2011

The application of data mining techniques to interrogate Western Australian water catchment data sets

Ajdin Sehovic
Edith Cowan University

Edith Cowan University

Copyright Warning

THE APPLICATION OF DATA MINING TECHNIQUES TO INTERROGATE WESTERN AUSTRALIAN WATER CATCHMENT DATA SETS

A thesis for a dissertation submitted in partial fulfilment of

the requirements for the degree of

Bachelor of Science (Software Engineering) Honours

Ajdin (Aiden) Sehovic

Supervisors: Dr Leisa Armstrong and Dr Dean Diepeveen

Faculty of Computing, Health and Sciences

School of Computer and Security Science

Edith Cowan University

2011

A. Sehovic

ABSTRACT

Current environmental challenges such as increasing dry land salinity, waterlogging, eutrophication and high nutrient runoff in south western regions of Western Australia may have both cultural and environmental implications in the near future. Advances in computer science disciplines, more specifically, data mining techniques and geographic information services provide the means to be able to conduct longitudinal climate studies to predict changes in the Water catchment areas of Western Australia.

The research proposes to utilise existing spatial data mining techniques in conjunction of modern open-source geospatial tools to interpret trends in Western Australian water catchment land use. This will be achieved through the development of a innovative data mining interrogation tool that measures and validates the effectiveness of data mining methods on a sample water catchment data set from the Peel Harvey region of WA. In doing so, the current and future statistical evaluation on potential dry land salinity trends can be eluded. The interrogation tool will incorporate different modern geospatial data mining techniques to discover meaningful and useful patterns specific to current agricultural problem domain of dry land salinity.

Large GIS data sets of the water catchments on Peel-Harvey region have been collected by the state government Shared Land Information Platform in conjunction with the LandGate agency. The proposed tool will provide an interface for data analysis of water catchment data sets by benchmarking measures using the chosen data mining techniques, such as: classical statistical methods, cluster analysis and principal component analysis.

The outcome of research will be to establish an innovative data mining instrument tool for interrogating salinity issues in water catchment in Western Australia, which provides a user friendly interface for use by government agencies, such as Department of Agriculture and Food of Western Australia researchers and other agricultural industry stakeholders.

Table of Contents

List of Figures

1. INTRODUCTION

The confluxes of innovative computer science disciplines, especially in data mining, as Johan (2009) states, can be regarded as a gift economy, compromised of values that contribute to knowledge and impact of thinking of others. Data mining has also contributed to the effective and robust modelling of business problems and along with advances in software development has lead to a number of modelling tools capable of examining large volumes of data, revealing hard to find patterns (Kowalski, 2000, p.1).

One of the factors contributing to the rise of popularity in data mining can be associated with the significance of knowledge discovery from databases (Kamber, 2006, p.5). Since the collection of data has grown over time, the management of data using knowledge in discovery databases (KDD) has become imperative in pattern discovery.

Before the development of spatial visualisation techniques, data mining was a difficult task to conduct in regard to interpretation and visualisation of geo-statistical data. The collaboration of various commercial and open source communities has resulted in establishment of specific spatial data mining methodologies which has improved the understanding of spatial type data. For example, a study by Kravchenko & Bullock (2002, p.804) of agricultural data sets found that geospatial topographic field analysis in conjunction with geo-statistical analysis could improve the crop ecology and agronomic management of soybean.

The following proposed research addresses the current industry problem of dryland salinity in Western Australia (WA). The government of Western Australia has expressed a need to find better predictive models to evaluate trends in land use amongst the states water catchments during the coming decade. The research will investigate and determine the most feasible spatial data mining techniques for conducting an analysis of water catchment data sets. Equally important, the results extracted during the analysis stage will help in assessing whether any significant spatial patterns are present and allow predictions to be made in relation to potential changes in climate and land use. This will be achieved through the use of an interrogative data mining tool that will be able to determine most appropriate data mining techniques for the proposed case study of the Peel Harvey water catchment area.

1.1. BACKGROUND TO THE STUDY

The proposed study is focused on water catchment issues in Western Australia, through an examination of the Peel Harvey region. The Peel Harvey region is approximately 70km south of Perth and covers an area of 3072 square kilometres (Rivers, 20002, p.1). Due to the vast land size and intensive agricultural practice, the region has a number of environmental sustainability problems, including increasing salinity. A national land and water resources audit assessment (ANRA, 2000) has forecasted that salinity will increase over the coming decades. It is estimated that approximately 8.8 million hectares (33%) by 2050 in the South West of WA will be at high risk of salinity damage. Furthermore, findings from (ANRA, 2000) indicate that approximately 81% agricultural land is at risk from dryland salinity. Consequently, this could lead to an estimated 1500 plant species being affected, with possibly 450 subject to extinction. As a result, the extent of increasing dry land salinity will greatly affect a large portion of Peel-Harvey inlet.

The Peel-Harvey catchment region is compromised of 27 large sub-catchments with 21 identified as residing in the coastal plain portion of the statutory boundaries, (Tom Rose 2003, p.7). To illustrate, Figure-1 shows the large physical size of the region, with gazetted plain portions outlined in purple. In this case, the region is coupled with rivers that feed directly into the Peel Harvey catchment.

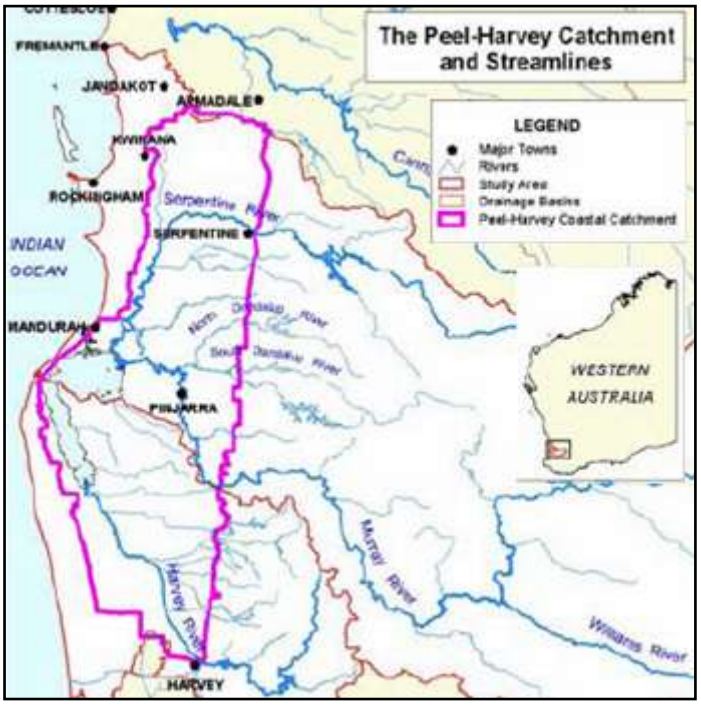


Figure 1: Total catchment and streamlines (Department of Environment and Conservation of WA, 2003, p7).

Mark Rivers (2000, p. 6-7) summarised the dominant conditions leading to the current problems experienced in the Peel Harvey catchment, these being:

- **Loss of Biodiversity:** as a result of historical agricultural development, native vegetation is lost and replaced by pasture or crops.

- **Water logging:** known to be common issue across Peel-Harvey catchment region, in particular across winter and spring due to high rainfalls and low evaporation.

- **Eutrophication:** caused by an enrichment of a water body with organic and inorganic plant nutrients which can increase biological activity in water, leading to algal growth. Also, eutrophication is a consequence of nutrient loss from the farming and urban areas;

- **Soil acidity:** a natural process that is enhanced by the use of fertilisers, leguminous crops and pastures, and

- **Irrigation salinity:** a predominant issue, increasing in occurrence across the southern portion of Peel Harvey catchment regions. Caused by an irrigating mixture of saline with water and chemical salts, it is believed that irrigation salinity has led to an increase of soil sodicity and ultimately promoted a decline in soil structure.

According to Ross, (2003, p. 10) these “issues have been exacerbated by rapid population growth, which has increased pressures for multiple uses of the land and waters in a sandy low relief high groundwater table and poorly flushed estuarine system”. An illustration (see Figure-2), a case study conducted by an Australian National Resource Atlas on Kamarooka catchment illustrates how salinity discharge may cause rises in water table salinity. Salinity discharge is compromised of two primary causes; first is the result of naturally occurring decomposition of oceanic salts produced from either rain or wind. Second, is related to land use, caused by people during irrigation or dryland management (“ANRA: Kamarooka Case Study Catchment”, n.d”). In short, this can be represented as a streamline discharge which may originate from the top of a hill, resulting in a flow of salinity that naturally descends through ground rocks and soil. The resulting discharge will eventually travel down the hill causing rises in water table salinity and ultimately impacting the agricultural land area (see Figure-2),.

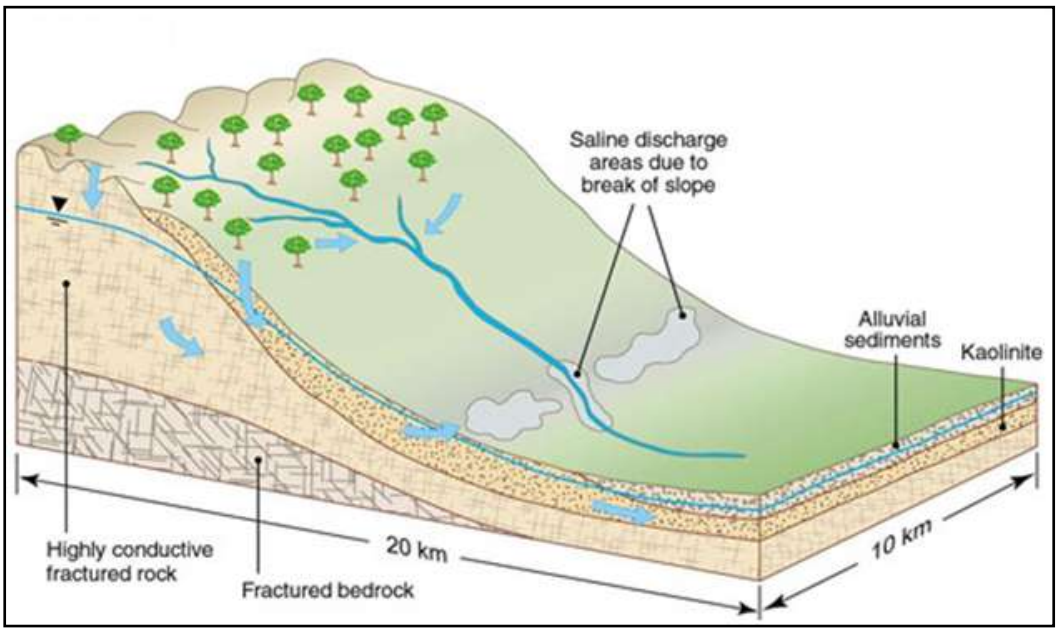


Figure 2: Salinity discharge caused in Kamarooka water catchment (ANRAa, 2000).

A comparative assessment has found the condition in Peel Harvey’s Murray River as extensively modified due to the salinity drainage (“Estuary Assessment”, 2000). With attention to salinity drainage concerning the Peel Harvey, Peel Inlet estuary, Figure-3, represents boundaries outlined in red as heavy discharge of salinity that flow to neighbouring sub-catchments.

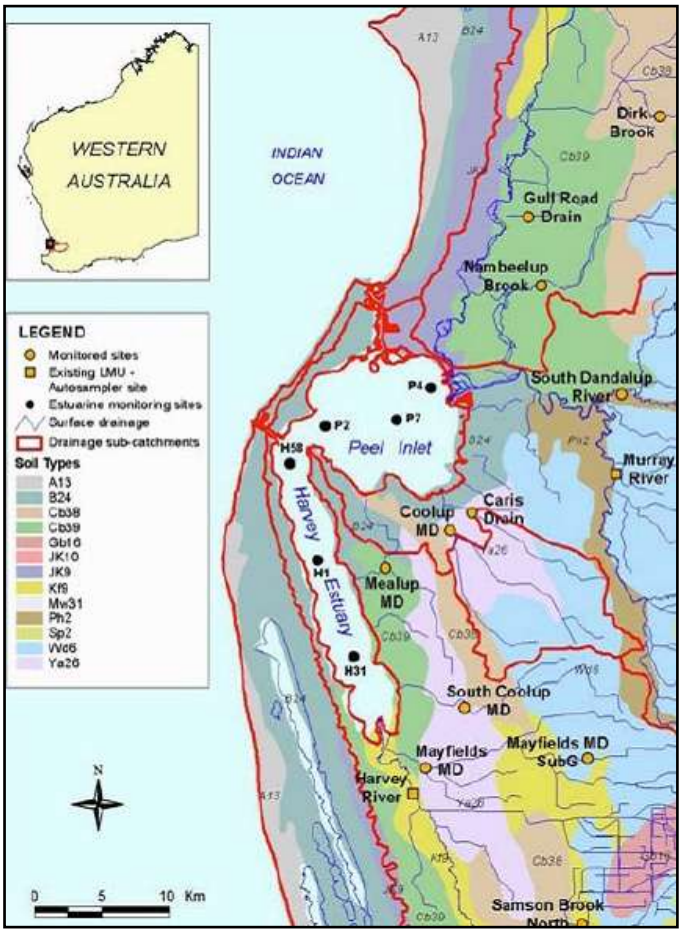


Figure 3: Peel inlet salinity discharge (Department of Environment and Conversations, 2003, p7).

The Natural Assessment Resource Group (NARG) has carried out an agricultural mapping of soil landscape data for the south west region of Western Australia. The mapped data is represented as a set of geographic spatial units or better known as datasets. The data is composed of hierarchies and stored in a GIS environment to allow spatial interrogation (ANRA, 2000, p. 17-18). The large proportion of the hierarchy is represented by different level of units, such as, regions, provinces, zones and systems. Importantly, each spatial unit is represented as a scaled map, for example, a zone displayed as a map consisting of hydrological characteristics. The hydrological attributes contain information, such as depth to water tables, trends and risk categories (Department of Environment and Conversations, 2000, p. 17). In addition, an Ag Bores database is used in conjunction with the soil-landscape system mapping to provide a spatial bore analysis (2000, p. 19).

1.2. SIGNIFICANCE OF THE STUDY

The climate of Western Australia is undergoing a period of change; with the current predicted climate trends and the impact of salinity indicating that south west Western

Australian water catchments are at great risk, posing critical economic impacts to infrastructure, biodiversity and agriculture (ANRA, 2000, p. 52).

The proposed research aims to determine the most appropriate data mining techniques for conducting an automated analysis on water catchment data sets. In doing so, this may help to establish grounds for carrying out the most effective decision-making process for changes in climate and land use in Western Australia. In fact, the proposed study, adheres to the current industry methodology, as outlined by the National Land and Water Resources Audit (2000, p.7).

1. Use existing spatial datasets for southwest WA
2. Identify and define areas at risk from dry land salinity
3. Predict current and future impacts of shallow water tables
4. Evaluate the effectiveness of predictions

The data mining techniques will enable the proposed methodology to provide the means of evaluation into the predictions of effects and risks of dry land salinity. This will consist of an innovative decision making platform that will contribute to solving and evaluating the effectiveness of the environmental problems and sustainability of productivity losses caused by dry land salinity in water catchment regions of Peel Harvey.

1.3. PURPOSE OF THE STUDY/STATEMENT OF THE PROBLEM

The primary purpose of the study is to determine how certain data mining techniques can be used to interpret trends in Western Australian water catchment land use. To that end, a study with contributions from various scientific disciplines and software engineering practices must be carried out on Peel Harvey data sets. The proposed study will use relevant information including spatial information on, boundaries, cadastre, geodetic, imagery, tenure, topography, roads and other relevant agricultural information on Peel-Harvey water catchment region.

The mining of large data sets is a difficult and time consuming task; this is certainly true when dealing with large volumes of spatial information. Therefore, it is imperative that custom software application is developed, which comprises the most appropriate data mining approaches, specific to the Peel Harvey water catchment case study. In order to justify the effectiveness of the data mining techniques, a

benchmark will be set against a simple statistical method. Any justification made from this investigation will contribute to evaluating the risk factors of water catchment attributes, for example, rainfall, crop yields, salinity nutrient and so forth. In addition, the ultimate goal is to conduct an interpretation of discovered patterns that will form as a basis of understanding various climate trends related to Peel Harvey water catchments.

