framework classification guide where several frameworks could be examined for selection or recommendation comparatively.

## 9.3 Limitations

It was inevitable that over the course of this research undertaking, several limitations emerged:

1. The rainfall and temperature *datasets were interpolations* calculated from original data from sparsely distributed weather stations. They were therefore mostly calculated estimates or approximations.

2. These datasets were *only sampled for the last decade* due to data availability and processing complexity.

3. The method of interpolation was *limited to ordinary krigeing*.

4. The study of the climatic effect was *limited to rainfall and temperature only* because of the exclusion of radiation and evaporation due to inherent complexity.

5. The research did not involve multiple case studies and instead, *used a single but complex case study* as the former would not have been feasible within the available resources.

Notwithstanding these limitations, the research was fruitful in that there were concrete benefits to involved parties and definite overall contributions to the broader field of data mining strategies.

## 9.4 Future directions

The work presented thus far was neither static nor sterile in that it inevitably led to the question of "*whither next*" that encapsulates productive research. In this regard, there are both lateral and vertical potential extensions to the current research. The former represents the immediate and obvious pathway of expanding the research through overcoming the limitations

mentioned in Section 9.4. Conversely, the latter extension to the research is the pathway that introduces the missing factors, based on water usage and retention that would have assisted the estimations of crop yield in this context. Potential lateral expansions include:

1. *Increase grid cell size (resolution)* from 1000 m to 50 m.
2. *Refine the interpolation* to cover other interpolation methods such as the cubic spline or inverse distance weighting methods.
3. *Increase the dimensionality of the climate variables* to cover a longer period of time i.e. four decades instead of one.
4. *Incorporate radiation and evaporation* into the climate variables dataset.
5. *Include other crops such as barley, canola, lupins and oats in the predictions.*
6. *Include other states or countries in the analysis.*

Whilst, the aforementioned extensions would give a wider lateral coverage to the research, the extension of *water usage and retention would add depth to the research*. Water usage for wheat cultivars would need to be gathered for the last decade and used to calculate the effective crop water use for its effect on crop yield potential, according to the French and Schultz equation (Anderson & Garlinge, 2000). This would probably require the use of sensors for periodic detection and recordings.

The future possibilities arising from both the lateral and vertical extensions to this research are likely to have a cascading and engendering effect on the research community and the exploration space both locally and internationally.

# ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| AR | Action Research |
| AR | Additive Regresssion |
| AEZ | Agro-Ecological Zoning |
| ANCOVA | Analysis of co-variance |
| ANOVA | Analysis of variance |
| ARMS | Army Remote Moisture System |
| BMP | Bit Map |
| BI | Business Intelligence |
| CAR | Canonical Action Research |
| CBR | Case Based Reasoning |
| CART | Classification And Regression Tree |
| CLI | Command Line Interface |
| CSV | Comma-Separated Value |
| CRS | Coordinate Reference System |
| CM | Crop Model |
| CVT | Crop Variety Testing/Trails |
| DEA | Data Envelopment Analysis |
| DM | Data mining |
| DW | Data Warehousing |
| DS | Decision Stump |
| DSS | Decision Support Systems |
| DAFWA | Department of Agriculture and Food of Western Australia |
| DEM | Digital Elevation Model |
| ENSO | El Nino – Southern Oscillation |
| EDW | Enterprise Data Warehouse |
| EDM | Exploratory Data Mining |
| EVDM | Exploratory Visual Data Mining |

| | |
|---|---|
| ETL | Extract, Transform and Load |
| GIGO | Garbage in, Garbage out |
| GP | Gaussian Processes |
| GCM | General Circular Model |
| GA | Generic Algorithms |
| GLM | Generalized Linear Modelling |
| GUIDE | Generalized Unbiased Interaction Detection and Estimation |
| GPS | Global Positioning System |
| GPL | GNU General Public License |
| GUI | Graphical User Interface |
| ha | hectare |
| HR | High Rainfall |
| HY | High Yield |
| HCI | Human Computer Interface |
| IDS | Intrusion Detection System |
| IDW | Inverse Distance Weighting |
| JPEG | Joint Photographic Experts Group |
| KDD | Knowledge Discovery through Databases |
| KDDM | Knowledge Discovery and Data Mining |
| LIS | Land Information System |
| LUCC | Land use, Cover and Change |
| LR | Low Rainfall |
| LY | Low Yield |
| MAE | Mean Absolute Error |
| M5R | M5 rules |
| MSS | Modified Selective Subset |
| MLP | Multilayer Perceptron |
| NASS | National Agricultural Statistics Service |
| NN | Neural Networks |

| | |
|---|---|
| NRARD | Non-Redundant Association Rule Discovery |
| NDVI | Normalized Difference Vegetative Index |
| PAR | Participatory Action Research |
| PNG | Portable Network Graphics |
| PA | Precision Agriculture |
| PC | Principal Components |
| PDF | Probability Distribution Function |
| PPR | Projection Pursuit Regression |
| RBF | Radial Based Function Network |
| RMSE | Root Mean Square Errors |
| SOLAM | Spatial Online Analytical Mining |
| SST | Sea Surface Temperature |
| SMO | Sequential Minimal Optimisation |
| SSCM | Site Specific Crop Management |
| STIFF | SpatioTemporal Integrated Forecasting Framework |
| TIFF | Tagged Information File Format |
| TLU | Tertiary Land Use |
| 3MT | Three Minute Thesis |
| UTM | Universal Transverse Mercator |
| VRA | Variable-Rate Application |
| VHR | Very High Rainfall |
| VDM | Visual Data Mining |
| VDM | Visual data mining |
| WEKA | Waikato Environment for Knowledge Analysis |

# GLOSSARY

**Action research** (AR) is defined by the cyclic process of "*observe, reflect, act, evaluate, modify – move in new directions*" that is commonly referred to as action-reflection.

**Association rules** are used to "show the relationship between a set of antecedents and its associated consequents" (Rodriguez et al., 2004) and are used to "predict any attribute, not just the class.

**Attribute Relation File Format** (ARFF) is the format of the input data required by the WEKA software which is similar to the spreadsheet or comma separated value (csv) format.

The **characteristic component** of data refers to observations and measurements of the data.

**Classification rules** are one of the most commonly used data mining techniques and are considered as a popular alternative to decision trees. Classification uses a set of pre-classified examples which are basically the precondition or antecedent. A model is developed from the preconditions which in turn generate a set of grouping rules. These rules form the consequent from which a new object is characterized.

**Clustering** is a process of grouping data items into sets or clusters based on some measure of similarity.

**Complex data** is data that emanates from various sources and is heterogeneous and can be variously characterised as multi-format, multi-structure, multi-source, multi-modal and multi-version depending on the heterogeneity.

**Data mining** is a semi-automated prediction and analytical process that enables raw data to be transformed into useful information and subsequently into practical knowledge.

**Data warehousing** is the consolidation of integrated and time-varying collection of data from different sources into a central repository that is

intended to be used for the support of management decision making through easy access and manipulation by analysts and decision makers.

**Decision trees** are classified as predictive models. Observations about a particular entity are mapped against conclusions to the items target value. Each node apart from the root corresponds to some variable. The arc to a child node represents a possible value of the variable. Each leaf represents the predicted value of the target variable traversed from the values of the variables in its path to the target from the root.

A **digital image** is made up of a number of points or picture cells (pixels) in a matrix. Examples of digital images include bit maps (BMP), tagged information file format (TIFF), joint photographic experts group (JPEG).

**Dirty data** can generally be characterized as missing, wrong or non-standard representations of the same data, data from legacy sources with inadequate or no metadata descriptors as well as data characterised by issues of pollution, outliers and noise.

**Framework** is a well-structured and refinable specification that permits the identification and understanding of common properties for the purpose of creating models from common abstractions.

A **generalised linear model** (GLM) is used to automatically predict the value of dependant variable from a series of independent variables that do not have a constant relationship with one another.

**A Geographical Information System** (GIS), consists of two major components which are the actual geographic data and the set of data-processing functionality that include collecting, storing, retrieving, transforming and displaying spatial or geographically referenced data within purpose-built software such as the open source QuantumGIS or the proprietary ArcGIS software suite.

**GRASS** is an acronym for Geographical Resources Analysis Support System and is essentially a public domain GIS used for the data management analysis and visualisation of GIS related data and is available

for download and free usage under the terms of the GNU General Public License (GPL).

**Knowledge discovery through databases** (KDD) is “the process of extraction and abstraction of any type of pattern, perturbation, relationship or association from analysed data”

**Machine learning** is the unearthing process of knowledge discovery through databases (KDD) where the acquisition of knowledge from source data is automated through the application of algorithms.

The **necessary view** of probability measures the extent to which one set possibilities supports another out of logical necessity.

**Numeric prediction** is considered as a special type of decision tree that stores a class value that is the average value of instances that reach the leaf node and where the outcome to be predicted is a numeric quantity rather than a discrete class or value.

The **objectivist view** of probability holds that some repetitive actions occur in the same probability as the mathematically calculated repetition of those random actions.

**Online analytical processing** (OLAP) organizes aggregate queries on data, such as sum and averages, for use in decision support algorithms where information is collected from detail tables and from which alternatives, trends and projections are able to be viewed through axes pivoting and changing aggregation.