Above all, as stated by Kargupta et al (2006), in view of the fact that data mining has undergone rapid development over the past decade it has become apparent just how important a multidisciplinary and application driven approach can be for a projects success. For this reason, the current research problem will see this study create an innovative rich user based application, to conduct spatial scientific investigation into climate factors on Peel Harvey water catchment land use.

1.4. THE RESEARCH QUESTIONS RELATED TO THE STUDY

Primary research question: "Can data mining techniques be used to interpret trends in Western Australian water catchment land use?"

i. ***Sub set questions one:*** "Which data mining techniques are the most appropriate for analysis of water catchment data sets"

ii. ***Sub set questions two:*** "How can data mining techniques be used to make informative predictions in relation to changes in land use and climate"

1.5. DEFINITIONS OF TERMS

Term	Description	Source
Algorithm	A set of instructions that aim to perform an action. It can consist of finite number of steps required to form an action.	(Howe, 2010a)
Architecture	Representations of a complex system design, including a set of components that make up the design.	(Howe,2010b)
Attribute	A value or a relationship based on an associated entity.	(Howe, 2010c)

Classification rules	Carries out a search on a dataset for a small number of rules which serve as classifiers for predicting results.	(Hui, et. el, 2008)
Cluster Analysis	Represents an overview of a region including clusters residing in the give region of space. It is used for classification purposes which illustrates a grouping of regions as a homogenous cluster.	(Stillwell & Scholten, 2001, p.160).
DAFWA	Depart of Agriculture and Food of Western Australia	
Data collection	A set of data gathered from either, surveys, or networked locations via data capture, data entry, or data logging.	(BusinessDictionary, 2010a)
Data Mining	A set of complex algorithms and techniques used for searching through large amounts of data.	(BusinessDictionary, 2010b)
ECU	Edith Cowan University	
Geospatial	A concept or illustration that represents a geographic location and characteristics of a natural or constructed feature, including boundaries on, above, or below the earth's surface.	(Dictionary, 2010a)
GIS	Geographic Information Systems is a technology which translates data into visual map that can make up geographical representations.	(Dees, 2002)
Hydrology	Distribution of geological matters that comprise the, distribution and effect of ground water.	(Answers, 2010a)
IDE	Integrated development environment are general user interface programming environments which are optimised for platforms that they support for creating applications and tools. For example, Java Eclipse or Microsoft Visual Studio.	(Wong, 2003)
Machine Learning	A computer machine which has an ability to improve its own performance through the use of software comprised of artificial intelligence	(BusinessDictionar, 2010c)

	techniques.	
Method	A procedure or a set of procedures comprised of techniques of a particular discipline or a systematic way of accomplishing a task.	(Answers, 2010b)
Methodology	A system or an approach that is composed of various rules and disciplines that aim to understand or study the given problem.	(BusinessDictionary, 2010d)
Model	A set of steps or schematics that describes a system, a theory.	(Answers, 2010c)
OGC	Open GIS Consortium (OGC) is an organisation for improve the interoperability process of spatial geographic standards.	(Gould & Hecht, 2001)
PCA	Principal Component Analysis is an algorithm used for reducing the dimensionality of a data set from a large number of related variables.	(Jolliffe, 2002)
Precision agriculture	A composition of applications and technologies aimed at managing spatial and farming practices with a purpose of improving farming crop production.	(Answers, 2010d)
R Environment	R environment or R Project is an advanced statistical computing application which provides high level of graphics quality and it is freely available as an open source platform.	(Ripley, 2001)
SLIP	Shared Land Information Platform is a government data custodian portal.	(SLIP, 2008, p.11)
Topology	Representation of a physical space, unlike geometry, it is not related to dimensions and angles but with the properties of a geographic surface. Such as, contiguity, order, and relative position.	(BusinessDictionary, 2010e)
WEKA	Waikato environment for knowledge analysis is an application workbench for conducting machine learning and data mining on data sets.	(Garner, 1995)
WA	Western Australia	

2. REVIEW OF THE LITERATURE

2.1. DATA MINING AND KNOWLEDGE DISCOVERY

Data mining is defined as an interdisciplinary approach, which comprises a number of paradigms including information science, statistics, artificial intelligence and database technologies (Kamber & Han, 2006, p.29). As stated by Abbas et al, (2002), the fundamental focus in data mining involves a heuristic approach that extracts greater value from the data than simple query and analysis approach (Hussein et. el, 2002, p 262 – 263). It provides a way of extracting useful and implicit knowledge from large sums of data, thereby supporting businesses in their decision making processes (Williams. 2006, p28). For example, Williams (2006) has reported the benefits of using data mining to improve customer segmentation and retention, credit scoring, product recommendation, marketing campaigns, cross selling and fraud detection.

Obtaining valuable information from large data sets using data mining is an essential step in process of knowledge discovery. Norton (1999), defines Knowledge Discovery in Databases (KDD) as the investigation and creation of knowledge, processes, algorithms, and the mechanisms for retrieving potential knowledge from data collections (Norton, 1999, p.1). Problems with handling large sums of data have resulted in unifying both the data mining with the KDD approach. In doing so, this has lead to an effective management of the data (Wojciech & Yao, 2001, p. 1 – 2).

Norton (1999, p.2) claims that KDD architecture was necessary since, an enormous increase in databases of all sizes and designs has created the necessity for more efficient grounds in data mining methods to access and analyse data. A data warehouse architecture or model allows the data to sit in a centralised repository (Inmon, 1995, p. 201). Inmon (1995) confers that although data warehouses are not essential they may greatly improve the effectiveness of the data mining process through the:

> ***Integration of data:*** allows the miner to easily and quickly look across vistas of data reduces the amount of time cleansing and conditioning.

Detailing and summarisation of data: necessary when the miner wishes to examine data in its most granular;

Historical data management: is important to the miner because important nuggets of information are hidden there; and

Metadata: serves as a road map to the miner, who uses metadata to describe not the content but the context of information.

Another benefit of data integration is that it allows the data mining analyst to concentrate on mining of data as opposed to the process of cleansing and integrating the data which as reported by Tamraparni (2003, p. 99), requires a considerable time to perform.

2.2 STAGES OF DATA MINING

General data mining process is comprised of several cyclic stages, these are outlined by Han and Kamber (2006, p.7): as data cleaning, data integration or selection, data transformation, data mining, pattern evaluation and knowledge presentation (see Figure-4).

Similarly, Ponce & Karahoca (2009, p.32) further narrows the process emphasising on the knowledge discovery, such as:

a. **Data cleaning, pre-processing and transformation:** involves preparing the data set for the data mining process. This includes, noise removal, missing data management and data sampling.

b. **Data mining:** carries out a basic data set analysis, such as, tasking, classification, association and cluster analysis of data.

c. **Knowledge extraction and interpretation:** used to describe the results from the analysis stage in a human-readable form. For example, an evaluation and understanding any patterns discovered.

2.1.1 DATA CLEANING

To begin with, cleaning of data, for the most part is an essential step that is used to carry out various maintenance procedures on data sets. Data maintenance is

compromised of filtering unwanted information, for example, modifying a set of records or attributes to achieve easy interaction during the pre-processing stage (Fayyad et. el, 2006, p. 301). This can include the mapping of data to achieve a single naming convention, for instance, the treatment of missing or noisy data (Fayyad et al., 2006, p. 40).

According to Tamraparni (2003, p. 99) data cleaning is time consuming, since between 30%, to 80% of the data analysis task is spent on cleaning and understanding the data.

2.1.2 DATA INTEGRATION AND TRANSFORMATION

Data integration is used to combine data from multiple sources into one coherent repository, for example, a database or data warehouse. Consequently, the next logical step involves carrying out transformation on data which can be comprised of several sub steps, like, smoothing, aggregation or generalisation of data (Soman et. al, 2006, p. 113). Fayyad et al. (2006, p.64) explains that data smoothing is used for performing data reduction, similarly, Ponce & Karahoca describe data reduction as the means of minimising the complexity in data (2009, p.50).

2.1.3 DATA MINING

The data mining stage, normally performed after pre-processing of data involves searching for hidden patterns; through the use of various algorithms. The data mining techniques, include, classification rules, decision trees, regression or clustering (Fayyad et al, 2006, B, p. 40).

2.1.4 PATTERN EVALUATION AND KNOWLEDGE PRESENTATION

Pattern evaluation and knowledge presentation is one of the final stages in the data mining, with the intention to interpreting and visualising the extracted patterns. In conjunction with data visualisation, it is essential to interpretation of the data mined knowledge. For example, this may consist of documenting; reporting or verifying previously believed or extracted knowledge (Fayyad et., 2006, B, p. 40).

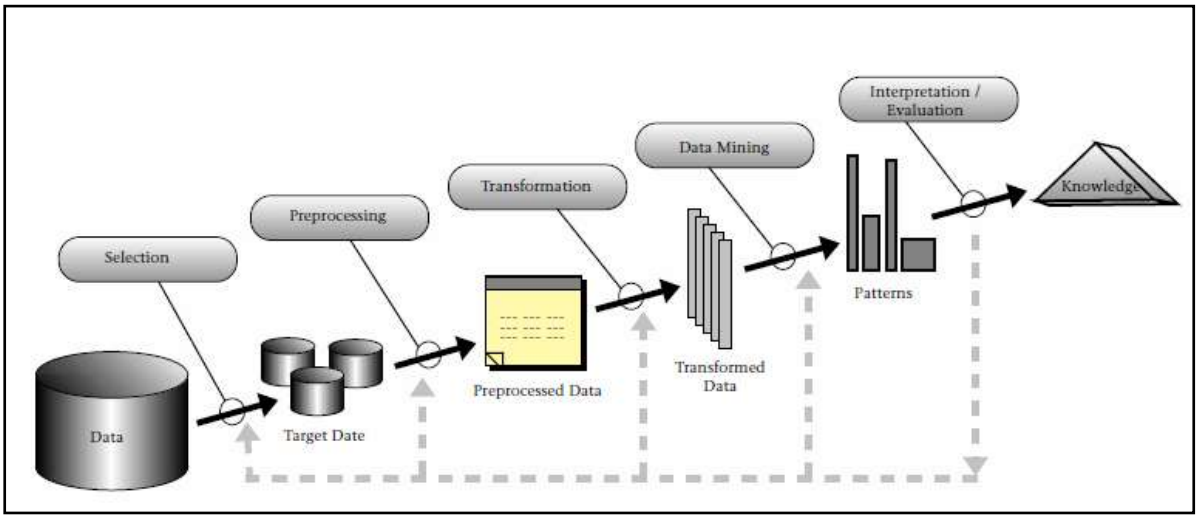


Figure 4: An overview of steps in data mining KDD process (Fayyad et. el, 2006).

2.2 DATA MINING TECHNIQUES

This section will examine the data mining techniques that could be considered the most appropriate for interrogation of the data sets. This review will concentrate on studies with an environmental or agricultural relevance.

2.2.1 OVERVIEW

Data mining techniques have been shown to aid in the description of data and prediction of future trends or variations in food grains (Mai, Krishna & Venugopal, 2006). For example, with the help of the technique rules described below, decision makers may suggest methods for improving agricultural productivity (Mai, Krishna & Venugopal, 2006). Technical approaches such as: clustering, learning classification rules, anomaly detection are some of the key techniques used for improving seed or crop varieties and converting an utilised land for use in agriculture (Mai, Krishna & Venugopal, 2006, p. 3).To ensure the correct techniques are chosen for a particular dataset, Aruns suggests that the analyst must determine whether the dataset used for analysis will serve the purpose in the prediction or description of the data (Arun, 2001, p.28).

> ***Prediction:*** Makes use of existing variables in the database in order to predict unknown or future values of interest
>
> ***Description:*** Focuses on finding patterns describing the data and the subsequent presentation for user interpretation

2.4.2 DISCOVERY AND PREDICTION RULES

Discovering and predicting rules have previously been used to represent demand in: water, fertilisation and pesticides. The effect of using the following rules on the chosen dataset assists in discovering new varieties of food grains (Mai, Krishna & Venugopal, 2006). An example of such rules is described below.

WATER DEMAND RULE

if("Supplyof_water(in%)" >= "Demand_of_water(in%)" ,~1 ,~0)

The water demand rule acts as an attribute in a dataset to verify if supply of water is greater than the demand of water (Mai, Krishna & Venugopal, 2006).

FERTILISER DEMAND RULE

if("Supplyof_Fertiliser(in%)" >= "Demandof_Fertilisers(in%)" ,~1 ,~0)

The fertiliser demand rule acts as an attribute in world dataset to verify if supply of fertilizer is greater than the demand of (Mai, Krishna & Venugopal, 2006).

PESTICIDES DEMAND RULE

if (“Supplyof_pesticides(in%)” >= “Demandof_pesticides(in%)”, ~1, ~0)

The pesticides rule acts as an attribute in world dataset to verify whether supply of pesticides is greater than the demand of pesticides (Mai, Krishna & Venugopal, 2006).

2.4. CLUSTER ANALYSIS

Cluster analysis technique can be used for presenting a general overview of the dataset. It is also used for classification purposes which can illustrate a grouping of regions into homogenous cluster. The clustered information can be visually illustrated with data mining tools (Stillwell & Scholten, 2001, p.160). There are a number of cluster analysis algorithms which can be applied for aggregation and

presentation of clustered data, for example, expectation-maximisation (EM), FartherFirst and k-means.

2.4.3.1 K-MEANS ALGORITHM

The k-means algorithm is widely used in the field of agriculture. According to Mucherino, there are several areas that k-means algorithm can be applied (A. Mucherino, Papajorgji, & Pardalos, 2009, p.68).

a. Atmospheric pollution forecast

b. Soil classification

c. Classification of plant, soil and residue in farming regions

d. Marking for apples using grades

e. Monitoring of water quality

f. Weed detection

2.4.3.2 EM AND FARTHERFIRST ALGORITHM

Researchers at the School of Computer Science and DAFWA have used large collection of data sets to carry out a cluster analysis with an open source WEKA tool. The analysis was conducted using the expectation-maximisation (EM) algorithm and FartherFirst algorithm on five soil data profiles to determine the accuracy of real values involved (Armstrong, Diepeveen & Maddern, 2007, p. 89 - 93).

2.4.3.3. PRINCIPAL COMPONENT ANALYSIS

“The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set" (Jolliffe, 2002, p. 1).

Mucherino claims that even though the PCA component can be used as a data mining technique itself, it may however be more effective if used in conjunction with the k-means algorithm for studying the wine fermentation in agriculture (Mucherino, Papajorgji, & Pardalos, 2009, p.68).

2.5 GEOSPATIAL DATA MINING

2.5.1 OVERVIEW

Data mining can be compromised of several disciplines include, information science, geographic information and visualisation. For that reason, Han & Kamber state that geo spatial data mining is the most efficient means of discovering and interpreting patterns (Han & Kamber, 2006).

Spatial data mining uses geographic information. A large proportion of the data mining activities carried out can be attributed to exploring high volumes of data sets comprised of geographic attributes and relations (Gahegan, p. 244). There are several important issues which must be addressed so that the analysis process can produce effective analysis of the large volumes of data sets. These include spatial data infrastructure standards and GIS interoperability.

2.5.2 THE SCIENCE OF SPATIAL DATA MINING

Geographic Information Services (GIS) is considered as the primary means of communicating geographic information or media (Sui & Goodchild, 2001, p.116). However, this process is a complex interaction that involves interfacing of large volumes of data from different scientific domain. Consequently, the rapid increase in geospatial data mining has created two fundamental problems that must be addressed to ensure consistency and efficiency among data sets. First, is the lack of interoperability applied in spatial data sets when you consider the large volumes of data sets that may be represented in different spatial formats and located in various geographical locations (Sun, et al., 2006, p.50). Second, there is a substantial lack of meaningful information available for describing the complexity and characteristics of the data set (Lee & Percivall, 2008, p.58).

2.5.3 GIS INTEROPERABILITY

The report conducted by Mark Reichardt represents a narrative into a potential havoc if interoperability to spatial data sets and general geographic information is not considered. Reichardt asserts that our world is going through a revolution of communication standards, and for this matter, many technologies are frequently the cause of confusion in the corporate standard decision making process (Reichardt, 2004, p.1). He further conveys his argument in conjunction to the Delphi study conducted by more than 800 users by concluding that “there is a clear and sudden shift in attitude towards software standards” (Delphi study, 2003, p.1).