The **personalistic view** of probability postulates that probability is a measure of the confidence that an individual holds for a particular outcome from a given set of other possible outcomes.

**Precision agriculture** (PA) is a general term that has arisen due to the use of advanced GPS and sensor technology being utilised in the agricultural sector for the purposes of crop and soil management and optimisation of crop yields.

**Predictive clustering** is the extension to clustering and predictive modelling where the clusters are predicted in two dimensions of the target properties as well as the description of object.

**Predictive modelling** or *supervised* learning is somewhat different to clustering. In this method, models are constructed with the ability to predict the value of a target attribute (dependent variable) from the given values for a set of input attributes (independent variables). In other words, the description of an object is used to predict a target property of an object.

**Principal Component Analysis** (PCA) is a method of reducing the dimensionality of a dataset by establishing a set of variables that are pertinent in summarising the attributes of the data and is classically used before clustering.

**QuantumGIS** is an Open Source Geographic Information System (GIS) that is a feature rich environment complete with applications and third-party plug-ins that offer various facilities for the creation, editing, mapping and conversion of a variety of raster and vector file formats that include ESRI shapefiles, ASCII grids, text and images

**Raster data** relevant to GIS are stored as numerically coded grid-cell or pixel data in the form of *n*-dimensional bit or pixel maps.

The **referential component** of data specifies the context of where the observations and measurements of the data were made such as time, location specifics.

The **Revolution R statistical package** is an Open Source software package with a large statistical capability organised into several packages each containing a set of different statistical functions or algorithms.

**Shapefiles** are non-topological data structures represented by rings usually in the form of polygons. They do not explicitly store topological relationships.

**Scaling down** of data can be regarded as data reduction most importantly due to the selection of relevant data through feature and instance selection and discretising, prior to input to the data mining algorithms.

**Scaling up** of algorithms is about refinement and optimisation in relation to insufficient and sparse data.

**Sparse databases** occur where there is limited amount of data available for forming associations between the various tables.

**Spatial data mining and knowledge discovery** (SDMKD) is about uncovering the implicit relationship and characteristics that may exist in large spatial databases.

**Statistics** is a subject that is concerned with inductive inference made from “the understanding of structure in data”

**Supervised learning** is used to find new patterns in well known cases to form generalizations.

**Unsupervised learning** searches for the patterns and the internal structure of a dataset at the outset by using certain characteristics inherent in the data inputs.

**Vector data** relevant to GIS is composed of a series of points, lines and polygons represented as unions or overlays as well as the partitions and networks that are formed by these components for the representation of objects digitally.

**Visual data mining** (VDM) is a process involving human intuitive interaction with data (Y. Vagh & Xiao, 2012) where it has been described as *overview first*, *zoom and filter*, followed by *details on demand,* otherwise known as Schneiderman’s (1996) information seeking mantra.

The **WEKA** software and machine learning toolkit is an excellent general-purpose environment for applying techniques of classification, regression, clustering and feature selection in bioinformatics research.

# INDEX

# APPENDICES

## Appendix A1: Script - Interpolation of the rainfall and temperature datasets.

```
# script to do the krigeing automatically for the full 120 files
# load the gstat library
library(gstat)
# store the current directory
# initial.dir<-getwd()
# setwd(initial.dir)

# change to the directory where the files list is held
setwd("E:/Mar23AgriData/NewClimateExtract/")
# read in the 10 year (120 months) files list for interpolation
ee <- read.csv("rainFilesList.csv")
i=120 # set  for single line testing

for (i in 120:dim(ee)[1]) {
 # change to the directory where the raw data is held
 setwd("E:/Mar23AgriData/NewClimateExtract/")
 filename=paste(ee$filesA[i],"csv",sep=".")
 # read the next file in the list
 e <- read.csv(filename)
 ## convert the rainfall to have no decimals and to increase the
number
 # rename Field4 to maxTemp, minTemp, varTemp as required
 #names(e)[names(e)=="Field5"] ="minTemp"
 e$rain=e$rain*100
 #e$minTemp=e$minTemp*100
 ## convert simple data frame into a spatial data frame object:
 coordinates(e) <- ~ x+y
 ## test result with simple bubble plot:
 #bubble(e, zcol='rain', fill=FALSE, do.sqrt=FALSE, maxsize=2)
 bubble(e, zcol='minTemp', fill=FALSE, do.sqrt=FALSE, maxsize=2)
 ## setup the study area grid for interpolation
 grd=expand.grid(x=          seq(          305000,1060000,1000),y=
seq(6354000,6217000,-1000))
 #combine the coordinates of the grid
 coordinates(grd) <- ~ x+y
 gridded(grd) <- TRUE

  ## make gstat object:
  g <- gstat(id="rain", formula=rain ~ 1, data=e)
  #g <- gstat(id="minTemp", formula=minTemp ~ 1, data=e)
  ## the original data had a large north-south trend, check with a
variogram map
  #plot(variogram(g,     map=TRUE,     cutoff=4000,     width=200),
threshold=10)

 ## another approach:
  # create directional variograms at 0, 45, 90, 135 degrees from
north (y-axis)
  v <- variogram(g, alpha=c(0,45,90,135))

  v.fit  <-  fit.variogram(v,  model=vgm(model='Lin'  ,  anis=c(0,
0.5)))
```

```
 ## plot results for checking:
  plot(v, model=v.fit, as.table=TRUE)
 ## update the gstat object:
 g <- gstat(g, id="rain", model=v.fit )
 #g <- gstat(g, id="minTemp", model=v.fit )
 ## perform ordinary kriging prediction:
 p <- predict(g, model=v.fit, newdata=grd,progress='text')
 ## divide the rainfall back to the original and round to figures
 p$rain.pred=round(p$rain.pred/100,2)
 #p$minTemp.pred=round(p$minTemp.pred/100,2)

 for (j in 1:dim(p)[1]){
       # reset any negative rainfalls to zero
       if (p$rain.pred[j] < 0.00) p$rain.pred[j] = 0.00
       #if (p$maxTemp.pred[j] < 0.00) p$maxTemp.pred[j] = 0.00 #
temp can be below zero
       }
 ## change to the output directory
 setwd("E:/Mar23AgriData/RainRpreditFix_csvs/")
 #setwd("E:/Mar23AgriData/MinTempRpredit_csvs/")
 #set the filename for automatic writing
 filename=paste(ee$filesB[i],"csv",sep=".")
 ## write out the file to the directory
 write.csv(p,filename)
}
```

## Appendix A2: Script used for extracting the interpolated data produced in the previous script (A1) and fit them onto the surface grid.

```
# script to read each of the 120 files and extract the predicted
rain for the grid cells
# set the input directory
 setwd("E:/Mar23AgriData/RainRpreditFix_csvs/")
#setwd("E:/Mar23AgriData/maxTempRpredit_csvs/")
#setwd("E:/Mar23AgriData/minTempRpredit_csvs/")

# create the matrix for output
#mat_e=matrix(0,nrow=104328,ncol=122)
mat_e=matrix(0,nrow=104328,ncol=123)
# add all the colnames for the matrix
colnames(mat_e)=c(substr(list.files(".",pattern=".csv"),1,7),"x","
y")
#colnames(mat_e)=c(substr(list.files(".",pattern=".csv"),5,9),"x",
"y")
# change the matrix to a dataframe
mat_e=as.data.frame(mat_e)
# load the x and y coordinates done once only after the first read
of e below
mat_e$x=e$x
mat_e$y=e$y
i=1
# read each of the 120 files to get the rainfall prediction in the
right column
for (i in 1:dim(ee)[1]) {
 filename=paste(ee$filesB[i],"csv",sep=".")
 colname=ee$filesB[i]
 e <- read.csv(filename)
 mat_e[[colname]]=e$rain.pred
```

```
 #mat_e[[colname]]=e$minTemp.pred
 #mat_e[[colname]]=e$maxTemp.pred
}
#write out the full file of the 120 months predictions
write.csv(mat_e,choose.files())
```

## Appendix A3: Script used to extract the rainfall (temperature) and production data for fitting onto the surface grid.

```
#script to extract the rainfall and production data for each of the
shires in cropping region
ee= read.csv(choose.files())  #Rain2001-2010
# sort the data frame by the shire name
ee=ee[with(ee,order(Shire)),]
ff=matrix(,nrow=1,ncol=61)
#ff=matrix(,nrow=1,ncol=150)
ff=as.data.frame(ff)
colnames(ff)=colnames(ee)
names(ff)=names(ee)
#setwd("E:/Mar23AgriData/Shires/")
setwd("E:/Mar23AgriData/Feb2012/Feb2012FinalFix/")
#total_rain02=0;total_rain03=0;total_rain05=0
total_rain = 0
no=0; j=0; ii=0
prev_shire=ee$Shire[1] # first shire
#avg_rain02=0; avg_rain03=0; avg_rain05=0; ii=1
avg_rain= 0; ii=1

# setup the matrix
dd=matrix(0,nrow=34,ncol=62)
dd=matrix(0,nrow=51072,ncol=12) # for counting up the soil types
# dd=matrix(0,nrow=34,ncol=53)

dd=as.data.frame(dd)
colnames(dd)=c("no_rows",colnames(ee))
#colnames(dd)=c("shire","no_rows","avg_rain02","avg_rain03","avg_r
ain05","Lupins02","Oats02","Barley02","Canola02","Wheat02","Lupins
03","Oats03","Barley03","Canola03","Wheat03","Lupins05","Oats05","
Barley05","Canola05","Wheat05")
#
colnames(dd)=c("shire","no_rows","avg_mxTJan03","avg_mxTFeb03","av
g_mxTMar03","avg_mxTApr03","avg_mxTMay03","avg_mxTJun03","avg_mxTJ
ul03","avg_mxTAug03","avg_mxTSep03","avg_mxTOct03","avg_mxTNov03",
"avg_mxTDec03","avg_mxTJan05","avg_mxTFeb05","avg_mxTMar05","avg_m
xTApr05","avg_mxTMay05","avg_mxTJun05","avg_mxTJul05","avg_mxTAug0
5","avg_mxTSep05","avg_mxTOct05","avg_mxTNov05","avg_mxTDec05","av
g_mxTJan05","avg_mxTFeb05","avg_mxTMar05","avg_mxTApr05","avg_mxTM
ay05","avg_mxTJun05","avg_mxTJul05","avg_mxTAug05","avg_mxTSep05",
"avg_mxTOct05","avg_mxTNov05","avg_mxTDec05","Lupins02","Oats02","
Barley02","Canola02","Wheat02","Lupins03","Oats03","Barley03","Can
ola03","Wheat03","Lupins05","Oats05","Barley05","Canola05","Wheat0
5")
#setup the new matrix for writing
max = dim(ee)[1]
pb <- winProgressBar(title = "progress bar", min = 0,max , width =
300)
for (i in 1:max) {
```

```
 if (ee$Shire[i]!= prev_shire) {
       # do the average rainfall of the start to finish
       #avg_rain02=total_rain02/no
       #avg_rain03=total_rain03/no
       #avg_rain05=total_rain05/no
       avg_rain=
       # write out the line
       print(ee$SHIRE_NAME[ii])
       print(no)
       print(avg_rain02)
       print(avg_rain03)
       print(avg_rain05)
       # reset the counters
       j=j+1
       dd$shire[j]=paste(prev_shire)
       dd$no_rows[j]=no
       dd$avg_rain02[j]=avg_rain02
       dd$avg_rain03[j]=avg_rain03
       dd$avg_rain05[j]=avg_rain05
       dd$Lupins02[j]=ee$X02.LUPINS[ii]
       dd$Oats02[j]=ee$X02.OATS[ii]
       dd$Barley02[j]=ee$X02.BARLEY[ii]
       dd$Canola02[j]=ee$X02.CANOLA[ii]
       dd$Wheat02[j]=ee$X02.WHEAT[ii]
       dd$Lupins03[j]=ee$X03.LUPINS[ii]
       dd$Oats03[j]=ee$X03.OATS[ii]
       dd$Barley03[j]=ee$X03.BARLEY[ii]
       dd$Canola03[j]=ee$X03.CANOLA[ii]
       dd$Wheat03[j]=ee$X03.WHEAT[ii]
       dd$Lupins05[j]=ee$X05.LUPINS[ii]
       dd$Oats05[j]=ee$X05.OATS[ii]
       dd$Barley05[j]=ee$X05.BARLEY[ii]
       dd$Canola05[j]=ee$X05.CANOLA[ii]
       dd$Wheat05[j]=ee$X05.WHEAT[ii]
       # reset the counters
       total_rain02=0;   total_rain03=0;   total_rain05=0;   no=0
       avg_rain02=0; avg_rain03=0;   avg_rain05=0
       filename=paste(ee$SHIRE_NAME[ii],"csv",sep=".")
       write.csv(ff,filename)
       rm(ff)
       ff=matrix(,nrow=1,ncol=147)
       ff=as.data.frame(ff)
       colnames(ff)=colnames(ee)
       names(ff)=names(ee)
       no=0
 }

 total_rain02=total_rain02 + ee$Rain02[i]
 total_rain03=total_rain03 + ee$Rain03[i]
 total_rain05=total_rain05 + ee$Rain05[i]

 ii=i
 no=no+1
 prev_shire=ee$SHIRE_NAME[i]
 ff[no,] = ee[i,]
 ff$SHIRE_NAME[no]= as.character(ee$SHIRE_NAME[i])
 ff$nvis.Lvl.3[no] = as.character(ee$nvis.Lvl.3[i])
 ff$gridClimate_1000m_FullSet_Prod_soils_south_DECODE[no]=as.char
acter(ee$gridClimate_1000m_FullSet_Prod_soils_south_DECODE[i])
 ff$TLU_DESC[no]=as.character(ee$TLU_DESC[i])
```

```
 Sys.sleep(0.000000001)
    setWinProgressBar(pb, i, title=paste( round(i/max*100, 0),
                                          "% done count ",i))

}
close(Campbell et al.)
        avg_rain02=total_rain02/no
        avg_rain03=total_rain03/no
        avg_rain05=total_rain05/no
        # write out the line
        print(ee$SHIRE_NAME[ii])
        print(no)
        print(avg_rain02)
        print(avg_rain03)
        print(avg_rain05)
        # reset the counters
        j=j+1
        dd$shire[j]=paste(prev_shire)
        dd$no_rows[j]=no
        dd$avg_rain02[j]=avg_rain02
        dd$avg_rain03[j]=avg_rain03
        dd$avg_rain05[j]=avg_rain05
        dd$Lupins02[j]=ee$X02.LUPINS[ii]
        dd$Oats02[j]=ee$X02.OATS[ii]
        dd$Barley02[j]=ee$X02.BARLEY[ii]
        dd$Canola02[j]=ee$X02.CANOLA[ii]
        dd$Wheat02[j]=ee$X02.WHEAT[ii]
        dd$Lupins03[j]=ee$X03.LUPINS[ii]
        dd$Oats03[j]=ee$X03.OATS[ii]
        dd$Barley03[j]=ee$X03.BARLEY[ii]
        dd$Canola03[j]=ee$X03.CANOLA[ii]
        dd$Wheat03[j]=ee$X03.WHEAT[ii]
        dd$Lupins05[j]=ee$X05.LUPINS[ii]
        dd$Oats05[j]=ee$X05.OATS[ii]
        dd$Barley05[j]=ee$X05.BARLEY[ii]
        dd$Canola05[j]=ee$X05.CANOLA[ii]
        dd$Wheat05[j]=ee$X05.WHEAT[ii]
        # reset the counters
        total_rain02=0;   total_rain03=0;   total_rain05=0;   no=0
        avg_rain02=0; avg_rain03=0;   avg_rain05=0
        filename=paste(ee$SHIRE_NAME[ii],"csv",sep=".")
        write.csv(ff,filename)
        rm(ff)
        ff=matrix(,nrow=1,ncol=147)
        ff=as.data.frame(ff)
        colnames(ff)=colnames(ee)
        names(ff)=names(ee)
        no=0
write.csv(dd,choose.files())
```

# REFERENCES


Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459.

Abdullah, A. (2009). Analysis of mealybug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool *Computers and Electronics in Agriculture, 69*(1), 59-72.

Abdullah, A., & Ansari, I. A. (2005a). *Discovery of cropping regions due to Global Climatic Changes using Data Mining.* Paper presented at the 3rd International Symposium on Intelligent Information Technology in Agriculture, Beijing.

Abdullah, A., & Ansari, I. A. (2005b, October). *Discovery of cropping regions due to Global Climatic Changes using Data Mining.* Paper presented at the CAIR Publications Beijing.

Abdullah, A., Brobst, S., Pervaiz, I., Umer, M., & Nisar, A. (2004). *Agri Data Mining/Warehousing: Innovative Tools for Analysis of Integrated Agricultural & Meteorological Data*. Paper presented at the The IASTED International Conference on Databases and Applications (DBA 2004).

Abdullah, A., Brobst, S., Umer, M., & Khan, M. F. (2004). The case for an Agri Data Warehouse: enabling analytical exploration of integrated agricultural data.

Abe, M., & Smith III, J. O. (2004). *Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolation*. Stanford, CA: Stanford University.

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture, 44*, 71-91.

Agarwal, R., & Joshi, M. V. (2000). *PNrule: A new framework for learning classifier models in data mining*. New York: University of Minnesota.

Akoka, J., Berti-Equille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., et al. (2007). *A Framework for Quality Evaluation in Data Integration Systems*. Paris: ESSEC France.

Allum, N., Sturgis, P., Tabourazi, D., & Brunton-Smith, I. (2008). Science Knowledge And Attitudes Across Cultures: A Meta-Analysis. *Public Understanding of Science, 17*(1(2008)), 35-54.

Alur, R., Madhusudan, P., & Nam, W. (2005). Symbolic Compositional Verification by Learning Assumptions *Lecture Notes in Computer Science, 3576*(2005,), 289-292.

Andersen, J. R. (1979). Impacts of climatic variability on Australian agriculture: a review. *Marketing and Agricultural Economics 47*, 147-177.

Anderson, W. K. (2010). Closing the gap between actual and potential yield of rainfed wheat. The impacts of environment, management and cultivar. . *Field Crops Research, 116*(1), 14-22.