However, to tackle the current issues trends of geographic information services, the Open GIS Consortium (OGC) was introduced over a decade ago with a goal to improve the interoperability process of spatial geographic standards. The open consortium has contributed to rapid improvement of interoperability among various geo-processing systems by introducing practical test beds and continuously conducting reviews on various consensus specifications (Gould & Hecht, 2001, p. 2).

2.5.3.1 SPATIAL DATA INFRASTRUCTURE STANDARD

Due to the evolution of geographic information in spatial data, a spatial data infrastructure (SDI) was introduced. The SDI is a subset process of the OGC which defines standard on interoperability across the climate sciences. (Woolf, et al., 2005, p. 9).The SDI contains developments which are compromised of several factors, for example, policies, data standards and human resources which are all necessary to achieve one unique goal. The goal is an infrastructure process consisting of achieving the following: process of spatial data, storage, distribution and utilisation of geospatial data (Woolf, et al., 2005, p. 9).

Woolf concludes an application that is diverse and vastly compatible with various interoperability standards will enable different GIS technologies and methods to use and represent data in the same manner (Woolf, et al., 2005, p. 9).

2.5.3.2 SPATIAL DATA INFRASTRUCTURE COMPONENTS AND LEVELS

Armenakis (2008, p. 328) describes the main components of the SDI, first, the ability to find relevant services and applications across virtual internet based SDIs. Second,

utilisation of tools and information that is independent of the supporting platform. And third, is the autonomous single vendor processing environment.

Amernakis (2008, p. 328) summarises the four levels which can be applied to achieve spatial interoperability.

1. ***Data interoperability:*** ensure that spatial data sets are shared independently despite the original format used;

2. ***Semantic interoperability:*** match data from different applications by sharing a common description that describes the properties of the data sets;

3. ***Network interoperability:*** provide communication between various services over the network; and

4. ***Interface interoperability:*** enable client applications to execute procedures on remote systems.

2.5.3.3 INTEROPERABILITY OF DATA IN CLIMATE SCIENCE

Various governments across the world have embraced the significance of spatial data standard by applying interoperable GIS applications. The United States, National Spatial Data Infrastructure (NSDI) Initiative was one of the first groups to embrace the spatial infrastructure. Additionally, this has led to interest from other countries to invest in interoperability of spatial standards, this includes, Canada, Australia, South Africa and the European Union countries (Woolf, et al., 2005, p. 9).

For example, the United Kingdom's, NERC Data grid project was developed to carry out scientific climate research on water salinity. To illustrate the interoperability process of this project, the project has conformed to standards of International Standard Organisation (ISO) by using compliant data model to measuring various salinity profiles across different marine locations. Furthermore, the information identified and collected in the process was collected and stored in data sets that abide by interoperable meta-data parameters (Woolf, et al., 2005, p. 12 - 14).

2.5.4 SPATIAL IDENTIFIER OBJECTS

Spatial data mining involves manipulation of attributes that are normally found in organizational databases, for instance, the database address table which may link to

other geographical information. This includes, namely qualitative data, which is identified using an indirect positioning system. Also, the spatial objects are normally represented by a location that is identified using a geographic Identifier (Santos & Amaral, 2004, p.374). Spatial identifier object is generally represented by a terrestrial position using two set of angles, the latitude and longitude (Ordance Survey, 2008).

2.5.5 QUALITATIVE SPATIAL REASONING

Qualitative reasoning is composed of various spatial algorithms which can establish relationship characteristics among various geospatial data. For example, these may be classified into the following three relations, namely, directional relation, distance relation and topological relation (Santos & Amaral, 2004, 376).

2.5.5.1 DIRECTIONAL RELATIONS

To estimate the orientation of objects, triangular model can be used to determine the relative orientation of corresponding centroids. Also, the triangular model is an effective approximation for handling objects which are far away from each other (Hernadez, 1994).

Hernadez (1994, p. 46-47) illustrates in the Figure-5 that if a centroid P2 object was to fall within the boundary of a triangular area in P1 than adjoining areas are assumed to have 50% overlap.

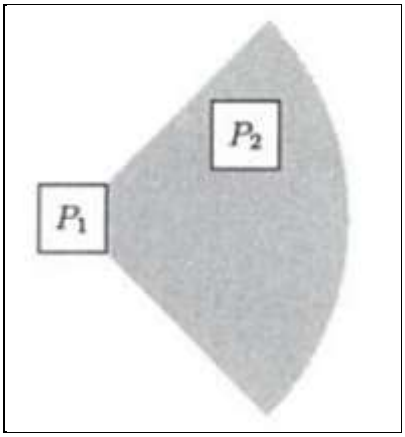


Figure 5: Triangular model (Henradez, 1994, p. 47).

2.5.5.2 DISTANCE RELATIONS

According to Renz (2002), distance relation is one of the most important aspects of spatial space which enables computation of distance between objects. The absolute

distance relation is calculated by dividing the real line into several sectors, known as: very close, close, commensurate, far and very far (Renz, 2002, p.39).

There are two main models which can be used to calculate the distance relationships, such as: absolute distance model and a relative distance model, as illustrated in Figure-6.

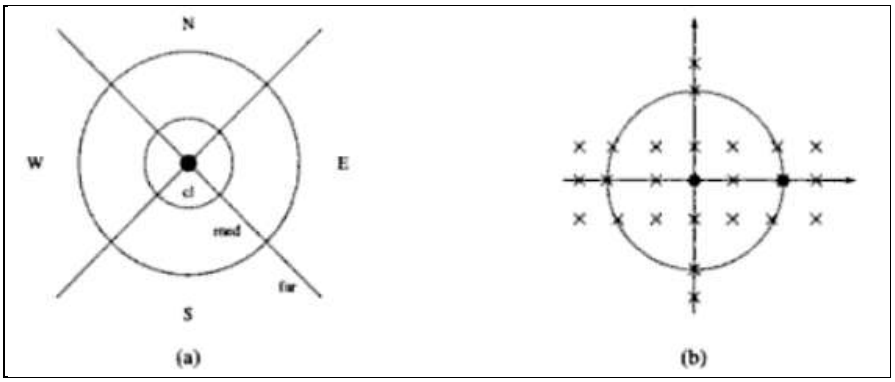


Figure 6: Illustration of distance relation models: (a) absolute distance, (b) relative distance (Renz, 2002, p. 40).

2.5.5.3 TOPOLOGICAL RELATIONS

Topological relation is used to represent a qualitative relationship between one or more regions and provide a way for distinguishing a regional representation (see Figure-7). For example, topological relation can determine whether or not two regions are disjointed, touched or inside each other boundaries (Forbus et el, 2004, p.65). In addition, such measures can determine when to compute and recognise different relationship types and provide a conceptual interpretations of regions (2004, p. 65-66)

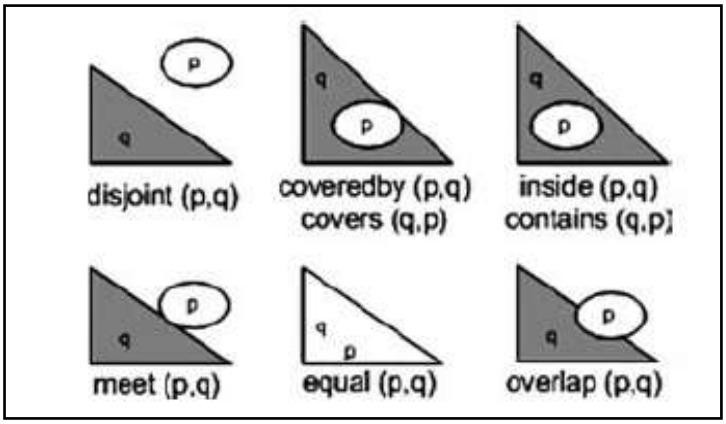


Figure 7: Topological regions (Santos & Amaral, 2004, p.376).

2.6 SPATIAL DATA MINING TECHNIQUES

2.6.1 OVERVIEW

Spatial data mining techniques can comprise various data mining techniques, generally speaking, these may include, classification, associations, clustering and principal component analysis. The following section outlines a set of useful case studies that examine different approaches in how, when, where, and what spatial data mining techniques could be used to when conducting analysis of various agricultural data mining, such as, soil mapping, land use, climate prediction and remote image sensing.

Large numbers of spatial techniques already exist; aside from a singular use of these techniques, it is not unusual to incorporate multiple techniques in the data mining process. That being said, integrating techniques such as: association, clustering and principal component analysis all form the basis in achieving a comprehensive and robust evaluation of spatial data sets.

2.6.2 SPATIAL INDUCTIVE CLASSIFICATION TECHNIQUES

There are various methods which can be applied to spatial data mining; according to Deren, Kaichang & Deyi (2000, p.31). Inductive learning is considered to be one of the most common methods. However various consequence may arise if inductive learning is not directly incorporated during classification of data. For example, a study by Deren, Kaichang & Deyi carried out spatial classification of land use on remote sensing images with a Bayes method (Deren, et el, 2000, p. 34). The first phase of the study was based without implicitly using an inductive approach. The classification was carried on various water based regions, such as: lakes, reservoir and ponds.Other non-water regions were also classified, including, vegetable fields, gardens and forests (Deren, et el, 2000, p. 32-34).

The study has found that Bayes method would yield a classification of 77.619% accuracy. Furthermore, it was found that an additional classification studied on dry land, garden and forests would result in reduced accuracy of 63.58%, 48.913% and 59.754% respectively. (Deren, et el, 2000, p. 32-34). However, despite an

inadequate accuracy of results, it was discovered when Inductive learning is integrated into the Bayes method, an increased accuracy of 88.857% can be achieved (Deren, et el, 2000, p. 34-35).

2.6.3 SPATIAL ASSOCIATION TECHNIQUES

“A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases. For example rules like most big cities in Canada are close to the Canada, US border is a spatial association rule” (Koperski & Han, 1995, p. 48).

In the study conducted by Ding and Perrizo (2008, p. 1513), applying a P-tree based Association Rule Mining (PARM) algorithm to spatial Remote Sensed Image (RMI) based dataset was found to be an effective method for identifying the following: crop yields, insect or weed infestations, nutrient requirements and flooding damage.

The structure of a P-tree algorithm is “a quadrant wise, Peano-order-run-length compressed representation of each (bit Sequential) bSQ file. The idea is to recursively divide the entire image into quadrants and records the count of 1 bits for each quadrant, thus forming a quadrant count tree” (Ding and Perrizo, 2008, 1514 – 1515). Figure 8 illustrates the benchmark associations of P-tree algorithm in comparison to a similar Peano Mask Tree (PM-tree) algorithm.

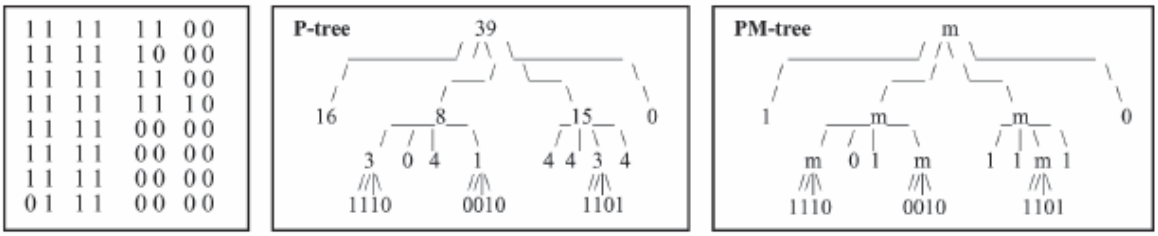


Figure 8: Benchmark comparison between PARM, P-tree and PM-tree (2008, p. 1515).

It is claimed that a P-trees is a major contributing factor in performance. For an example, a benchmark experiment conducted against other algorithms like, Apiori and FP-growth has found that PARM is more scalable with large spatial data sets. In summary, it is evident that the PARM algorithm, as illustrated in Figure 9, proves to be vastly superior in scalable performance when compared to other contending algorithms (Ding and Perrizo, 2008, p. 1520 – 1522).

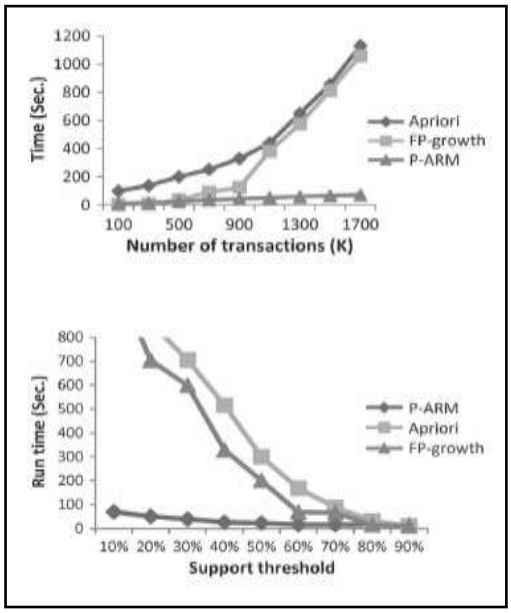


Figure 9: Scalable performance benchmark comparison between Apiori, FP-growth and P-Arm (2008, p. 1521).

2.6.4 SPATIAL CLUSTERING TECHNIQUES

Spatial cluster analysis is an important data mining technique with an essential role in quantifying geographic variation patterns (Jacqueez, 2008). Jacqueez (2008, p. 395) explains the spatial cluster analysis is "commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields, but the underlying principles are the same".

The study by Ng and Han, (1994) conducted a research in a development of a unique Clustering Large Applications based upon Randomised Search (CLARANS) algorithm which is based on various spatial cluster analysis techniques. Since CLARANS technique was specifically suited for analysing large sample of spatial and non-spatial attributes (2008, 144), for that purpose, it was utilised to study the distribution of 2500 luxurious housing units in Vancouver, Canada. However, the Ng & Han, (1994, p. 145) argue that in the past cluster analysis has been applied unsuccessfully to general data mining and machine learning problems. For this reason Ng and Han state that this can be linked to a lack of a natural notion in similarities among the clustered objects.. In fact, the CLARANS method has taken a direct approach in challenging this issue by successfully establishing natural notions of similarities between objects. By the same token, such similarities can be compared to the Euclidean or Manhattan relation concepts, as discussed in research methodology section.

In order to justify the effectiveness of CLARAN, other cluster algorithms, like, Partioning Around Medoids (PAM) and Clustering Large Applications (CLARA) were tested. Ng & Han outline (1994, p. 144 – 149) some of the characteristics found in the following algorithms, during an experimental stage:

1. ***PAM* cluster analysis**: was used to establish a location of a **medoid** object in each cluster node. Medoid objects are specifically processed by being centrally allocated to each clustered region. An experiment has found that PAM would perform satisfactorily on small datasets, however, PAM would not perform effectively enough on medium or large data sets (Ng & Han, 1994, p. 144).

2. ***CLARA* cluster analysis**: is primarily used for analysing large data sets by searching for representative objects in conjunction with the PAM algorithm. Findings prove that a lack of distance accuracy was present during an analysis of 5 clustered regions. As a result, the clustered regions would yield a displacement gap of 30% during an association of 1000 objects. Furthermore, when number of objects was increased to 2000, an association of unsatisfactory 20% improvement were found. (Ng & Han, 1994, p. 146).

The experiment as illustrated in Figure-10 was based on evaluating the efficiency in number of objects during a runtime and relative distance benchmarks between CLARAM, PAM and CLARA spatial clustering techniques.