Anderson, W. K., & Garlinge, J. R. (2000). *The Wheat Book: Principles and Practice*. Perth: Agriculture Western Australia.

Andrienko, N., & Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach: A systematic approach*. Berlin Heidelberg: Springer-Verlag.

Apaydin, H., Sonmez, F. K., & Yildirim, Y. E. (2004). Spatial interpolation techniques for climate data in the GAP region in Turkey. *Climate Research,, 28*(1), 31-40.

Armstrong, L., Diepeveen, D., & Vagh, Y. (2007). *Data mining can empower growers' crop decision making*.

Asseng, S. (2010, 26 March 2010). *Climate adaptation for Western Australian agriculture.* Paper presented at the Climate Science Update 2010, South Perth.

Asseng, S., Foster, J., & Turner, N. C. (2011). The impact of temperature variability on wheat yields. *Global Change Biology, 17*(2), 997-1012

Asseng, S., & Pannell, D. J. (2012). Adapting dryland agriculture to climate change: Farming implications and research and development needs in Western Australia. *Climatic Change*, 1-15.

Atwood, D. A. (1991). Aggregate food-supply and famine early warning. *Food Policy, 16*, 245–251.

Audi, R. (2011). *Epistemology: a contemporary introduction to the theory of knowledge*. New York: Routledge.

Baesens, B. (2009). Data Mining: New Trends, Applications and Challenges. *Review of Business and Economic, 1*(1), 46-61.

Bakewell, O., & Garbutt, a. (2005). *The Use and Abuse of the Logical Framework Approach*: International NGO Training and Research Centre.

Barandela, R., Valdovinos, R. M., Sanchez, J. S., & Ferri, F. (2004). The Imbalanced Training Sample Problem: Under or over Sampling. *Computer and Information Science, 3138*(Training (2004)), 806-814.

Basso, B., Richie, J. T., Pierce, F. J., Braga, R. P., & Jones, J. W. (2001). Spatial validation of crop models for precision agriculture. *Agric. Syst., 68*, 97–112.

Batts, G. R., Morison, J., Ellis, R. H., Hadley, P., & Wheeler, T. R. (1997). Effects of $CO_2$ and temperature on growth and yield of crops of winter wheat over four seasons. *European Journal of Agronomy, 7*, 43-52.

Bayardo, R. J., Ma, Y., & Srikant, R. (2007). *Scaling Up All Pairs Similarity Search.* Paper presented at the WWW 2007, Banff, Alberta, Canada.

Beard, D. A., Gray, D. M., & Carmody, P. (2010). Farmers' management of seasonal variability and climate change in WA. *Crop Updates, 2010*(141).

Bell, D. E., Raiffa, H., & Tversky, A. (1995). *Decision Making: Descriptive, normative and prescriptive interactions*. Cambridge: Cambridge University Press.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques In *Grouping Multidimensional Data* (pp. 25). Berlin Heidelberg.

Berry, et., & al. (1997). *Data Mining Techniques*: John Wiley and Sons inc.

Berzal, F., Blanco, I., Cubero, J. C., & Marin, N. (2002). Component-based Data Mining Frameworks. *Communications of the ACM, 45*(12).

Besemer, et, & al. (2004). *Improving Coherence in Agricultural Information Systems*. Rome: Food and Agriculture Organization of the United Nations (FAO).

Bhattacharyya, R., & Bhattacharyya, B. (2007). Confidence Association Mining Without Support Pruning *Pattern Recognition and Machine Intelligence, 4815*(2007), 332-340.

Bocchia, S., & Castrignanò, A. (2007). Identification of different potential production areas for corn in Italy through multitemporal yield map analysis *Field Crops Research, 102*(3, 20 June 2007), 185-197.

Bolstad, P. (2005). *GIS Fundamentals*. White Bear Lake : MN: Eider Press.

Boulos, M. N. K., & Honda, K. (2006). Web GIS in practice IV: publishing your health maps and connecting to remote WMS sources using the Open Source UMN MapServer and DM Solutions MapLab. *International Journal of Health Geographics, 5*(6), 1-7.

Boumaa, J., Stoorvogela, J., et, & al. (1999). Pedology, Precision Agriculture, and the Changing Paradigm of Agricultural Research *Soil Science Society of America Journal*(63), 1763-1768.

Boussaid, O., Tanasescu, A., Bentayeb, F., & Darmont, J. (2007). Integration and dimensional modeling approaches for complex data warehousing. *J Glob Optim, 2007*(37), 571-591.

Brand, M. (2010). *Analysis avoidance techniques of malicious software.* Edith Cowan University, Perth.

Bransford, J. D., Sherwood, I. L. D., & et.al. (1990). Anchored instruction: Why we need it and how technology can help. In D. Nix & I. Spiro (Eds.), *Cognition, education and multimedia: Exploring ideas in high technology* (pp. 115-141). HiUsdale, NJ: Lawrence Erlbaum.

Brennan, L. E., Hochman, Z., McCown, R. L., Darbas, T. M., Carberry, P. S., Fisher, J. R., et al. (2007). *Using Computer-based Technologies to Support Farmers' Decision Making*.Unpublished manuscript.

Brighton, H., & Mellish, C. (2002). Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery, 6*(2002), 153-172.

Brooks, R. J., Semanov, M. A., & Jamieson, P. D. (2001). Simplifying Sirus: Sensitivity analysis and development of a meta-model for wheat yield prediction. *European Journal of Agronomy, 14*, 43–60.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(2), 32-42.

Brown, M. L., & Kros, J. F. (2003). Data Mining and the Impact of Missing Data. *Industrial Management & Data Systems, 103*(8), 611-621.

Brown, R. A., & Rosenberg, N. J. (1997). *Agric. Forest. Meteor, 83*, 171-203.

Brunner, R. D. (2006). A paradigm for practice. *Policy Sciences, 2006*(39), 135-167.

Burgess, P., & Lamond, M. (2010). National Variety Trials. Retrieved from http://www.nvtonline.com.au/state-news-wa.htm

Camara, G., Monteiro, A., et., & al. (2000). Towards a Unifying Framework for Geographic Data Models. *Computers and Graphics, 20*, 390-403.

Campbell, R., Crowley, P., & Demura, P. (1983). Impact of drought on national economy and employment. *Quarterly Review of the Rural Economy, 5*(Bureau of Agricultural Economics, Canberra, Australia), 254-257.

Cano, J. R., Herrera, F., & Lozano, M. (2006). On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining *Applied Soft Computing, 6*(3 March 2006), 323-332.

Cantelaub, P., & Terres, J. M. (2004). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A*(57A), 476-487.

Cao, L., & Zhang, C. (2007). The Evolution of KDD: Towards Domain-Driven Data Mining. *International 1 Journal of Pattern Recognition and Artificial Intelligence, 21*(3 (2007)), 1-16.

Cao, L. J., Keerthi, S. S., Ong, C., Zhang, J. Q., Periyathamby, U., Fu, X., et al. (2006). Parallel Sequential Minimal Optimization for the Training of Support Vector Machines. *IEEE TRANSACTIONS ON NEURAL NETWORKS, 17*(4), 1039-1049.

Carter, L. M. (2000). *Arguments in hypertext: a rhetorical approach.* Paper presented at the Proceedings of the eleventh ACM on Hypertext and hypermedia, San Antonio, Texas, United States.

Casa, R., & Castrignanò, A. (2008). Analysis of spatial relationships between soil and crop variables in a durum wheat field using a multivariate geostatistical approach *European Journal of Agronomy, 28*(3, April 2008), 331-342.

Castrignanò, A. (2010). Different Approaches To Delineate Management Zones Perth: CRA.

Cereghini, P. M., & Ordonez, C. (2002). United States Patent No.: U. S. Patent.

Cfar. (1998). Data Mining for Site-specific Agriculture Retrieved May,22, 2006, from http://www.gis.uiuc.edu/cfardatamining/default.htm

Challinor, A. J., Slingo, J. M., Wheeler, T. R., Craufurd, P. Q., & Grimes, D. I. F. (2003). *Toward a Combined Seasonal Weather and Crop Productivity Forecasting System: Determination of the Working Spatial Scale*.Unpublished manuscript.

Chase, J. M., & Leibold, M. A. (2002). Spatial scale dictates the productivity-biodiversity relationship. *Nature, 416*(6879), 427-430.

Chien, C. F., & Chena, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry *Expert Systems with Applications, 34*(1, January 2008), 280-290.

Chris, S., & Coryn, L. (2007). The Holy Trinity of Methodological Rigor: A Skeptical View. *Journal of MultiDisciplinary Evaluation, 4*(7), 26-31.

Clark, D. (2009). Understanding and Performance. Retrieved February,8, 2010

Cleveland, H. (1982). Information as Resource. *The Futurist, December, 1982,* 34-39.

Cohen, J. (2004). Bioinformatics—An Introduction for Computer Scientists. *36*(2), 122-158.

Collins, A. (1988). *Cognitive apprenticeship and instructional technology*. Cambridge, MA.: BBN Labs Inc.

Congalton, R. G. (1997). Exploring and Evaluating the Consequences of Vector-to-Raster and Raster-toVector Conversion. *American Society for Photogrammetry and Remote Sensing, 63*(4), 424-435.

Cooper, M., Brennan, P. S., & Sheppard, J. A. (1996). A strategy for yield improvement of wheat which accommodates large genotype by environment interactions. In M. Cooper & G. L. Hammer (Eds.), *Plant Adaptation and Crop Improvement* (pp. 487 - 511): CAB International, UK.

Couclelis, H. (1992). People Manipulate Objects: Beyond the Raster-Vector Debate in GIS. *Springer-Verlag New York, 639*, 65-67.

Craik, W., & MacRae, A. (2010). *Wheat Export Marketing Arrangements*. Melbourne.

Creswell, J. W. (1994). *Research Design :Qualitative and Quantitative Approaches*. London, UK: Sage Publications.

Cunningham, S. J., & Holmes, G. (2001). Developing innovative applications in agriculture using data mining. *SEARCC Journal*.

D'Inverno, M., Justo, G., et., & al. (1996). A formal framework for specifying design methodologies. *Software Process Improvement Pract., 2*, 181-195.

DA Roshier, P. W., RJ Allan, AI Robertson (2001). Distribution and persistence of temporary wetland habitats in arid Australia in relation to climate. *Austral Ecology, 2001*(26), 371-384.

Dadax. (2013). Worldometers: Real Time World Statistics. Retrieved 9 February 2013, 2013, from http://www.worldometers.info

DAFWA. (2009). Tools To Assist Decision-Making - Details. *Department of Agriculture and Food - Farm Systems*. Retrieved from http://www.agric.wa.gov.au/PC_92630.html

Darmont, J., Boussaid, O., Ralaivao, J. C., & Aouiche, K. (2005). *An architecture framework for complex data-warehouses.* Paper presented at the 7th International Conference, DaWaK 2005, Copenhagen.

DeFalco, I., Della-Cioppa, A., & et.al. (2005). An evolutionary approach for automatically extracting intelligible classification rules. *Knowledge and Information Systems, (2005)*(7), 179-201.

deHaro-Garcia, A., delCastillo, J. A. R., & Garcia-Pedrajas, N. (2009). Scaling up instance selection algorithms by dividing and conquering. *Data Mining and Knowledge Discovery, 18*(3), 392-418.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of Qualitative Research*. London, UK: Sage Publications.

Deressa, T. T. (2007). *Measuring The Economic Impact of Climate Change on Ethiopian Agriculture: Ricardian Approach*: The World Bank.

Drew, J. (2010). Operating In A Change Environment – NEAR/Drought Reform. Retrieved from http://www.agric.wa.gov.au/objtwr/imported_assets/content/lwe/cli/near_summary.pdf

Duckworth, W. M., & Stephenson, W. R. (2002). Beyond Traditional Statistical Methods. *American Statistical Association, 56*(3).

DuDoit, S., Gentleman, R. C., et., & al. (2003). Open Source Software for the Analysis of Microarray Data. *BioTechniques, 34*(March 2003), 45-51.

Dunstan, D. (2009). *Hierarchies of sustainability in a catchment.* Paper presented at the 4th International Conference on SustainableDevelopment and Planning, Cyprus.

Dunstan, D., Despi, I., & Watson, C. (2009). *Anomalies in multidimensional contexts.* Paper presented at the 10th International Conference on Data Mining, Protection, Detection and other Securities, Crete.

Durieux, L., Lagabrielle, E., & Nelson, A. (2008). A method for monitoring building construction in urban sprawl areas using object-based analysis of Spot 5 images and existing GIS data. *International Journal of Photogrammetry and Remote Sensing, 63*(4), 399-408.

Eakin, H. (1999). Seasonal climate forecasting and the relevance of local knowledge. *Physical Geography, 20*, 447–460.

Ekasingh, B. S., Ngamsomsuke, K., Letcher, R. A., & Spate, J. M. (2005). *A Data Mining approach to simulating land use decisons: Modeling farmer's crop choice from farm level data for integrated water resource management.* Paper presented at the Proceedings of the 2005 International Conference on Simulation and Modeling.

Fernandez, G. (2003). *Data Mining using SAS applications*. Boca Raton:Florida: Chapman & Hall/CRC.

Foulkes, M. J., Slafer, G. A., Davies, W. J., Berry, P. M., Sylvester-Bradley, R., Martre, P., et al. (2011). Raising yield potential of wheat. *Journal of Experimental Botany, 62*(2), 469-486.

Francis, L. A. (2005). *Dancing With Dirty Data*. Paper presented at the Casualty Actuarial Society Forum.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics, 20*(15), 2479-2481.

Frédérique Lisacek, F., Cohen-Boulakia, S., & Appel, R. D. (2006). Proteome informatics II: Bioinformatics for comparative proteomics. *Proteomics 2006, 6*(20), 5445–5466.

Freitas, A. A. (2000). Understanding the crucial differences between classification and discovery of association rules. *SIGKDD Explorations, 2*(1), 65-69.

Fu, X., & Wang, L. (2005). Data dimensionality reduction with application to improving classification performance and explaining concepts of data sets. [Refereed]. *Int. J. Business Intelligence and Data Mining, 1*(1), 65–87.

Garcia, R. E., DeOliveira, F. M. C., Maldonado, J. C., & Mendonc, M. (2004). *Visual Analysis of Data from Empirical Studies*. Paper presented at the International Workshop on Mining Software Repositories.

Garlin, D. (1990). The role of formal reusable frameworks. *ACM SIGSOFT:Software Eng. Notes, 15*, 42-44.

Garlin, D., & Notkin, D. (1991). *Formalising design spaces: Implicit innovation mechanisms*. Paper presented at the Formal Software Development Methods : Int Symposium VDM Europe.

Gentleman, R. C., & Ihaka, R. (2010). The R Project for Statistical Computing. *R Journal*

Geunon, R. (2004). *Symbolism of the Cross*. New York: Sophia Perennis.

Godwin, R. J., & Miller, P. C. H. (2003). A review of the technologies for mapping within-field variability. *Biosystems Engineering, 84*, 393-407.

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *SIGKDD Explorations, 1*(1), 20-33.

Gold, J. M., Mundy, P. J., & Tjan, B. S. (2012). The Perception of a Face Is No More Than the Sum of Its Parts. *Psychological Science April 2012 vol. 23* (4), 427-434

Goodchild, M. (1992). Geographic Data Modelling. *Computers and Geosciences, 18*, 401-408.

Gooding, M. J., Ellis, R. H., Shewry, P. R. and Schofield, J. D. (2003). Effects of restricted water availability and increased temperature on the grain filling, drying and quality of winter wheat. *Journal of Cereal Science 37*(3), 295-309.

Gray, J. (2008). Quantum GIS: the Open-Source Geographic Information System *Linux Journal, 2008*(172).

Gray, J., Chaudrhuri, A., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., et al. (1997). Data cube: a relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery, 1*, 29-54.

Greenfield, J., & Short, K. (2003, October 26–30). *Software Factories :Assembling Applications with Patterns, Models, Frameworks and Tools.* Paper presented at the OOPSLA'03, Anaheim, California, USA.

Guarneri, F., Vaccaro, M., & Gaurneri, C. (2008). Digital Image Compression in Dermatology: Format Comparison. *TeleMedicine and e-Health, 14*(7), 666-670.

Guba, E. C. (1990). The Alternative Paradigm Dialog. In *The Paradigm Dialog*: Sage Publications.

Gulati, A., Joshi, P. K., & Cummings, R. (2007). The way forward: Towards greater agricultural diversification and greater participation of smallholders. In *Agricultural diversification and smallholders in South Asia*. New Delhi: Academic Foundation.

Habersack, H. M. (2000). The River Scaling Concept (RSC) : a basis for ecological assessments. *Hydrobiologia, 422-423*(0), 49-60.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1), 10-18.

Han, J., & Kamber, M. (2011). *Data mining: concepts and techniques* San Francisco: Morgan Kaufman Publishers.

Han, J., Nishio, S., et., & al. (1998). Generalization-based data mining in object-oriented databases using an object cube model *Data & Knowledge Engineering, 25*(1-2), 55-97.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Massachusetts: MIT Press.

Hayman, P. (2004). *Decision support systems in Australian dryland farming: A promising past, a disappointing present and uncertain future* Paper presented at the Australian Agronomy Conference.

He, H. S., Dey, D. C., Fan, X., Hooten, M. B., Kabrick, J. M., Wickle, C. K., et al. (2007). Mapping pre-European vegetation at fine resolutions using a heirarchical Bayesian model and GIS. *Plant Ecology, 2007*(191), 85-94.

Hernon, P. (1991). The Elusive Nature of Research LIS. In C. R. McCLure & P. Hernon (Eds.), *Library and Information Science Research*. USA: Ablex Publishing Corporation.

Hernon, P. (2009). *What Really Are Student Learning Outcomes?* Paper presented at the ACRL Fourteenth National Conference.

Herrington, J., & Oliver, R. (2000). An Instructional Design Framework for Authentic Learning Environments. *ETR&D, 48*(3), 23-48.

Herrington, J., Oliver, R., & Herrington, A. (2007). Authentic Learning on the Web: Guidelines. Flexible learning in an information society. In (pp. 26-35). Hershey PA: Information Science Publishing.