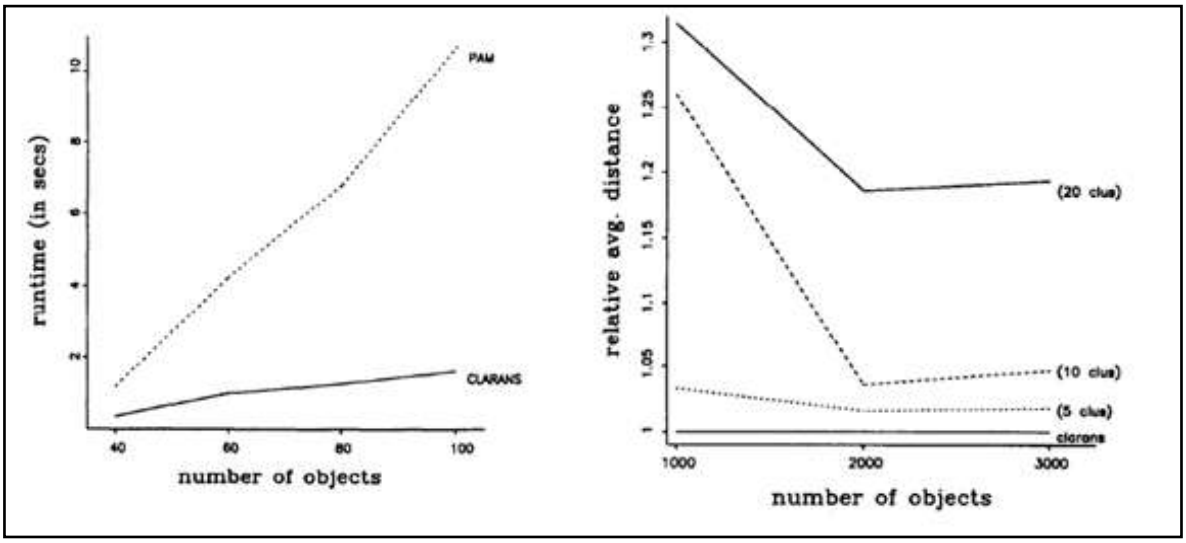


Figure 10: Experiment result of CLARAM vs. PAM and CLARA (Ng & Han, 1994, p. 152).

In conclusion, satisfactory results have been achieved when the CLARANS technique was applied to a real estate dataset with a set of filtered rules to discover spatial distribution of various luxury houses in city of Vancouver, Canada. The results have discovered realistic outcomes, in particular when representing clustered tuples or condo regions (Ng & Han, 1994, 153). Ng & Han in support of Figure-11, outline the results for the distribution of four universally different spatial regions (1994. 152).

Top right rectangular region: represents 1200 condos uniformly distributed.
Bottom left hand region: represents 320 mansions and 80 single houses.
Bottom right polygonal region: represents 800 single houses.
Middle scattered region: represents 100 single uniformly distributed houses.

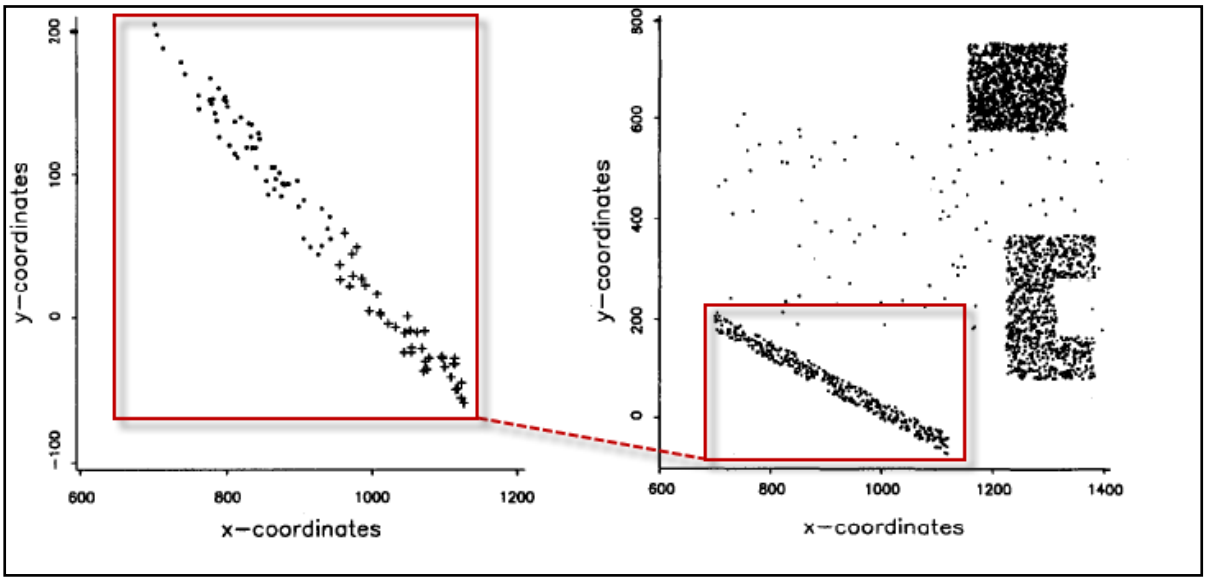


Figure 11: Spatial distribution of the 2500 luxurious houses (Ng & Han, 994, p. 152).

2.6.5 SPATIAL PRINCIPAL COMPONENT ANALYSIS TECHNIQUE

Principal Component Analysis (PCA) normally used for multivariate statistical purposes, can also be applied spatially. For example, it was proven that PCA techniques can be used to identify parameters of maximum variations by monitoring spatial and temporal changes of water quality, elevation of the water and land use (Zeilhofer, Lima & Lima, 2006). Similarly a recent study by Asanobu Kitamoto (2002) into spatio temporal data mining study has successfully demonstrated that PCA technique can be an effective method for extracting spatial attributes, such as, latitude, structures and spiral bands from a large collection of satellite images of typhoons.

2.6.5.1 SPATIO TEMPORAL DATA MINING USING PCA

According to Andrienko et al., (2006) spatio temporal data mining is an emerging research area which undertakes development of novel data mining applications. The fundamental aspect of this research is subdivided into two areas, spatial data and temporal data. Spatial data is normally obtained by GIS and robotic or mobile applications, while the temporal data is obtained by registered events, such as, telecommunication or network traffic data (Andrienko, 2006, p.187).

Also, the spatio-temporal data mining can be used to study the dynamics and patterns in spatial analysis (Kitamoto, 2002). For example, research by Kitamoto (2002, p.31-35) has effectively applied the Principal Component Analysis for studying the patterns of typhoon clouds of northern and southern earth hemisphere. Similarly, it was demonstrated that geographic visualisation (GeoVIS) methods such as, cluster analysis in conjunction with knowledge discovery in databases can provide an effective means for extraction, correlation analysis, anomaly detection, pattern recognition and filtering of spatio temporal patterns in environmental data (Wachowicz, 2002, p.483).

2.7 DATA MINING TOOLS

2.7.1 OVERVIEW

There are a number of data mining tools that have been developed by different communities (Schneiderman, 2000, p.9). The following section describes the data mining applications and tools which are used for machine learning, data mining and statistics and geospatial paradigms. The popularity of data mining has lead to both the public and consumer market in developing various open source and commercial platforms, these include, WEKA, Project R, commercial SPLUS tools, and more described in the following section

Current modern data mining applications are compromised of rich user interfaces with an emphasis on problem solving via graphical means, such as simulations. Although, the extra complexity may be an additional overhead, however according to Schneiderman (2002, p. 8), it is an essential process if the user is directly involved with the graphical user interface. The visualisation of problems in modern data mining applications can therefore provide the best means of pattern discovery. They

may assist in effective decision making for highly inhomogeneous and noisy data (Keim, 2002, p. 100).

By the same token, Schneiderman (2002, p 10 – 12) has outlined three fundamental recommendations in developing graphical data mining applications:

> **Recommendation 1:** deeper understanding of the data sets may be achieved by creating various graphical charts, like, scattergrams; in doing so this may contribute in creating an innovative data mining tool;
> **Recommendation 2:** provide the user with an opportunity to specify what patterns they are looking for in the data set. Also, ensure the user is given the ability to perform an interesting evaluation of the results; and
> **Recommendation 3:** create an atmosphere of social interaction between the researchers and practitioners by sharing data from multiple sources. Create a social application domain that is focused on allowing the user to present their findings with other domain experts.

2.7.2 WEKA

The Waikato environment for knowledge analysis (WEKA) as described by Garner (1995) is an application workbench capable of carrying out machine learning on real world data sets. In addition, Garner adds, the WEKA analysis was designed to bring a range of machine learning techniques and schemes under one unified repository (Garner, 1995, p.1). Equally important, the WEKA's primary interface known as an Explorer is comprised of user interface panels that allow the data analyst to conduct various data mining operations. For example, these may include, conducting the pre-processing of data set which resides on an SQL server, in either remote or local mode (Frank et. el, 2004). However, Frank et. el, (2004, p. 2479) argues that WEKA's memory is not effectively maintained. As a result, Frank suggests that this may cause unwanted high memory usage and potential performance reduction during the sampling of large data sets.

2.7.3 PROJECT R

Project R originally developed by Ross Ihaka and Robert Gentlemen at the University of Auckland, is an advanced statistical computing application which provides high level of graphics quality and it is freely available as an open source platform (Ripley, 2001, p.1-p.2). Ripley claims (2001, p.2), one of the key benefits of

Project R is related to high public support of various packages which harvest different data mining and scientific algorithms. The packages also known as extensions can be obtained free of charge from the Comprehensive R Archive Network (CRAN).

However, Ripley (2001, p.2) has compared several differences to another similar commercial statistical platform, named S or S-PLUS. One of the benefits of Project R when compared to S environment is the application size and notable performance speeds on less powerful computer machines. However, Ripley suggested that project R performs unsatisfactorily during three dimensional graphical processing. Nevertheless, both platforms provide rich two dimensional graphical analyses. In particular, for an open source platform, the Project R is proven to be far more superior in computing complex mathematical annotations.

2.7.4 S ENVIRONMENT

Unlike the project R, Bates (1999, p.266) claims that S environment was established to be one of the principal computing applications for statisticians and serious data analysts. Bates argues that S is greatly as a result of being able to provide user extensibility and flexibility of graphical user interfaces to users.

2.7.5 MATLAB

Matlab is considered as one of the predominant commercial software language tools in technical computing. Likewise, Matlab provides strong support of expressive language, specific to running rich data analysis, simulation and mathematical modelling inclusive of many mathematical algorithms (Travinin Bliss & Kepner, 2007, p.336).

The users of matlab are primarily said to be engineers and scientists, for this purpose, the users are given resourceful design features. Travinin Bliss & Kepner clarify that Matlab software is valuable in so that is allows the user to concentrate on the core scientific tasks while less time is being spent during an implementation phase (2007, p.336). Furthermore, Matlab contains powerful functionalities that enable complex data processing, specifically cantered for large data sets (Travinin et. al, 2007, p.336-337).

2.7.6 GRASS GIS

Geographical Resource Analysis Support System (GRASS) is a powerful open source software environment that focuses on geospatial data mining and other scientific studies. Neteler & Mitasova (2008, p.3) argue that Grass software provides vast support of raster and vector data formats which are capable of carrying out powerful image processing and data mining tasks. Additionally, designed in mind to support GIS interoperability features, specifically compliant with OGC industry standards.. Consequently, GRASS provides approximately 350 mathematical modules, in particular catered for analyses on geo-referenced analytical features, such as: data management, data processing, spatial analysis and spatial visualisation (Neteler et al, 2000, p.3).

Neteler & Mitasova claim that GRASS, originally developed to run under a UNIX operating system (OS), though it is said that existing knowledge of UNIX system is an essential requirement in order to make the most of application capabilities. Nevertheless, gradual support for other operating systems have been incorporated, including Sun Solarix and MacOS X. In addition, it was proven that GRASS can surprisingly run on constrained limited devices such as portable data assistant (PDA) devices (Neteler et al, 2000, p.7).

In a study by Woolf, A., et al., (2004, p. 14) they showed that geological algorithms can effectively be used with Grass software to aid in the process of constructing and visualising three-dimensional land surface boundaries. For example, using raster data sets from Honjyo, Akita regions in Japan, GRASS was able to visualise various three-dimensional boundaries that represent various detail of surface elevations.

2.7.7 DATA BIONIC ESOM

According to Ultsch & Morchen (2005, p.5) the Data Bionic Emergent Self-Organising Maps (BIONIC ESOM) is an open source tool that is specifically developed to conduct common data mining tasks in support of classification and other cluster analysis techniques. Moreover, ESOM tool is reported to comprise of a rich graphical user interface for carrying out various algorithms that deliver visual interpretation of data sets through, clustering, outlier removals and construction of non-redundant map views (Ultsch et al, 2005, p.6).

2.7.8 SHARP MAP

SharpMap is a spatial development library tool that assists software developers to query geo-spatial functions and render geographic information to maps. The rendering engine is based on two common GIS data formats, vector data and raster data. Furthermore, some of the examples of vector data types supported by SharpMap, include, ESRI Shape files, PostGIS, Oracle and CSV. Likewise, the support of raster data types, includes, Grass raster format, JPEG, Arc and many other common formats and many other common industry formats.

A research by Lamas et al. (n.d, p.722), has demonstrated how Microsoft Visual Studio 2005 in conjunction with Sharp Map library can aid the process of constructing a rich geospatial PDA application by generating spatial map attributes and carry out a request of geometrical information in real time. For example, Lamas et al (n.d, p. 722-723), explains that Sharp Map was used effectively to query geometrical spatial data through the PostgressSQL database provider and notify the user of their location of interest.

2.7.9 POST GIS

The significance of centralised data mining storage and management of geospatial data is a crucial task, in particular when dealing with large sums of data in an online web application. An investigation into the use of spatial database tools by Brovelli & Magni (2003) has demonstrated various capabilities of Post GIS tool. It was established that Post GIS can be configured to run an online data analyses of cultural heritage data sets and illustrate the analysed outcomes in a digital format. The results could then be used to render various topographic properties, such as: pathways, highlights, and hydrology and land use characteristics. Brovelli & Magni (2003, p.93) concluded that Post GIS, originally being an extension to Postgresql SQL database is an indispensable system from which an online geospatial data can be used to render and evaluate digital maps in an online application using the Map Server, the component of a Post GIS tool.

2.7.10 SUMMARY

The current research will revolve around several dominant and inclusively appropriate data mining tools for current study, such as: Project R and Post GIS. For this purpose the Project R will serve for carrying out data analysis on selective data sets, including pre-processing and a Post GIS tool will provide storage and relational management of spatial data sets. However, in order for such tools to interface and communicate effectively, the proposed methodology as explained in section 4 will employ a collaboration of several other equally significant spatial data mining and data visualisation tools. Refer to section 4.3 for an in depth explanation into the apparatus and conceptual context of the research tools involved.

2.8 CASE STUDIES SIMILAR TO THE CURRENT RESEARCH

2.8.1 OVERVIEW

The following section introduces case studies related to the proposed data mining research. These include research studies into various decision support system (DSS) and current best management practices for developing software applications which may assist various geo-spatial data mining research groups. Equally important, the following section examines the concept of precision agriculture and how it may contribute in discovering useful patterns, in particular, in farming fields which could reduce farmer's costs and therefore keep many struggling farms viable (Luntz, 1998).

2.8.2 SELECTING AREAS FOR LAND USE IN WATER CATCHMENTS

Dunstan, Armstrong & Diepeveen (2009) emphasise on current impacts of dry land, in particular, the effects that pose threats to land use and water catchments of Australia. Therefore, It is concluded that salinity is a,

> "... major problem in arable areas of Australia where past farming practices have led to rises in the ground water table that result in stored salt being transported to the surface. High levels of salinity at or near the soil surface diminish crop yield and result in runoff into creeks and streams with high salt content".

For this purpose, a research into the study of Land Use Cover Change (LUCC) was established to illuminate the effects of human activities on the landscape and environment, as well as to predict the trends in environmental impacts. Furthermore, the LUCC model was created for carrying future trends simulations of dry land salinity by a set of hydrology inputs. For example, inputs included in the process are, rainfall, land use and soil type which all serve the purpose of discovering useful patterns (Dunstan, Armstrong & Diepeveen, 2009, p.1-3).

One of the unique aspects of LUCC is the application use, which has the ability to calculate a rate increase in the ground of water table rises using an aggregation measure model, namely, Depth of Water Table (DWT). Dunstan, Armstrong & Diepeveen (2009) outline an aggregation of measures which comprise an evaluation of a DWT model:

a. A minimum DWT.
b. A maximum DWT.
c. An average DWT.
d. Percentage of blocks with Depth Water Table at the surface.
e. An average decrease in the DWT.
f. Percent of blocks with rapid decrease in DWT.
g. Total amount of count of blocks or dimensional area of a water catchment region.