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications : a comprehensive reference for science*. Tulsa, Oklahoma.

Hoadley, C. M. (2004). Methodological Alignment in Design-Based Research. *Educational Psychologist, 39*(4), 203–212.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics 24*(2), 225-232.

Howden, S. M., Soussana, J. F., Tubiello, F. N., Chhetri, N., Dunlop, M., & Meinke, H. (2007). Adapting agriculture to climate change. *Proceedings of the National Academy of Sciences of the United States of America 104*(50), 19691-19696.

Hsu, J. (2002). *Data Mining Trends And Developments : The Key Data Mining Technologies and Applications for the 21st Century.* Paper presented at the ISECON, San Antonio.

Hughes, L. (2003). Climate change and Australia: Trends, projections and impacts. *Austral Ecology, 28*(4), 423-443.

Hunter, J., & Cheung, K. (2005). *Generating eScience Workflows from Statistical Analysis of Prior Data*. Paper presented at the APAC'05.

Illian, J., Penttinen, A., et., & al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. West Sussex: Wiley.

Ines, A. V. M., & Hansen , J. W. (2006). Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and Forest Meteorology, 138*(1-4), 44-53.

Inmon, W. H. (1996). The Data Warehouse and Data Mining. *Communications of the ACM, 39*(11), 49-50.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. [Review]. *ACM Computing Surveys, 31*(3), 265-321.

John, M., Pannell, D., & Kingwell, R. (2005). Climate Change and the Economics of Farm Management in the Face of Land Degradation: Dryland Salinity in Western Australia. *John, M., Pannell, D., & Kingwell, R. (2005). Climate change and the economics of farm management in the face of land degradation: dryland salinity in Western Australia. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, , 53*(4), 443-459.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer.

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *Transactions on Visualization and Computer Graphics, 7*(1), 100-107.

Keim, D. A., & Hermann, A. (1998). *The gridfit algorithm: an efficient and effective approach to visualizing large amounts of spatial data.* Paper presented at the IEEE Visualization, Research Triangle Park, NC.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Zeigler, H. (2008). Visual Analytics: Scope and Challenges. In S. J. Simoff, M. H. Böhlen & A. Mazeika (Eds.), *Visual data mining: theory, techniques and tools for visual analytics* (pp. 76-90). Berlin, Heidelberg: Springer-Verlag.

Keim, D. A., Panse, C., Sips, M., & North, S. C. (2004). Pixel based visual data mining of geo-spatial data *Computers & Graphics, 28*(3), 327-344.

Kim, S., Song, Y., Kim, K., & Lee, G. G. (2006). *MMR-based Active Machine Learning for Bio Named Entity Recognition.* Paper presented at the Human Language Technology Conference of the North American Chapter of the ACL, New York.

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2001). A Taxonomy of Dirty Data. [classification]. *Data Mining and Knowledge Discovery,, 2003*(7), 81-99.

Kimball, R. (1997). A dimensional modeling manifesto. *DBMS and Internet Systems, August 1997*.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit*. New York: John Wiley & Sons.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2011). *The data warehouse lifecycle toolkit*: Wiley.

Kitchenham, B., Linkman, S., & Linkman, S. (2005). Experiences of using an evaluation framework. *Information and Software Technology, 47*(2005), 761-774.

Kittler, J. (1998). Combining Classifiers: A Theoretical Framework. *Pattern Analysis & Applic., 1998*(1), 18-27.

Konen, W. (1999). Learning to Generalize. In G. E. Hinton & T. J. Sejnowski (Eds.), *Unsupervised Learning: Foundations of Neural Computation*: Library of Congress.

Koperski, K., Han, J., & Adhikary, J. (1999). Mining Knowledge in Geographic Data. *Communications of ACM, 1998*.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Miining : A Recent Overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71-82.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data Preprocessing for Supervised Leaning. *International Journal of Computer Science, 1*(1 2006), 111-117.

Kovalerchuk, B. (2004). Decision Process and its Visual Aspects. In K. B & J. Schwing (Eds.), *Visual and spatial analysis: advances in data mining, reasoning, and problem* Netherlands: Springer.

Krauss, S. E. (2005). Research Paradigms and Meaning Maker: A Primer. *The Qualitative Report, 10*(4), 758-770.

Kuba, P. (2001). *Data Structures for Spatial Data Mining*: FIMU.

Labib, K., & Vemuri, V. R. (2006). An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks. *Journal of Telecommunications*.

Lauer, J. (1995). Software SELECT : Crop Variety Selection Software for Microcomputers. *Production Agriculture, 8*(3), 327-437.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leduc, T., Bocher, E., Fernando, G. C., & Moreau, G. (2009). GDMS-R: A mixed SQL to manage raster and vector data *GIS Ostrava 2009*(1), 25-28.

Li, C., & Maguire, D. (Eds.). (2003). *The handheld revolution: towards ubiquitous*. Redlands:California: ESRI.

Li , D., Di, K., & Li , D. (2000). Land Use Classification Of Remote Sensing Image With GIS Data Based On Spatial Data Mining Techniques. *International Archives of Photogrammetry and Remote Sensing, 33*(B3), 238-245.

Li, D., & Wang, S. (2005, August 27-29). *Concepts, Principles And Applications Of Spatial Data Mining And Knowledge Discovery.* Paper presented at the ISSTM, Beijing, China.

Li, D., & Wang, S. (2008). Spatial data mining and knowldge discovery. In T. et.al. (Ed.), *Advances in Spatio-Temporal Analysis*. London: Taylor & Francis Group.

Li, J. (2006). On Optimal Rule Discovery. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 18*(4), 460-471.

Li, Y., Li, Z., Chen, Y., Li, X., & Lin, Y. (2003). Raster space with relativity. *Kybernetes, 32*(5/6, 2003), 629-639.

Li, Z., Dunham, M. H., & Xiao, Y. (2003). STIFF: A Forecasting Framework for Spatio-Temporal Data. *Mining Multimedia and Complex Data, LNAI*(2797), 183-198.

Liao, S. H., & Wen, C. H. (2007). Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005. *Expert Systems with Applications, 32*(1), 1-11.

Lincoln, Y. S., Lynham, S. A., & Guba, E. G. (2011). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In L. Y. Denzin HK (Ed.), *Handbook of qualitative research*. USA: SAGE Publications, Inc.

Liu, H., & Motoda, H. (2002). On Issues of Instance Selection. *Data Mining and Knowledge Discovery,, 2002*(6), 115-130.

Liu, L., Y, C., Shan, S., & Yin, L. (2008). *Mining Condensed and Lossless Association Rules by Pruning Redundancy.* Paper presented at the FSKD '08 Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery Washington, DC, USA

Liu, L. M., Bhattacharyya, S., Sclove, S. L., Chen, R., & Lattyak, W. J. (2001). Data Mining On Time Series: An Illustration Using Fast-Food Restaurant Franchise Data. (Publication.:

Liu, Y., & Goodchild, M. (2008). Towards a general Field Model and its order in GIS. *International Journal of Geographical Information Science, 22,*(6), 623-643.

Lobell, D. B., & Field, C. B. (2007). Global scale climate crop yield relationships and the impacts of recent warming. *Environ. Res. Lett, 2*.

Longley, P., Goodchild, M., et., & al. (2005). *Geographic information systems and science* West Sussex: John Wiley & Sons Ltd.

Lu, K., Jong, K., Rajasekaran, A., Cloughesy, T., & Mischel, P. (2004). Upregulation of tissue inhibitor of metalloproteinases (TIMP)-2 promotes matrix metalloproteinase (MMP)-2 activation and cell invasion in a human glioblastoma cell line. *Lab Invest, 84*, 8 - 20.

Lucchese, C., Orlando, S., & Perego, R. (2006). Fast and Memory Efficient Mining of Frequent Closed Itemsets. *IEEE Transactions on Knowledge and Data Engineering, 18*(1), 21-36.

Luck, M., & D'Inverno, M. (2001). A Conceptual Framework for Agent Definition and Development. *The Computer Journal, 44*(1), 1-20.

Luger, G. F. (2005). *Artificial Intelligence* (5th ed.). London: Addison Wesley.

Luo, L. (2012). Value of the Research Methods Course: Voices from LIS Practitioners. *SLIS Student Research Journal, 2*(1).

Lymann Ott, R., & Longnecker, M. (2010). *An Introduction to Statistical Methods and Data Analysis*. Belmont, CA.

Ma, C., Chou, D. C., & Yen, D. C. (2000). Data warehousing, technology assessment and management. *Industrial Management & Data Systems, 100*(3), 125-135.

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues In Educational Research, 16*(2006).

Maliappis, M. T. (2006). Technological Aspects of Using Agricultural Ontologies.

Manly, B. F. J. (2005). *Multivariate statistical methods: a primer, Volume 2004* Boca Raton ,Florida: Chapman & Hall/CRC.

Mannila, H. (2000). Theoretical Frameworks for Data Mining. *ACM SIGKDD 1*(2), 30-32.

Manouselis, N., Kastrantas, K., & Tzikopoulos, A. (2006). An IEEE LOM application profile to describe training resources for agricultural & rural SMEs. *Informatics Library*,

March, J. G. (1983). *Rationality, ambiguity and the engineering of choice*. Paper presented at the Decision making: Normative, Descriptive and Prescriptive Interactions.

Marinho, T., Costa, E. B., Dermeval, D., Ferreira, R., Braz, L. M., Bittencourt, I. I., et al. (2010). An ontology-based software framework to provide educational data mining. *SAC'10*, 1433-1437.

Markovic, N., Stanimirov, A., & Stoimenov, L. (2009). *Sensor Web for River Water Pollution Monitoring and Alert System*. Paper presented at the 12th Agile International Conference on Geographic Information Science.

Martin, R. V., Washington, R., & Downing, T. E. (2000). Seasonal maize forecasting for South Africa and Zimbabwe derived from an agroclimatological model. *Journal of Applicable Meteorology, 39*, 1473–1479.

Matthews, P., & McCaffery, D. (2012). Wheat Variety Guide for WA 2012.

McBratney, A., Whelan, B., Ancev, T., & Bouma, J. (2005). Future Directions of Precision Agriculture. *Precision Agriculture, 6*, 7-23.

McBrien, P., & Poulovassilis, A. (2001). *A Semantic Approach to Integrating XML and Structured Data Sources.* Paper presented at the 13th International Conference on Advanced Information Systems Engineering (CAiSE 01).

McInerny, J. (2002). The production of food: from quantity to quality. *Proceedings of the Nutrition Society, 61*(02), 273-279.

McLellan, H. (1993). Evaluation in a situated learning environment. *Educational Technology Research and Development, 33*(3), 39-45.

Mearns, L., Rosenweig, O. C., & R. Goldberg, R. (1997). Mean and variance change in climate scenarios: Methods, agricultural applications, and measures of uncertainty. *Climatic Change, 35*, 367–396.

Michalski, R. S. (2000). Learnable Evolution Model: Evolutionary processes guided by Machine Learning. *Machine Learning, 38*(2000), 9-40.

Mielikainen, T. (2003). *An Automata Approach to Pattern Collections.* University of Helsinki, Finland.

Miller, D. C., & Salkind, N. J. (2001). *Handbook of Research Design*. London: SAGE.

Miller, H. J., & Wentz, E. A. (2003). Representation and Spatial Analysis in Geographic Information Systems. *Annals of the Association of American Geographers, 93*(3), 574–594.

Miller, J., & Han, J. (2009). *Geographic data mining and knowledge discovery* (2nd ed.). New York: Taylor & Francis Inc.

Mitra, S., & Acharya, T. (2003). *Data Mining: Multimedia, Soft Computing, and Bioinformatics*. New Jersey: John Wiley & Sons.

Moore, A. D., Angus, J. F., Bange, M., Crispin, C. J., Donnelly, J. R., Freer, M., et al. (2004). *Computer-based decision support tools for Australian farmers.* Paper presented at the Australian Agronomy Conference.

Morse, J. M., Hupcey, J. E., Penrod, J., Spiers, J. A., Pooler, C., & Mitcham, C. (2002). Issues of validity: Behavioral concepts, their derivation and interpretation. *International Journal of Qualitative Methods, 1*(4).

Morse, J. M., & Mitcham, C. (2002). Exploring qualitatively dervied concepts: Inductivedeductive methods. *International Journal of Qualitative Methods, 1*(4).

Mosley, R. (2005). Detecting A Pattern. *Best's Review, 106*(1), 68-70.

Murray, R. (2012). Short-Range Weather Forecasting. *Weather, 42*(11), 346-350.

Nazzal, J. M., El-Emary, E. M., & Najim, S. A. (2008). Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale. *World Applied Sciences, 5*(5), 546-552.

Neuman, W. L. (2005). *Social research methods, quantitative and qualitative approaches*: Allyn and Bacon.

Ng, R. T., & Han, J. (2002). CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions on Knowledge and Data Engineering, 14*(5), 1003-1016.

NIAB. (2010). Variety Results. Retrieved from http://www.niab.com/pages/id/245/Variety_Results

Nicholls, N. (1985). Towards the prediction of major Australian droughts. *Australian Meterological Magazine, 33,* 161-166.

Oates, B. (2007). *Researching Information Systems and Computing*. London: Sage Publications.

Ormsby, T., Napoleon, E., Burke, R., Groessl, C., & Feater, L. (2004). *Getting to Know ArcGIS Desktop*. Redlands: California: ESRI Press.

Ossimitz, M. L. (2009). *Data Warehousing - Concepts and Applications in SAP BI*. Bondi Beach, Sydney: Auxilia Pty Ltd.

Papadimitriou, C. H. (2003). *Computational complexity*: John Wiley and Sons Ltd.

Parent, C., Spaccapietra, S., & Zimány, E. (1999). Spatio-Temporal Conceptual Models:Data Structures + Space + Time. *ACM GIS, 11*(99), 26-33.

Parthasarathy, B., Kumar, K. R., & Munot, A. A. (1992). Forecast of rainy-season foodgrain production based on monsoon rainfall. *Indian Journal of Agricultural Science, 62*, 1-8.

Payyappillil, H. (2005). *Data Mining Framework.* Unpublished Masters, West Virginia University, Morgantown, West Virginia.

Perry, D. E., Sim, S., & Easterbrook, S. M. (2004). Case Studies for Software Engineers. *International Conference On Software Engineering, 26*, 736-738.

Phillips-Wren, G. E., Hahn, E. D., & Forgionne, G. A. (2004). A multiple-criteria framework for evaluation of decision support systems. *OMEGA The International Journal of Management Science, 32*(2004), 323- 332.

Phillips, S. J. (2008). Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al.(2007). *Ecography, 31*(2008), 272-278.

Piaget, J. (1970). *Structuralism*. New Yordk: Basic Books.

Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., Kloesgen, W., & Simoudis, E. (1996). *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications.* Paper presented at the KDD-96.

Pimentel, D. (2009). Energy Inputs in Food Crop Production in Developing and Developed Nations. *Energies, 2*(1), 1-24.

Pitman, A. J., Narisma, G. T., Pielke Sr, R. A., & Holbrook, N. J. (2004). Impact of land cover change on the climate of southwest Western Australia. *Journal of Geophysical Research, 109*(18), 1-12.

Pittock, A. B. (1984). On the reality, stability, and usefulness of southern hemisphere teleconnections. *Australian Meterological, 32,* 75-82.

Potgieter, A. B., Hammer, G. L., Meinke, H., Stone, R. C., & Goddard, L. (2005). Three putative types of El Nino revealed by spatial variability in impact on Australian wheat yield. *Journal of Climate, 18*(May 2005), 1566-1573.

Priya, S., & R, S. (2001). National spatial crop yield simulation using GIS-based crop production model. *Ecological Modelling, 136*(2-3), 113-129

Purchase, H. C., Andrienko, N., Jankun-Kelly, T. J., & Ward, M. (2008). Theoretical Foundations of Information Visualization. In A. Kerren, J. T. Statsko, J. D. Fekete & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives.* (pp. 46-64): Springer.

Pyle, D. (1999). *Data Preparation for Data Mining*. London: Academic Press.

Rasmassun, E. M., & Wallace, J. M. (1983). Meterological aspects of the El Nino/Southern Oscillation. *Science, 222*, 1195-1202.

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. *Springer-Verlag Berlin, 2004*, 63-71.

Revesz, P. (2010). Interpolation and Approximation. In *Introduction to Databases* (pp. 435-483 ). London: Springer-Verlag.

Robinson, T. P., & Metternicht, G. (2006). Testing the performance of spatial interpolation techniques for mapping soil properties. *Computers and Electronics in Agriculture 50*(2), 97-108

Roddick, J. F., Hornsby, K., & Spiliopoulou, M. (2001). *An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining Research.* Paper presented at the TSDM 2000, Berlin Heidelberg.

Rodriguez, A., Carazo, J. M., & Trelles, O. (2004). Mining Association Rules from Biological Databases. *Journal of American Society for Information Science and Technology, 56*(5), 493-504.

Rola-Rubzen, M. F., Storer, C., & Pringle, M. (2005). *Which chain is best for me? The case of supply chain for wheat in Western Australia* (Consultancy). Northam, W.A: Curtin University of Technology.

Rozinat, A., de Medeiros, A. K. A., Gunther, C. W., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2008). *Towards an Evaluation Framework for Process Mining Algorithms*. Eindhoven, The Netherlands: Eindhoven University of Technology.

Ruiz, D. A., Becker, K., et, & al. (2005). *A Data Warehousing Environment to Monitor Metrics in Software Development Processes.* Paper presented at the 16th International Workshop on Database and Expert Systems Applications (DEXA'05), Copenhagen.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2008). *Optimizing Wheat Yield Prediction Using Different Topologies of Neural Networks.* Paper presented at the IPMU'08, Malaga.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2009). *Visualization of agriculture data using self-organizing maps.* Paper presented at the Proceedings of AI-2008.

Sadowski, V. (1997). Object-Oriented Application Frameworks. *Communications of the ACM, 40*(10), 32-38.

Savage, L. J. (1972). *The Foundations of Statistics* (Dover ed.). New York: Dover Publications Inc.

Schlenker, W., Roberts, M. J., & Smith , V. K. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of Sciences, 106*(37), 15594-15598.