In conclusion, it was determined that an application was a viable simulation environment, with capabilities in providing meaningful and informative data mining predictions in DWT rate scenarios. As a result, this has lead to establishing a better understanding of the consequences for an overall water catchment planning (Dunstan et al, 2009, p.8).

2.8.3 SIMULATING CROP DECISIONS FOR WATER RESOURCE MANAGEMENT

A study conducted by Ekasingh et al. (2005) has demonstrated the benefits of using decision trees with the WEKA, a data mining tool which may assist in more efficient management of water resources. The approach was comprised of data collections carried out on three catchment areas (Ekasingh et al, 2005, p. 317), including a comprehensive crop study of house hold characteristics surrounding the water catchment areas. The study has concluded that decision trees have made a great impact to the classification of various crop-types and categories of crops. Ultimately, it was discovered, when a decision tree technique is employed to analyse socioeconomic and biophysical variables, for example, income, subsistence production, erosion and water yield characteristics (Ekasingh et al, 2005, p.325) can simulate effective agricultural socio economic land decisions.

2.8.4 PRECISION FARMING FOR AGRICULTURE

In October 2001, Wilcox (p. 1) described precision farming as coupled with many challenges especially since,

> "...the goal of precision agriculture is to manage your farm land better. The challenge is to find the best way. If you talk to 10 different people, you may get 10 different tasks. It's cause for much debate".

Wilcox (2000, p.1) reported on an agricultural study undertaken by farmers into precision farming at Nebraska Agricultural Technologies Association (NeATA). One of the unique aspects of precision farming, in particular to the NeATA study context, is how best farming practices can be incorporated with modern software applications. For example, Wilcox (2001, p.1 – 2) provides some useful insights into how the Trimble AGIS system has inherited various GIS and data analysis software packages to normalise dataset of yield data. Furthermore, the AGIS system was used to compare: soil images, soil conductivity and topography in order to produce an evaluation for profit and loss scenarios in conjunction with topographic GIS layers. In particular, the precision farming was also used as a long term plan in determining future instability of areas that may suffer from an ongoing dryland.

Wilcox (2001, p.2) outlines various characteristics of the AGIS system and other software tools contained into Trimble AGIS system.

> **AgLeader SMS Basic software:** was used for organising the yield data, also, it was used to import Arc View shape files and TIFF images for creating custom map layers;
>
> **Farm YES software:** obtained from Red Hen Systems for conducting yield data analysis;
>
> **Arc View GIS ESRI software: commonly** used since it is regarded as preferable software among government agencies;
>
> **Arni Hinksons software:** commercial used primarily since it was well priced, and above all, it provided high end geographic information (GIS) by quickly handling data at any resolution without an over head of re-sampling of data sets;
>
> **Micro Excel and Red Hen's MapCalc Learner:** for statistical analysis of data sets; and
>
> **S-plus Math Soft and GS+ design software:** for carrying out more complex data analysis.

A similar study by Zhang et al., (2009) has proven that precision farming is for online decision making. ZoneMap was developed as a web based decision support tool that allows end users, farmers, ranchers or extension specialists to conduct precision agriculture of crop management (Zhang et al., 2009, p. 105). In addition, Zhang et

al., (2009, p. 105-108) suggest that ZoneMap is capable of gathering multiple source of data from any part of the world for carrying out remote sensing observation which was proven to be extremely effective in capturing field variability.

2.8.5 BEST MANAGEMENT CRITERIA FOR PEEL-HARVEY WATER CATCHMENT

Australian government coastal catchment initiative (CCI) of Western Australia has committed an investigation into the best management practices (BMP) for Peel-Harvey water catchment in order to evaluate and improve water quality of various estuary ecosystems. The importance of the study was to evaluate the effects of various BMP criteria's that could serve as a model for future decision support systems (DSS) and for spatial optimisation of Peel-Harvey catchment data sets.

The study suggests that integration of best management practices into decision making tools is an essential process which may result in the reduction of excessive loads of nutrients reaching the receiving waters (Keipert et. al, 2008, p. 1749). For example, the current plan suggest that DSS framework described by Weaver et al (2005) may assist in examining possible nutrient scenarios, for example, a decision tool that studies the adoption of P indicators (2008, p.1752). For this purpose, the P indicator, an original approach introduced by Heathwaite et al (2003) outlines three types of contributing factors, as follows:

> ***Source factors:*** nutrient input characteristics including mineralisation inputs;
> ***Transfer factors:*** rainfall and erosion risks;
> ***Delivery factor:*** land drainage factors including hydrological connectivity.

The following Sanken diagram (see Figure-12) illustrates a model for estimation of relevant nutrient flows and stores for the Peel-Harvey water catchment. Generally, the diagram illustrates each bar which represents a relative contribution from each land use sector. For the most part, the applicability of a Sankey model may provide many benefits in better understanding the impacts of nutrient flows and stores for various land use sectors (Keipert et. al., 2008, p.4 - 5).

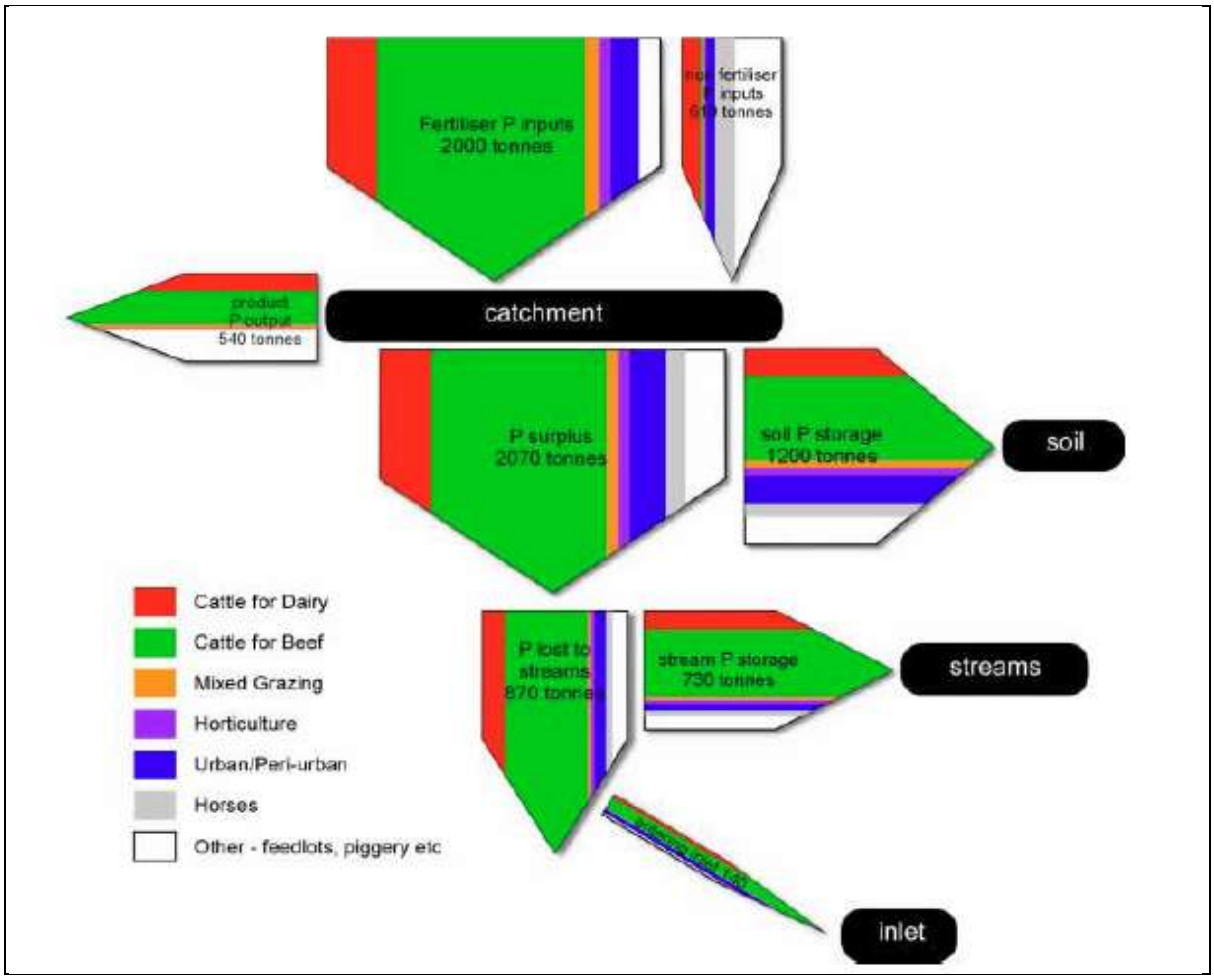


Figure 12: Sankey's diagram of P flows and stores in Peel Harvey water catchment (Keipert et. el, 2008, p.5).

In conclusion, Keipert et al., (2008, p. 1754) outlines various guidelines of water catchment criteria's that can be implemented, for instance a DSS, to provide a guide priority setting and investment planning to achieve the following:

a. Establishing water quality with a statistical approach using median values for each P concentrations factor;
b. Incorporation of a cost benefit for different types of BMP's;
c. Provide effectiveness by reducing nutrient loads to all estuaries; or
d. Merge all the following criteria's as a, one unified BMP criteria.

2.8.6 HYDROLOGICAL MODELLING WITH JGRASS SOFTWARE

An application that inhibits various Decision Support System (DSS)features has been developed in conjunction with other system to study topographic data and represent visualisation of hydro-geographic scenarios (Sengonul & Yilma, 2001). For example, a scenario depicting large stream of waters was developed with an open

source; JGRASS software. According to Sengonul and Yilma (2001), JGrass, being hydrological modelling software, was proven to be successful in carrying out effective management and modelling of river basins, For instance, the following diagram (see Figure-13) illustrates a composition of other applications used for collaborative and centralised communication with the Grass or JGRASS system.

In addition, Sengonul & Yilma (2001, p. 249 – 250) justify that digital elevation model (DEM) can be directly applied to the JGRASS for representing geospatial information. For this purpose, a range of DEM rules when incorporated into the GRASS environment, can govern the implementation for the following: basic analysis, network measurements, hill slope analysis and visualisation of hydro geomorphic index (2001, p. 249).

In summary, according to the figure 16, it is conceptually clear, that a direct integration of a database management system and other equally important data mining tools could be constructed to achieve a unified environment. Comparatively, the current research poses some similarities as proposed in the section 4. For example, an integration of external GIS data, generally used for data analysis can be directly interfaced with an R environment, as shown in the figure below (see Figure-13).

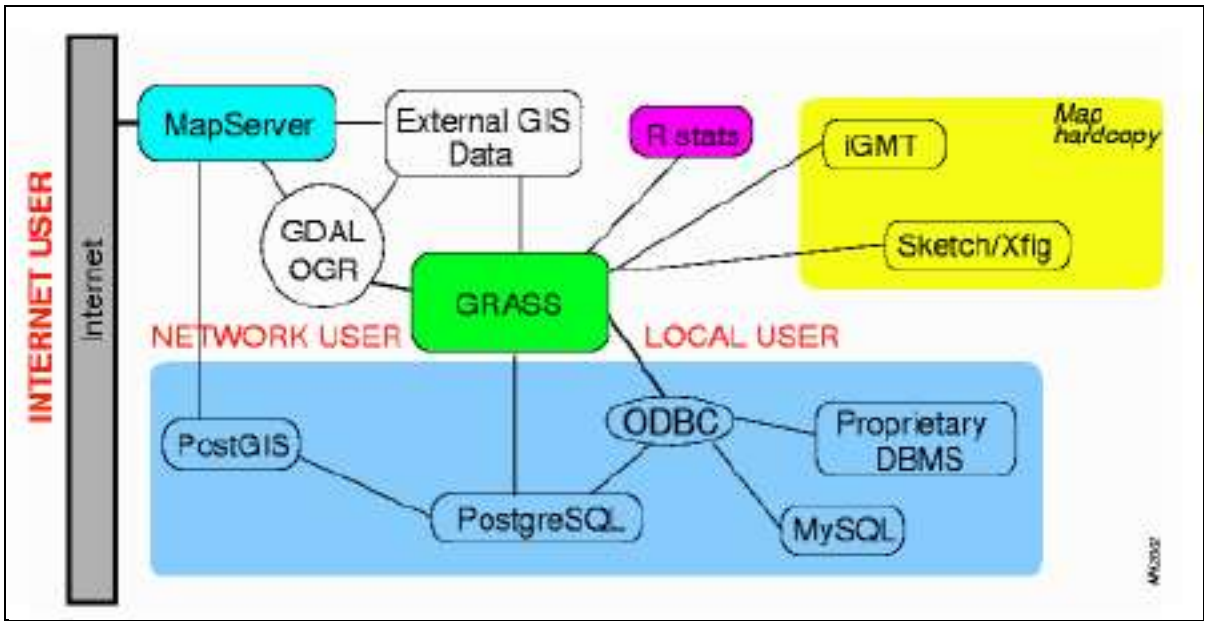


Figure 13: GRASS (JGRASS) and integrated applications (Sengonul & Yilma, 2001, p.248).

3 MATERIALS AND METHODS

This section describes the proposed research methodology with a focus on the industry problem. In addition, this section includes a description of relevant equipment required to create a proposed data mining application. The section will also include an examination of how the data sample was obtained with the proposed research strategy and application process to conduct data analysis.

3.1 INDUSTRY PROBLEM

The primary focus of this research is to find ways to solve the dryland salinity problems facing the Western Australia, in particular, the south west region of Peel-Harvey. For this purpose, the WA government has conducted surveys to discover the extent and impacts of dryland salinity trends and establish how ground water table salinity may cause dramatic effects on our climate and agriculture over the next decade (ANRA, 2000). As a result, government and industry has proposed a methodological approach to combat the effects and risks of dryland salinity (Australian National Resource Australian, 2000, p.7).

3.2 RESEARCH METHODOLOGY

The methodology employed in the proposed research will investigate an appropriate data mining techniques using a software instrument comprised of various data mining tools. The instrument will be created using a general software engineering process. Though, software engineering processes is said to have gained little knowledge into its appropriateness of use (Basili, 1992, p.1-2). For this reason, it was determined that scientific method, originally introduced by Basili (1992, p.2) and quasi-experiment research will be the most applicable methodology to be undertaken for the current research.

There is a lack of effective data mining frameworks available which to create an effective research instrument software for validation of data mining methods. As a result, the quasi research approach was deemed a feasible approach which will involve a manipulation of quantitative data from the Peel Harvey water catchment, spatial data sets. The research will perform a benchmark comparison followed by a detailed data analysis and validation into the effectiveness of data mining techniques, such as: cluster analysis, PCA and classic statistical methods. Furthermore, the software, an instrument tool, will assist in an inductive process by interpreting the results using various open source geospatial tools. Therefore, the

instrument tool aims to integrate various proven data mining tools capable of interpreting the salinity trends on water catchment data sets. Consequently, this will determine whether any useful patterns or correlations in climate trends exist across longitudinal data collection.

In conclusion, to ensure an effective creation of an instrument tool is carried out to validate current research activities, the following measures will be applied, as follows (Basili, 1992):

a. ***Observe the research*** on dryland salinity using Peel-Harvey water catchment data sets. And, evaluate the effectiveness of current data mining techniques used to discover meaningful patterns;

b. ***Propose a model*** by developing an integrated geospatial data mining application domain in conjunction with latest, industry standard GIS tools;

c. ***Measure and analyse*** *the data sets* using statistical tools and geospatial frameworks;

d. **Validate hypotheses** by carrying out a benchmark evaluation and interpretation of analysed results; and

e. ***Repeat the procedure*** in a software engineer manner until reasonable justification is made to establish a meaningful conclusion to an overall industry problem.

3.3 DESIGN

The following section is comprised of required software and hardware which will assist in development of a proposed data mining application. Also, this section contains a conceptual context diagram which aims to support the current software engineering process, this includes, a visual design illustrating an overall application domain of tools.