Schulz, H. J., Nocke, T., & Schumann, H. (2006). *A Framework for Visual Data Mining of Structures*. Paper presented at the Twenty-Ninth Australasian Computer Science Conference (ACSC2006).

Sdorra, P. B., Hafez, A. M., & Raghavan, V. (2001). *A Theoretical Framework for Association Mining basedon the Boolean Retrieval.* Paper presented at the DaWaK 2001, LNCS.

Sean, K. (1997). *Data Warehousing in action*. Brisbane, NY, Chichester: John Wiley & Sons Ltd.

Seifert, J. W. (2006). Data Mining: An Overview. In D. D. Pegarkov (Ed.), *National Security Issues*. New York: Nova Science Publishers Inc.

Semanov, M. A., & Porter, J. R. (1995). Climatic variability and the modelling of crop yields. *Agriculture for Meteorology, 73,*, 265–283.

Sen, A., & Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Communications of ACM, 48*(3), 79-84.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shah, P., & Akhtar, T. (2011). Varietal improvement of rice under rainfed conditio of Parwnipur, Bara, Nepal. *International Research Journal of Applied and Basic Sciences, 2*(11), 423-425.

Shekhar, S., Zhang, P., Huang, Y., & Vatsavai, R. R. (2003). Spatial data mining. In *Data Mining and Knowledge Discovery*.

Shneiderman, B. (1996). *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.* Paper presented at the Proceedings of the IEEE Symposium on Visual Languages.

Simoff, S. J., Bohlen, M. H., & Mazeika, A. (2008). *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Heidelberg: Springer-Verlag Berlin.

Sivakumar, M. V. K. (2006). Climate prediction and agriculture: current status and future challenges. *Climate Research, 33*(2006), 3-17.

Smith, A., Cullis, B., & Gilmour, A. (2001). The Analysis of Crop Variety Evaluation Data in Australia. *Australian and New Zealand Journal of Statistics, 43*(2), 129-145.

Smith, M. B., and S. Davies. (1995). Famine Early Warning and Response:The Missing Link. *Intermediate Technology Publications,* 320.

Stahla, K., Moorea, R. D., Floyer, J. A., Asplina, M. G., & McKendrya, I. G. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agricultural and Forest Meteorology, Volume 139*( 3–4, 12 October 2006), 224–236.

Stefanakis, E., & Prastacos, P. (2008). Development of an open source-based spatial data infrastructure. *Applied GIS, 4*(4), 1-26.

Tabios, G. Q., & Salas, J. D. (1985). A Comparative Analysis of Techniques for Spatial Interpolation of Precipitation. *JAWRA Journal of the American Water Resources Association, 21*(3), 365–380.

Thomas, C. S., Skinner, P. W., Fox, A. D., Greer, C. A., & Gubler, W. D. (2002). Utilization of GIS/GPS-Based Information Technology in Commercial Crop Decision Making in California, Washington, Oregon, Idaho,and Arizona. *Journal of Nematology, 34*(3), 200-206.

Tomlin, C. D. (1994). Map algebra: one perspective. *Landscape and Urban Planning, Elsevier Science, , 30*, 3-12.

Tomlinson, R. F. (2007). *Thinking about GIS*. Redlands: California: Ingram Publishers.

Trnka, M., Dubrovsky, M., Semera´dova, D., & Zalud, Z. (2004). Projections of uncertainties in climate change scenarios into expected winter wheat yields. *Theoretical and Applied Climatology, 77*(2004), 229-249.

Trochim, W. M. K. (2006). Qualitative Measures *Research Measures Knowledge Base*( ), 361-9433.

Tsang, I. W., Kwok, J. T., & Cheung, P. M. (2005). Core Vector Machines: Fast SVM Training on Very Large Data Sets. *Journal of Machine Learning Research, 6*(2005), 363-392.

Ullmer, B., & Ishii, H. (2001). Emerging Frameworks for Tangible User Interfaces. *Human-Computer Interaction in the New Millenium,* 579-601.

Utts, J. (2003). What Educated Citizens Should Know About Statistics And Probabiity. *The American Statistician*(May 2003).

Vagh, Y. (2012). *An investigation into the effect of stochastic annual rainfall on crop yields in South Western Australia.* Paper presented at the International Conference on Knowledge Discovery.

Vagh, Y., Armstrong, L., & Diepeveen, D. (2010). *Application of a data mining framework for the identification of agricultural production areas in WA.* Paper presented at the 14th Pacific-Asia Conference of Knowledge Discovery and Data Mining, Hyderabad, India.

Vagh, Y., & Xiao, J. (2012). Mining temperature profile data for shire-level crop yield prediction. *IEEE, 39*(2), 301-309.

van Gool, D. (2011). *Wheat yield potential and land management constraints in the South West of Western Australia*. Perth: DAFWA.

van Ittersum, M. K., Ewert, F., Heckelei, T., Wery, J., Olsson, J. A., Andersen, E., et al. (2008). Integrated assessment of agricultural systems – A component-based framework for the European Union (SEAMLESS). *Agricultural Systems, 96*(1-3, March 2008), 150-165.

van Schaalk, J. G. J., & van der Kemp, J. (2009). Real Crime on Virtual Maps: The Application of Geography and GIS in Criminology. In H. J. Scholten (Ed.), *Geospatial Technology and the Role of Location in Science*. Amsterdam: Springer.

Vassilios S. Verykios, V. S., Bertino, E., et., & al. (2004). State-of-the-art in Privacy Preserving Data Mining. *SIGMOD, 33*(1), 50-57.

Vaus, D. (2001). What is Research Design. In *Research Design in Social Research*. London: SAGE.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer Science + Business Media Inc.

von Braun, J. (2007). *The World Food Situation: New Driving Forces and Required Actions*. Washiington D. C.: International Food Policy Research Institute.

Wang, J. (2003). *Data mining: opportunities and challenges*. Hershey, PA: IRM Press.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer, 40*(9 ), 96-99.

Wegman, E., & Solka, J. L. (2005). *Statistical Software for Today and Tomorrow*. Vancouver.

Wegman, E. J. (2003). Visual data mining. *Statistics in Medicine, 22*(2003), 1383–1397.

Wheeler, T. R., Craufurd, P. Q., Ellis, R. H., Porter, J. R., & Prasad, P. V. V. (2000). Temperature variability and the annual yield of crops. *Agricultural Ecosystem Environment, 82*, 159–167.

Wiemer, J. C. J. C., & Prokudin, A. (2004). Bioinformatics in proteomics: application, terminology, and pitfalls *Pathology - Research and Practice, 200*(2 April 2004), 173-178

Wilcox, R. R. (2010). *Fundamentals of Modern Statistical Methods*. Los Angeles: Springer.

Witten, I. H., & Frank, E. (2005). *Data Mining. Practical Machine Learning and Techniques*. San Francisco: Morgan Kaufman.

Witten, I. H., Franke, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques

Wong, M. T. F., Corner, R. J., & Cook, S. E. (2001). A decision support system for mapping the site-specific potassium requirement of wheat in the field. *Australian Journal of Experimental Agriculture, 2001*(41), 655-661.

Worboys, M. P., & Duckham, M. (2004). *GIS : a computing perspective*. Boca Raton: Florida: CRC Press.

Xiong, H., & Kumar, V. (2006). Hyperclique pattern discovery. *Data Mining and Knowledge Discovery, 13*(2), 219-242.

Yang, Y., Adelstein, S. J., & Kassis, A. I. (2009). Target discovery from data mining approaches. *Drug Discovery Today, 14*(3-4 Febrary 2009), 147-154.

Yarrow, J., Perlman, Z., Westwood, N., & Mitchison, T. (2004). A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnology, 4*(1), 21.

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Computer Science and Engineeering*.

Yost, M. (2000). Data Warehousing and Decision Support at the National Agricultural Statistics Service. *Social Science Computer Review*(18), 434.

Young, M. F. (1993). Instructional design for situated learning. *Educational Technology Research and Development, 41*(1), 43-58.

Zaicou-Kunesch, C., Penny, S., Shackley, B., Ellis, S., Miyan, S., Dhammu, H., et al. (2010). Wheat variety guide 2010 Western Australia. *Bulletin 4795*. Retrieved from http://www.agric.wa.gov.au/PC_91997.html?s=1410546987,Topic=PC_91997

Zaki, M. J. (2004). Mining Non-Redundant Association Rules. *Data Mining and Knowledge Discovery, 2004*(9), 223-248.

Zenko, B., Dzeroski, S., & Struyf, J. (2006). Learning Predictive Clustering Rules. *KDID 2005, 2006*, 234-250.

Zhang, X., Pan, F., & Wang, W. (2008). *CARE: Finding Local Linear Correlations in High Dimensional Data*. Paper presented at the 2008 IEEE 24th International Conference on Data Engineering

Zhang, X. C., Nearing, M. A., Garbrecht, J. D., & Steiner, J. L. (2004). Downscaling Monthly Forecasts to Simulate Impacts of Climate Change on Soil Erosion and Wheat Production. *Soil Science Society of America Journal, 68*(4), 1376-1385.

Ziervogel, G., P, J., Matthew, M., & Mukheibir, P. (2010). Using climate information for supporting climatechange adaptation in water resource managementin South Africa. *Climate Change, 103*(3), 537-554.