3.3.1 DESCRIPTION OF INSTRUMENTS EMPLOYED

The following section includes relevant instrument types which are required to support the development of a proposed data mining application. The following section is divided into five sections. The first section will describe the required computer programming tools, including, integrated development environment (IDE) tools and various software development kits required for integrating with other tools, as illustrated in section 4.3. The second section will describe the database management tools used for storage and manipulation of the spatial data sets. This

will be followed by a section that details public and third-party proprietary components, normally considered as plug-ins to the IDE for geospatial and computational processing of data. The following section will describe the simulated environment that will aid the process of visualising the data mined results in 3D space. Finally the last section will outline the physical hardware components and platforms required for carrying out the primary software development for conducting data mining analysis by chosen research methods, as explained in section 4.5.

3.3.2 REQUIRED SOFTWARE INSTRUMENTS

Sections, 1, 2, 3 and 4 describe various non-commercial tools which can be obtained without any charge and used for development purposes under an open source license. Refer to Appendix 1 for a list of website links to the following open source files.

3.3.3 Computer development tools

Java SE Development Kit (JDK) Bundle, version 6: Java SE (JDK) is a pre-requisite and as is the java computer programming language that will serve as a development platform to the Eclipse IDE and any subset components required by the IDE. Generally speaking, the JDK consists of various Application Programming Interface standards that will aid the process of programming the proposed application. In addition, Java was chosen as a primary language since it is predominantly an open-source platform and capable of interoperability on multiple operating systems. Interoperability may be useful to the end product, especially, during future OS changes proposed by DAFWA or ECU stakeholders.

Tool 1, Eclipse, Jee Galileo, version SR2: Eclipse is an integrated development environment comprising various tools for java developers to create enterprise and online applications;

Tool 2, Project R, version 2.1.1.0: Project R will serve as a primary statistical and data mining analysis tool. Several extension or library packages for carrying out the proposed data mining techniques will be installed within the Project R environment. The extensions outlined below, are composed of predefined

algorithms and mathematical formulas for running data analysis operations on the Peel-Harvey spatial data sets;

R extensions for Cluster Analysis method: k-means, pvclust, mclust and fpc. R extensions for PCA method: psych, nFactor, and FactorMiner;

Tool 3, rJava, version 0.8-3: The rJava tool is a Project R, interface bridge and is based on Java Native Interface (JNI) technology. Using rJava will assist the development process by exposing native Project R operations within the java application.

3.3.4 Spatial database management systems

Tool 4, PostgresSQL, version 8.2: PostgresSQL SQL is an Open source object-relational database system which has a proven architecture and a record of reliability in, data integrity, and correctness. It is capable to run on all major operating systems, including Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), and Windows;

Tool 5, PostGIS, add on to the PostgresSQL: The PostGIS tool, exists as an installation add on, as part of the PostgresSQL installer package, and provides additional support for spatial geographic object manipulation of the PostgresSQL object-relational database.

3.3.5 Graphical GIS framework tools

For development purposes, both of the outlined tools, uDIG and jGRASS, are based on Eclipse, Rich Client Platform (RCP) technology. The RCP is an architecture that allows various open tool platforms, described as plug-in components which are capable of integrating into one unified client application.

Tool 6, uDIG, version 1.2 – RC2 software development kit: UDIG, abbreviated which stands for User friendly, Desktop located, Internet oriented

and Geographic information system. It is comprised of complex analytical functionalities with a flexible graphical user environment. UDIG is composed of various installer packages, for example, stand-alone Windows OS and Linux OS files and many others, it is important that the latest and most stable version is obtained from the public subversion repository.

Tool 7, jGRASS, version 2.0.20060730: JGRASS is an free open source GIS tool based on the uDIG framework, built and maintained by HydroloGIS in collaboration with CUDAM. The jGrass tool consists of various visual and built in algorithms for that navigating spatial data specifically related to hydrology.

3.3.6 Visual simulation of an interpreted data

Tool 8, Processing expert, version 1.0.9: Processing is an open source programming language and environment that allows individuals to developer artistic images, construct computer simulated animations and allow user interactions. Generally processing is used by students, artists, designers, researchers, and hobbyists for learning and teaching fundamentals of computer graphic programming and data visualisation within a creative context.

3.3.7 REQUIRED HARDWARE INSTRUMENTS

The following hardware specifications will be required for the development environment used in this research. The platforms used will be windows platforms, such as: Windows XP, Windows Server 2003 and Windows Vista. The requirements include the following:-

Central Processing Unit (CPU): Intel Quad Core, 2.67 Gigahertz
Random Access Memory: 3.00 GB/s
Hard drive storage: 500 Gigabytes
Hard drive speed: 5400 RPM
Input and output components: Generic mouse, keyboard and an external USB port.
Internet Connection Speed: 12Mbit ADSL connection.

For software development purposes, the outlined hardware is deemed adequate. However, for any testing, design and run time development, the hardware will only meet the minimum required specification. For this reason, section 4.6.1 has outlined a hardware limitation and recommendations to mitigate this problem.

3.3.8 CONCEPTUAL CONTEXT

According to Victor Basili (1992, p.1), one of the important issues in software engineering is related to a complex to engineering process; that being, complex to develop and complex understand. Likewise, this may lead to complexities in development, resulting in errors and development estimation turning into a difficult task, since there is a lack of understanding of the implications of the changes. To prepare for prevention of any unforseen complexities in the current proposed development phase, a careful strategy must be employed to ensure a planned integration of chosen software instruments is required. For this purpose, the current application will follow the chosen software paradigm in conjunction with a conceptual context diagram.

The conceptual diagram (see Figure-14) is comprised of five different contextual regions which are designed to provide a visual design overview describing various applications [see section 4.2] involved that make up an overall application domain.

Context 1: Data set

The data set context illustrates how the spatial data sets are consumed and collected for storage using a centralised database system.

Context 2: Visual

Representation of a main application that will allow the user to interact with spatial data sets in a visual spatial manner in conjunction with existing functions from uDig and jGrass geospatial frameworks. Also, the existing geospatial functions will enable interaction and manipulation of: spatial map layers and water catchment catalogues. In addition, the utilisation of a processing component will provide animated simulation of the water catchments. For example, the effects and impacts of future trends surrounding the salinity issues, such as streamline of salinity chemical streamlines. However, the simulation may only be performed upon a completed data mining analysis of data sets.

Context 3: Data mining

The representation of a primary data mining process for conducting proposed data mining methods, as explained in detail in section [4.6].

Shared context 1: Data set, visual and data mining

The shared context 1: illustrates a shared functionality of database management between context 3 and context 2. For example, the visual functionalities of an application may require non-data mining database functionalities for performing query or transactional operations such as: add, delete, view and update of records.

Shared context 2: Visual and data mining

Represents the functionality shared between visual, context 2 and data mining, context 3. Aside from the data mining tasks carried out in the following context, the user need may request the Project R environment to create various graphical outputs, for example, graphical charts, sequence of GIF images and other graphical functions supported by Project R.

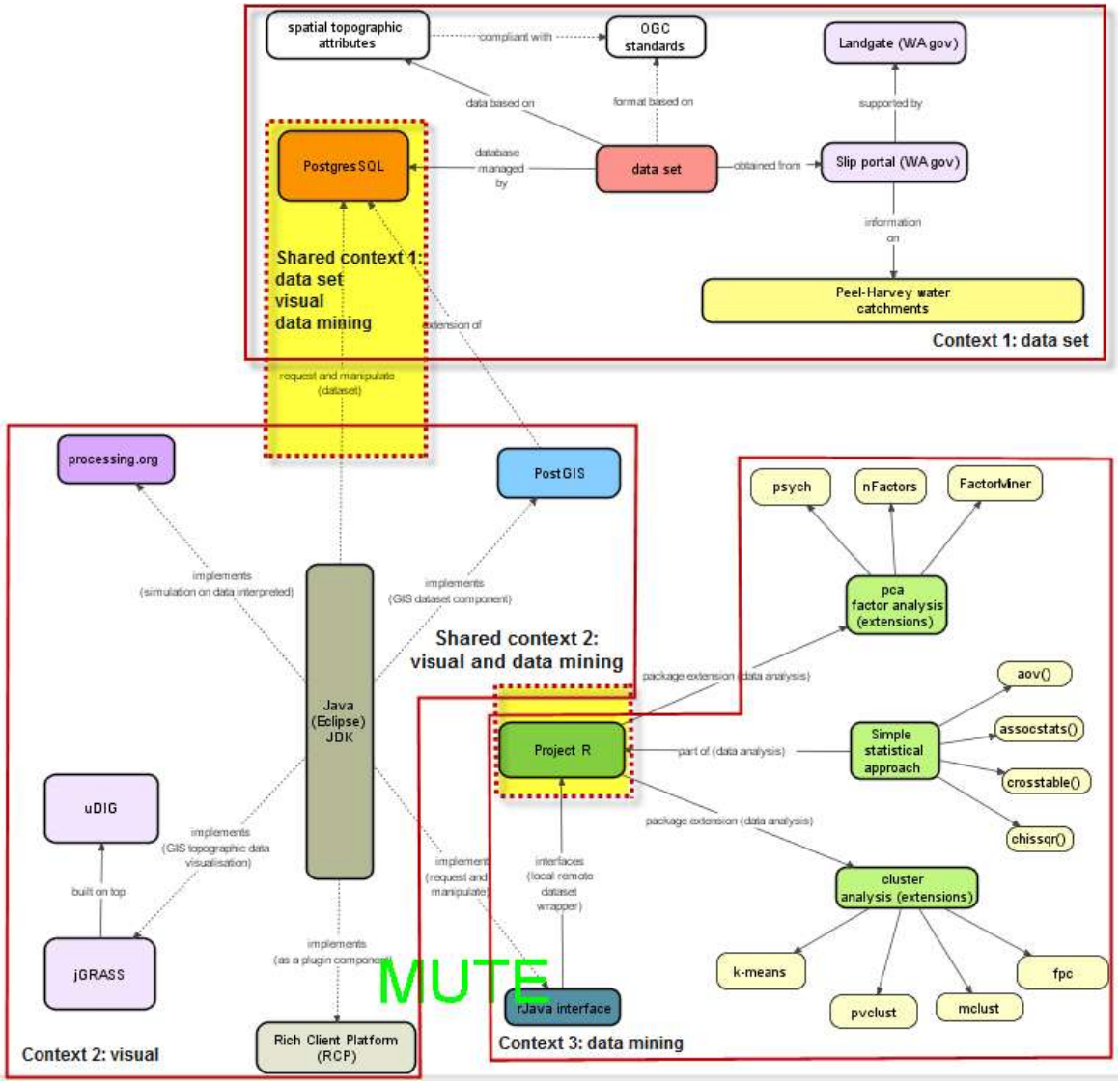


Figure 14: Conceptual context of an application domain.

3.4 DATA COLLECTION PROCEDURE

The collection of the Peel-Harvey water catchment data was previously conducted by various Western Australian government data custodian groups, such as DAFWA (www.agric.wa.gov.au) and the Landgate agency (www.landgate.wa.gov.au). The Landgate agency, being responsible for maintaining the geo-scientific land and property data and their services provides an online services using: Shared Land Information Platform (SLIP) portal (https://www2.landgate.wa.gov.au/web/guest). In addition, the SLIP (2008, p.11) is a service comprised of centralised sharing facilities which provide the following services to data consumers or users.

- Service for consumption by a web browser, such as: GIS applications or Business Solution Applications;

- Data Service that allows various data consumers to view and interrogate different land and geospatial information, however, this may depend on the given access roles; and

- Supports compliant OGC standards for services which feature both Web Map Service (WMS) and Web Feature Service (WFS).

The SLIP agency in collaboration with DAFWA's spatial requirements is a primary data custodian capable of data conversions of Peel Harvey water catchment data. For example, this requires the existing water catchment data samples to be converted into an electronic spatial data set file formats; this enables various GIS and data mining tools to perform data analysis. In detail, this conversion process involves, geographic data being represents topographic layers and regions as compliant geospatial format. Further detail is explained on the GIS file formats in section 4.4.2.

Consequently, an access to SLIP portal must be established using an authorised agreement to establish authorised data downloaded between involved researcher of ECU and DAFWA. In doing so, this will provide a sufficient online service to the user for downloading an electronic version of spatial file formats or better known as data sets.

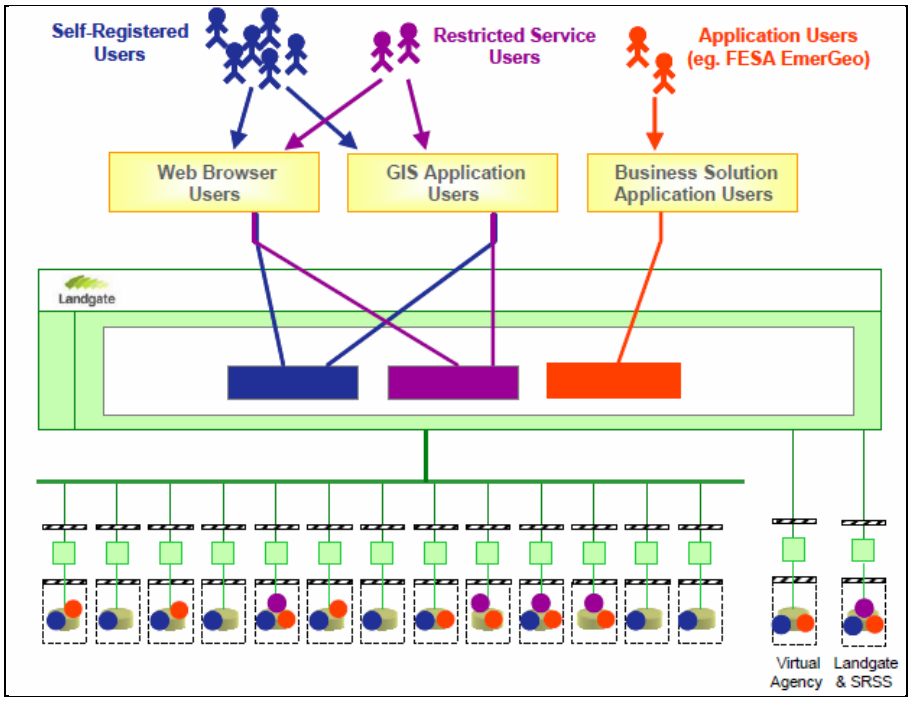


Figure 15: Multiple data services available to data consumers (Shared Land Information Platform, 2008, p11).

3.4.1 SPATIAL META-DATA FEATURES

SLIP service holds a repository with over 350 data sets, stretching over thousand of megabytes in spatial information. Each data set is categorically allocated, depending on the spatial format and region. Some of the categories may include spatial meta-data information on the following areas: cadastre, geodetic, imagery, tenure, topography, roads and fire affected areas (SLIPc, 2010). For the purpose of this project, it is essential that topographic category is obtained, especially since, the current study will emphasise the complex cultural and agricultural matters of Peel-Harvey water catchment regions. For example this will be achieved by ensuring that specific Hydrographical (SLIPa, 2010) and ground surface (SLIPb, 2010) meta-data features are extracted from a data set, as outlined below.

Hydrographical meta-data features: The hydrographical features include measurement and description of waters such as, water resources that are useful or potentially useful to humans (SLIPa, 2010).

> ***Coastal flat:*** Water features which describe water areas along the coast, such as: saline coastal areas, mangroves and intertidal areas;

Framework: Topographic features that specific to the Western Australian land and sea boundaries;

Fuzzy water: Image features that represent a spatial extent on hydrographical features in, water lines, water points, swamps, water polygon, rivers, seas and many more;

Inland flat: Water features that relate to low lying areas in the interior of a country, for example, areas subject to flooding and inundation; and

Inland water: Water features that represent the interior features of the country, such as: water course, rivers, channels, drains. Also, water points feature that represent water bodies such as: clay pans, earth dam, estuaries, lakes, pools, reservoir, water course and wash, waterfalls and water point structure.

Ground surface meta-data features: ground surface features represent characteristics of land surface areas of Earth, both which may be exposed and underwater (SLIPb, 2010).

Elevation: a topographic feature which describes the elevation, such as: land surface contours, bathymetric contours, spot heights and sounding elevation points;

Fuzzy land: Imaginary features which represent spatial landforms, such as: bank lines, beaches, capes likes, cape likes, depression likes and many more; and

Morphology: a set of topographic features which describe various landform characteristics, such as: breakaways, cliffs, ledges, san ridges, reefs, rocks, craters and many more.

3.4.2 VECTOR GIS FILE FORMATS

Current SLIP portal service integrates various data sets file formats which follow an Open Geospatial Consortium (OGC) standards, for this purpose, SLIP services are strictly conformed in compliance to the OGIS document 01-068r3 Web Map Service (SLIP, 2008, p.15). One of benefits of this relates to integrating existing spatial GIS standards features, specifically the ability that provides a unification of disparate data sets to represent a complete picture of an overall situation (ESRI, 2003, p.3).

The required GIS file formats for the current research are dominantly based on Vector data formats. For example, using the SLIP data custodian portal, the following GIS file formats will be obtained:

General Mark-up File (GML): an open and already published GIS format that is used by various vendors and is considered a logical choice for encoding common spatial data exchanges. The GML file format consists of XML schema packages that can be used to encode geometry and its properties attributes independent of a data or content model (OGC, 2003, p.3).

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema targetNamespace="http://ogr.maptools.org/" xmlns:ogr="htt
<xs:import namespace="http://www.opengis.net/gml" schemaLocation="ht
<xs:complexType name="FeatureCollectionType">
  <xs:complexContent>
    <xs:extension base="gml:AbstractFeatureCollectionType">
      <xs:attribute name="lockId" type="xs:string" use="optional"/>
      <xs:attribute name="scope" type="xs:string" use="optional"/>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
<xs:element name="DMP_011_4283_data_SHP" type="ogr:DMP_011_4283_data
<xs:complexType name="DMP_011_4283_data_SHP_Type">
  <xs:complexContent>
    <xs:extension base="gml:AbstractFeatureType">
      <xs:sequence>
<xs:element name="geometryProperty" type="gml:GeometryPropertyType"
```

Figure 16: GML XML schema file content of Geological Map of WA (SLIPd, 2010).

Shapefile (SHP): The shapefile (SHP) is composed of non-topological geometry and meta-data information for the spatial feature characteristics in a data set. In addition, the shapefile is compromised of various vector coordinates. For example, the Cartesian coordinates (ESRIa, 1998, p.32), generally recognised by x and y identifiers which assist the shapefile in representing geographic features using vertices. Likewise, the shapefile requires less disk space and therefore is considered easier to work with. For instance, shape files are a preferred file format for carrying out complex manipulation of spatial data using computer programming tools (ESRIa, 1998, p.3)

3.5 DATA ANALYSIS

The following section introduces sequential steps for conducting the various data mining approaches. The proposed methods will be developed and packaged as a component in order for the proposed application to utilise its software engineering

plug-in features. Likewise, the proposed application will utilise the proposed data mining techniques in conjunction with the research strategy, as described below. In doing so, the following research activities are envisioned to form a general guide of steps for carrying out a data mining data analysis in relation to research hypothesis.

Primary question: *"How can the use of data mining techniques be used to interpret trends in Western Australian water catchment land use?"*

3.5.1.1.1 ***Subset question one:*** *"Which data mining techniques are the most appropriate for analysis of water catchment data sets"*

Subset question one will be answered by research activity 1 and 2.

3.5.1.1.2 ***Subset question two:*** *"How can data mining techniques be used to make informative predictions in relation to changes in land use and climate"*

Subset question two will be answered by research activity 2 and 3.

3.5.2 RESEARCH METHOD STRATEGY

The following matrix diagram illustrates a strategy of methods employed in the current study in order to interpret future trends of salinity impacts across a longitudinal time periods. The methods chosen will feature the following techniques: classical statistics, cluster analysis and principal component analysis. The matrix diagram illustrates a cross selection of benchmarks. Furthermore, the benchmark will extract the result conducted during the data analyses in order to establish validation of results and evaluate the effectiveness of impacts in relation to a topographic category, such as: rainfall meta-data, crop yield meta-data, nutrient run-off meta-data and many more features, as describes in section 4.4.

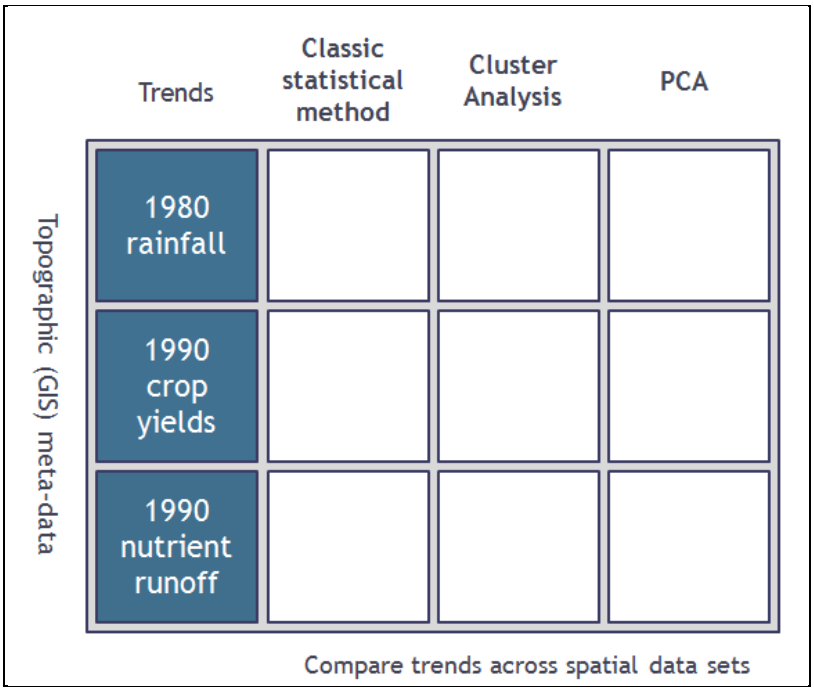


Figure 17: Research method strategy employed for data set analysis.

3.5.3 PRELIMINARY STEPS

The following section defines the preliminary steps that must be performed to the water catchment data sets. Steps 1 to 3 are a mandatory process; however the data cleaning step 4 is an optional pre-processing data mining task, since, it is assumed that not all data is required for a cleanup process. This may be true, in the case of a data set been already pre-processed by the data custodian or the current application user.

Step 1: Spatial data set retrieval

All the data sets, related to the Peel-Harvey water catchment will be retrieved using the SLIP portal (see section 4.5) and then stored in either a PostgresSQL server or stored locally in a computer file system directory. Using locally stored data may be a quicker process; however this will result in double-handling and during the course of these actions, may becoming corrupted, in particular during preliminary data manipulation. Therefore, it is recommended that data is stored directly to a database. Consequently, Project R environment will assist in the initialisation and cleaning of data sets, as explained in step 3 and step 4

Step 2: Data set file integrity check

Binary size of each spatial file will be verified against an original consumed file. Generally, this will be carried out using existing java file utility packages.

Step 3: Initialisation of GIS data sets

The initialisation will be performed in conjunction with the existing geospatial tools found in uDIG and jGRASS framework by an automated process on data set content. For example, a data structure process will be programmed to parse all the relevant topographic attributes located inside of GML and SHP files.

Step 4: Data set cleaning

The next step will involve an invocation of a Project R interface environment through the rJava component. In doing so, common data mining pre-processing processes of data cleaning, as outlined in section 2.3 will be carried out. This will evolve, invoking various Project R functions, such as, R scripts comprised of command for performing a removal of empty longitude and latitude attributes. Furthermore, default empty fields will be replaced with default predefined complaint values or, in this case, depending on the complexity of a data set, the entire data set row may be removed.

3.5.4 RESEARCH ACTIVITY 1: CLASSIC STATISTICAL APPROACH

Since the majority of pre-processing and data analysis tasks in relation to data mining process, will be utilised, using Project R environment through an invocation of an rJava component. In doing so, this will provide complete environment of Project R functions required for conducting a classic statistical data analysis. For example, the following R functions describe specific steps that will be incorporated into an application and performed autonomously. This involves:

sapply() function: will be used to return a vector or matrix of data (R, 2009, p.225), depending on the statistical functions used. For example, sapply will call the following functions: mean standard deviation, variance, minimum, maximum, median and range.

table(),ftable and xtab functions: will aid in establishing a frequency of a data set, for example, this includes creating a variance in two or three dimensional space of data. Using the xtab function the chi-squared test will be run to establish the independence of factors found in the data set (R, 2009, p. 458).

rchisq() function: will be used to establish a chi-squared distribution for statistical tests of categorical quantitative data, for instance, various

topographic features, as explained in section 4.4.1. However, depending on the spatial data length, the n distribution may be set to run in 1000 sample records per test.

$$\chi^2 = \sum_{i=1}^{n} Z_i^2$$

Figure 18: Chi square distribution test formula (Weisstein, E. W, 2003, p.995).

3.5.5 RESEARCH ACTIVITY 2: CHOSEN DATA MINING TECHNIQUES APPROACH

In addition to the classical statistical functions, as explained earlier, the research hypothesis will perform further validation of effectiveness using Cluster Analysis and PCA methods. This will help in validating whether any useful patterns can be established, specific to dryland salinity in water catchment of Peel Harvey region.

3.5.5.1 APPLYING CLUSTER ANALYSIS

It is optional for the preliminary data cleaning steps to be carried out; especially since data pre-processing may share some similarities. Cluster analysis will be broken up into five steps, in order, as follows:

Cluster analysis, step 1: (Data preparation): before a benchmark is conducted, the data is required to undertake a data standardisation process (see Figure-19). This involves an invocation of the following R functions on geospatial data set file formats, as explained in section 4.4.

Run a ***na.omit function*** to handle the missing values from any attribute objects found in the data set, In this case, the data set is chosen as being a GML and a SHP file (R, 2009, p.1208).

Run a ***scale function*** to perform data unique data standardisation *by* cantering or scaling the data set columns as a numeric matrix representation (R, 2009, p.375).

Figure 19: Cluster analysis, data preparation step.

Cluster analysis, step 2: *(**Partitioning):*** using the previously transformed matrix data, the next step will perform a k-mean clustering on the data matrix (R, 2009, p. 1152). For this purpose following steps must be performed in order (see Figure-20).

Partitioning, *step 1:* aggregate regions of clusters in a data set.
Partitioning, step 2: run the k-means function for each clustered region.
Partitioning, s*tep 3:* run a mean variation on each cluster region

Figure 20: Cluster analysis, partitioning step.

Cluster analysis, step 3: *(**Hierarchical agglomeration clustering):*** will create an agglomeration of clusters using hierarchical representations to classify dissimilarities of objects. This will involve a twofold invocation process as displayed in Figure 21.

Run a ***dist function*** on the data set and pass a ***Euclidean*** *distance* measure to compute the distance between the clusters in each matrix dimension (R, 2009, p. 1078).

Usage: dist(PeelHarveyDataSet, method = "euclidean")

Run a ***hclust function*** and provide a relational distance parameter and an agglomeration centroid parameter type (R, 2009, p. 1127).

Usage: hclust(d, method = " centroid ", members=NULL)

*Parameter **d**:* is a dissimilarity structure as produced by dist in step a.
*Parameter **method**: is* a set of agglomeration types, it can also support other relational distance types, such as: ward, single, complete and average.

Figure 21: Cluster analysis hierarchical agglomeration clustering step.

Cluster analysis, step 4: *(**Model clustering**)*: will utilise a ***mclust function*** to produce clustered regions and pass an expectation maximisation (EM) for final cluster analysis computation (see Figure-22).

Run a ***mclust function*** to create a **Gaussian** mixture model according to the EM initialisation (Ra, 2010).

Usage: mclust(PeelHarveyDataSet)

Figure 22: Cluster analysis, model clustering

Cluster analysis, step 5: (plotting of clusters): will use the ***clusplot* function** from a cluster package to construct a ***bivariate*** plot for visualising specified clustered regions of a data set (see Figure-23). Each representation will be represented as a set of points in the graph against each found principal component (Project R, 2010b).

Run a ***clusplot function*** on an entire summarised data sets, this includes, analysed data set summary of data from steps 1 to 4.

Usage: clusplot(PeelHarveyDataSet)

Figure 23: Cluster analysis, plotting of clusters.

3.5.5.1.1 APPLYING PRINCIPAL COMPONENT ANALYSIS

The final step will involve applying the PCA method that aims to establish a variance amount of principal components found in a data set. This section is broken up into two steps, in order as follows:

PCA, step 1: (Produce Principal Components): will utilise a ***princomp function*** to determine possible principal components from a given data matrix (R, 2009, 1287). Consequently, this will invoke a ***biplot function*** to visualise the principal components as plots on a graph (see Figure-24).

Principal component, step 1: run ***princomp function*** on an entire aggregated data set in conjunction to previously performed steps, as explained in cluster analysis section: 4.5.4.1.

Usage: princomp(*PeelHarveyDataSet*, cor=TRUE)

Principal component step 2: Invoke the ***biplot function*** on an aggregated data set.

Usage: biplot(*PeelHarveyDataSet*)

Figure 24: PCA, produce principal components.

PCA, step 2: (Factor Analysis): will utilise a ***factanal function*** (see Figure-25) to produce a maximum-likelihood in factor analysis on a covariance matrix data set (R, 2009, 1091).

Run the **factanal function** on an “n” entire data set collection by specifying the amount of factors required for factor analysis. By default the varimax option will be used for the analysis for rotation of factored data.

Usage: factanal(PeelHarveyDataSet, factors, rotation="varimax")

*Parameter **factors:*** number of factors to be analysed.
*Parameter **rotation:*** function to be used to rotate the factors.

Figure 25: PCA, factor analysis.

3.5.6 RESEARCH ACTIVITY 3: VISUAL GIS FILTERING

Different geographic regions of a data set may need to be re-sampled to ensure that only specified map regions are used for effective data analysis. This will enable the user to choose a specific water catchment region for data analysis. Such functionality will be inherited from an existing uDig and jGrass GIS framework tools. Generally, the selection process will be carried out using the “pan” selection functions on a visual map area of south west, as illustrated in Figure 27 which represents topographic series map, with map dimensions of 1:25 000.

The following visual interaction steps will take place, as follows, in order:

Visual GIS Filtering, step 1: allow the user to select specific region located in the Peel-Harvey water catchment and selected corresponding boundaries by predefined topographic region ID’s, as provided by the SLIP data custodian service.

Visual GIS Filtering, step 2: the data mining application will compute and aggregate only the filtered regions in a series of automated steps.

Visual GIS Filtering, step 3: notify the user that selected regions are filtered with an option to commence data analysis steps, as explained in earlier research activity sections, 4.5.3 and 4.5.4.

Figure 26: Visual GIS filtering process steps.

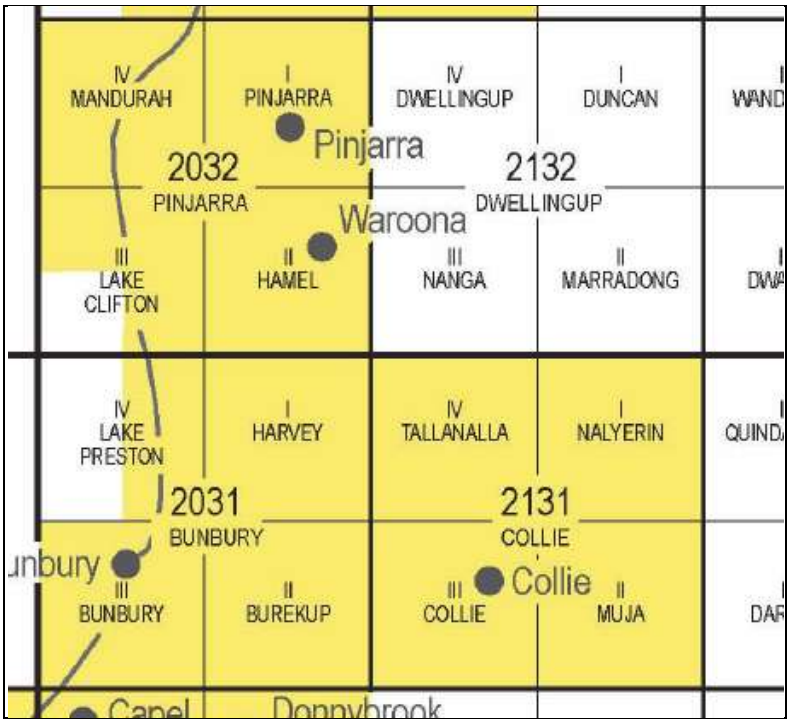


Figure 27: Dola topographic series 1:25 000 map dimension (SLIPf, 2009).

3.5.7 TIMELINE OF RESEARCH ACTIVITIES

	Feb 2010	March 2010	Apr 2010	May 2010	Jun 2010	Jul 2010	Aug 2010	Sep 2010	Oct 2010	Nov 2010	Dec 2010	Dec 2012
1. Establish Project and data mining techniques research												
2. Review of Literature into data mining techniques and tools												
3. Conceptual data mining instrument setup												
4. Acquisition and Collation of Data Sets												
5. Data Analysis and mining of Data Sets												
6. Standardise data sets												
7. Analyse and data mine data sets												
8. Incorporate research activities into the instrument												
9. Analysis and interpretation of Research Findings												
10. Software test maintenance and deployment	`											
11. Thesis writing and completion												

Figure 28: Timeline of research activities

3.6 LIMITATIONS

3.6.1 HARDWARE LIMITATIONS

It is assumed that higher hard drive speed may be required, for instance, it is assumed that anything above an aging 5400RPM specification is preferred to achieve robust performance of data computation. For example, a 7000RPM or 10000 RPM is preferred. This will ensure that larger amounts of traffic are analysed more quickly and robustly, especially during an analysis stage.

In addition, the random access memory (RAM) may also be a limited factor to an overall application performance. Especially, since the application may be set up to run data analysis on a local machine. Similarly, it would be preferred for the pre-processing and data mining process to be scheduled remotely using the rJava remote method invocation requests. In doing so, this will prevent the local client system from being exposed to stressful situation in prioritising hardware resources.

3.6.2 SIMULATION OF GIS DATA

Due to the current project limited time line, it is envisaged that visual simulation of spatial data may not be carried as originally planned for interpreting the simulated salinity current and future trends. Originally it was established that processing expert tool would provide this experience to the user, thus, this functionality may not be fulfilled during the current time frame. However, due to the integration of an existing spatial framework tools, uDig and jGRASS and conformance to object-oriented (OO) standards, therefore it is believed that current implementation process will enable future work to be continued incrementally (see Figure 28).

4 RESEARCH ANALYSIS

The contents of this chapter is based on the peer reviewed special paper presented at the data mining conference for the 14th Pacific-Asia conference on Knowledge Discovery and Data Mining for interrogation of water catchment data sets using data mining techniques (Sehovic, Armstrong and Diepeveen, 2010, Appendix B).

A component based software tool has been designed and prototyped which integrates the tools described above (as shown in Fig. 2). This tool will integrate a data set component, visualization component, data set, visual and data mining components.

Data set context: The data set context illustrates how the spatial data sets are consumed and collected and used for storage using a centralised database system.

Visual Context: Representation of a main application will allow the user to interact with spatial data sets in a visual spatial manner in conjunction with existing functions in uDig and jGrass geospatial frameworks. Also, the existing geospatial functions will enable interaction and manipulation of: spatial map layers and water catchment catalogues. In addition, the utilisation of a processing component will provide animated simulation of the water catchments, for example, the effects and impacts of future trends surrounding the salinity issues, such as streamline of salinity chemical streamlines. However, the simulation may only be performed upon a completed data mining analysis of data sets.

Data mining context: The representation of a primary data mining process for conducting proposed data mining methods.

Data set, visual and data mining: The shared context 1: illustrates (see Figure 29) a shared functionality of database management between context 3 and context 2. For example, the visual functionalities of an application may require non-data mining database functionalities for performing query or transactional operations such as: add, delete, view and update of records.

Visual data mining: Figure 29 represents a functionality shared between visual, context 2 and data mining, context 3. Aside from the data mining tasks carried out in

the following context, the user may request the Project R environment to create various graphical outputs such as: graphical charts, sequence of GIF images and other graphical functions supported by Project R.

4.1.1 APPROACH ON KEY ACTIVITIES

The key activities in carrying out the interrogation of the Peel Harvey catchment datasets was undertaken primarily with the assistance of uDig software and PostgreSQL database. For this purpose, the processed activities involved a pre-process of a large geospatial shapefile of Peel-Harvey region and transporting the data into the postgresql database for post-process data mining tasks which involve utilising Project R and running various R packages for visualising with cluster analyses. Figure-29 demonstrates this conceptual approach, followed by the step-by-step process as illustrated and explained in the proceeding section of this paper.

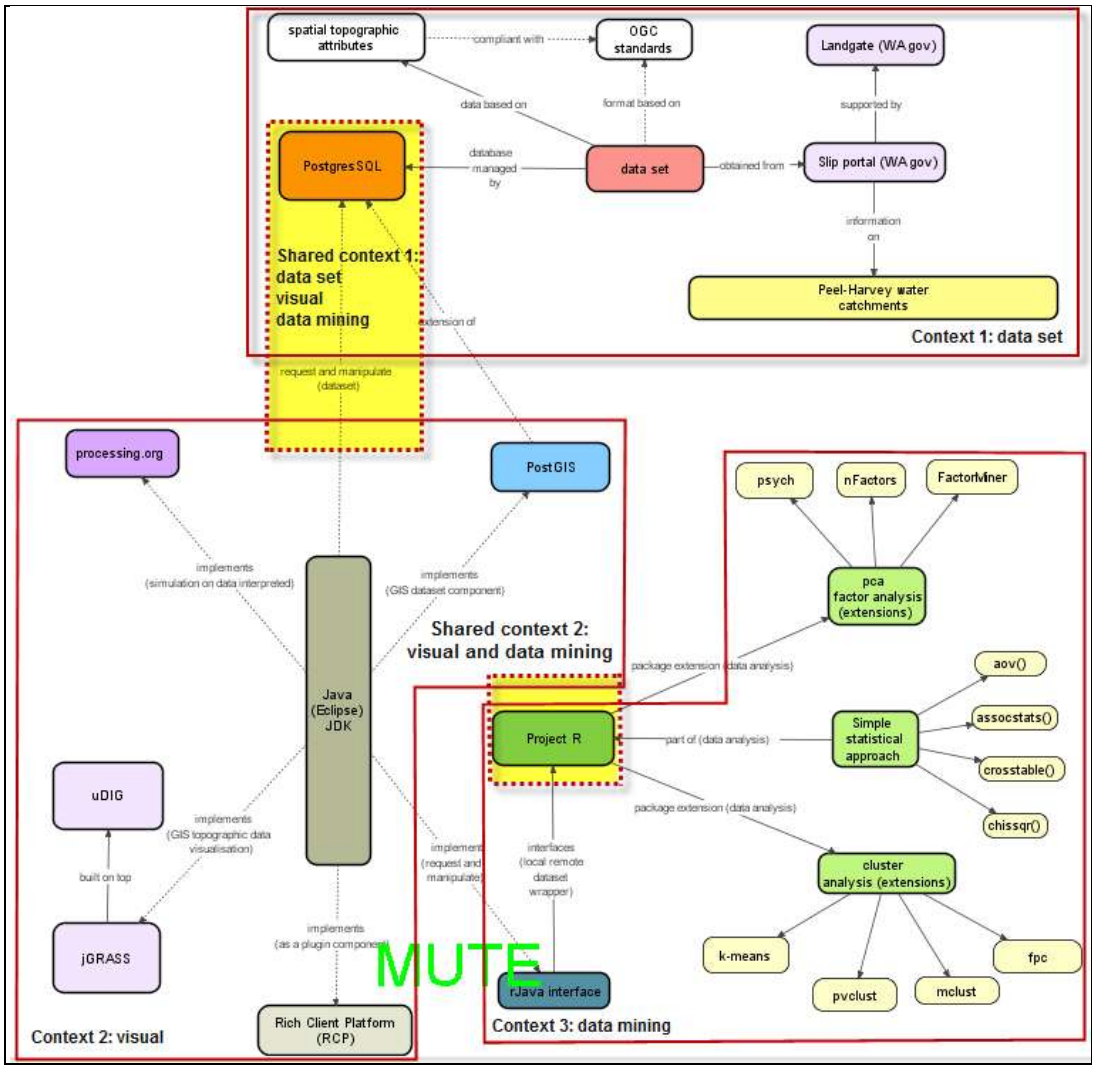


Figure 29: Overview of the data mining context

The Peel Harvey data set was prepared for two regions, Collie and Pinjarra sub region of the Peel Harvey catchment. This data set was composed of SHP files with 2:250 000 resolution. The files were imported and catalogued using the uDIG software. This import process can inter-connect multiple layers with the parent layer. The parent layer was collie2 which is represented in green colour (see Figure 30). While the Pinjara2 layer is placed on top of the parent layer and represented in the yellow colour (see Figure 30). Shapefiles are imported into the uDig software.

4.1.2 SELECTING BOUNDED BOX REGIONS FOR PINJARA LANDSCAPES

A specific subregion can be selected using the uDig software. For example, using Collie 2m 250K - Shape file meta-data, it is possible to select specific regions using the “Info” function and selecting a region on the map. Alternatively, using the “border region” selection function from a toolbar section, we can select a boundary (see Figure-31 and Figure-32). Also, note reach time the region is selected; the corresponding meta-data is also selected and highlighted in yellow. Once the table section is accessed, all the data being selected is temporarily aggregated for further manipulation, for example, data extraction.

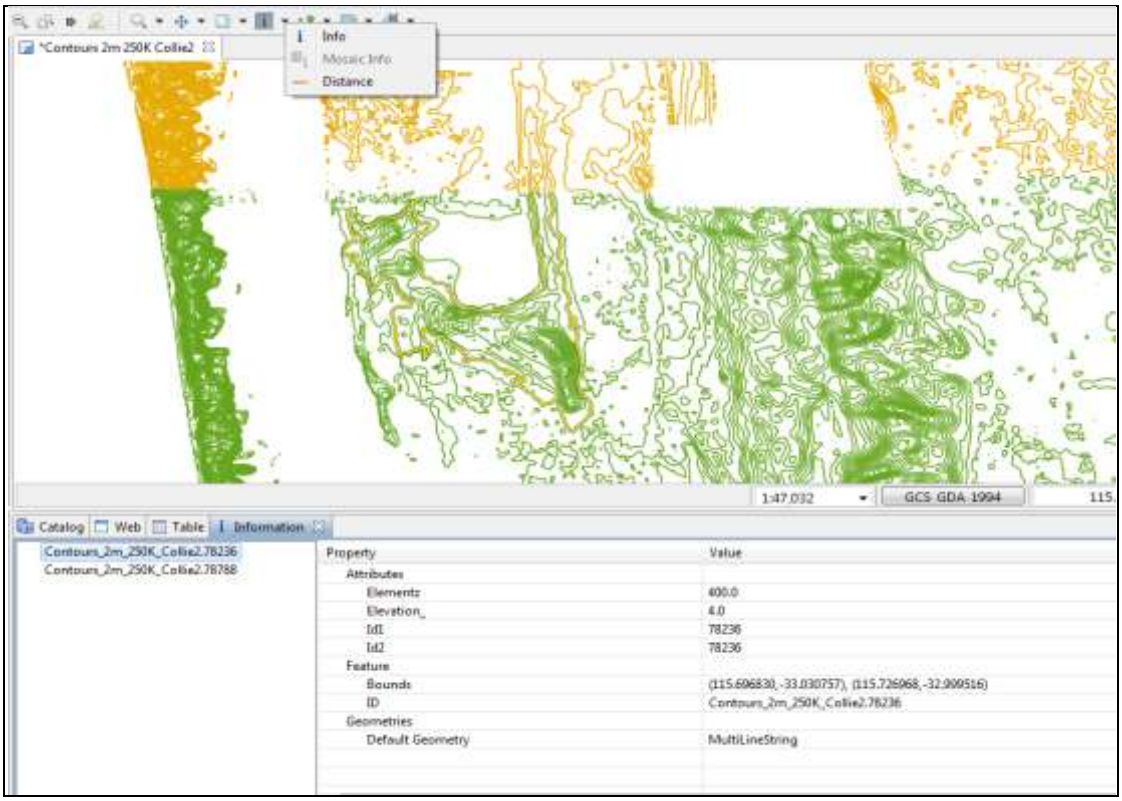


Figure 30: Map information identifier feature using uDIG software

Figure 31: Border selection tool

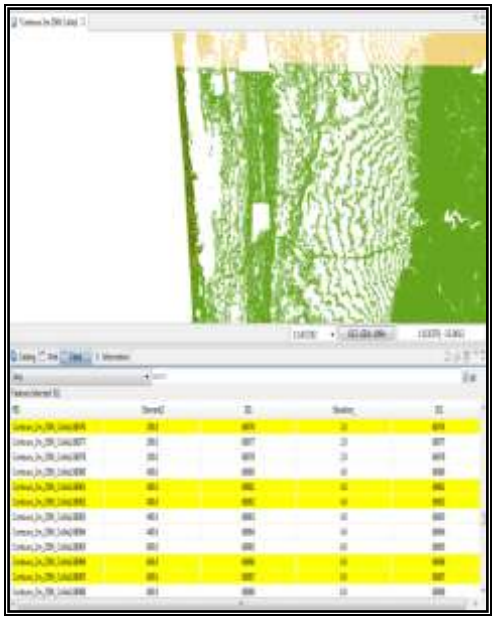


Figure 32: Data extraction using uDIG

The data set was exported using uDIG as a resource shape file. Other data extraction formats are also possible included image files. Data export will result in the production of prj file, wld files and shp files for each layer file.

4.1.3 ANALYSIS OF WATER CATCHMENT DATA

To demonstrate the possible data analysis that can be used to interrogate the water catchment data as a series of cluster analyses were carried out using the R scripting package. Project R Packages required for carrying out cluster analysis , Hclust, mclust, stats, pfc, shapefiles, cluster R packages. The following section details the

processes used to perform sample data analysis manipulation on the selected region of a Collie data set:

4.1.4 Parsing the shape-file using Project R

We first parse the shape file and assign the collie dataset to an object for further manipulation (see Figure-33). It is important that the correct location of the shape file is provided, also, the process may take several second or at most half several minutes, depending on the size of a shape file. For this purpose the shape file is reasonably small, less than a megabyte. In addition, there is no need to provide the extension of a shape file, especially since the shapefile package has distinct features to recognise the file format.

```
#read the shape file and assign it to a an collieDS object
collieDS<-read.shapefile("Contours_2m_250K_Collie2")
```

Figure 33: R-script

4.1.5 Setup a variable table list

Return an actual list that the shape file package has processed. Note that shape file automatically processes the corresponding dbf files. For this purpose, the following list will display a set of dbf objects that correspond with the shapefile (see Figure-34).

```
#returns the list dbf content list of header information. ElementZ, ID1, Elevation_, ID2
list(collieDS$dbf)
```

Figure 34: R-script

4.1.6 Assign variables of interest for clustering the data on

We proceed to create two variables as unique list of data objects. This is required in order for the data frame to be constructed. In addition, assign the graph values as, ID and Elevation. Note, appending the dbf%ID2 or Elevation_ keywords to the collieDS string will implicitly access the meta-data attributes (see Figure-35).

```
# variable list
varID   <- list(collieDS$dbf$dbf$ID2)
varElev <- list(collieDS$dbf$dbf$Elevation_)
```

```
# aggregated data into a frame object consisting of (ID and Elevation)
collieDSFrame <- data.frame( a=varID, b=varElev, c=c('ID','Elevation'))
```

Figure 35: R-script

Using the hclust and stat package, a hierarchical agglomerative graph was created using the Euclidean distance matrix representation (see Figure-36).

```
#Create HIERARCHICAL AGGLOMERATIVE
distanceMatrix <- dist(collieDSFrame, method = "euclidean") # distance matrix
fit <- hclust(distanceMatrix, method="ward")
```

Figure 36: R-script

4.1.7 Running MClustering Data Mining

A model based clustering was created using the mclust R package library as described following (see Figure-37).

```
# Model Based Clustering
library(mclust)
fit <- Mclust(collieDSFrame[-3])
plot(fit, collieDSFrame[-3]) # plot results
```

Figure 37: R-script

As a result of this, the following four diagrams are displayed (see Figure-38, Figure-39, Figure-40), mclust, Bayesian Information Criterion (BIC) classification, direct classification plot, uncertainty classification and density contour plot. These plot provide an example of the clustering techniques that can be used to interrogate the spatial data sets.

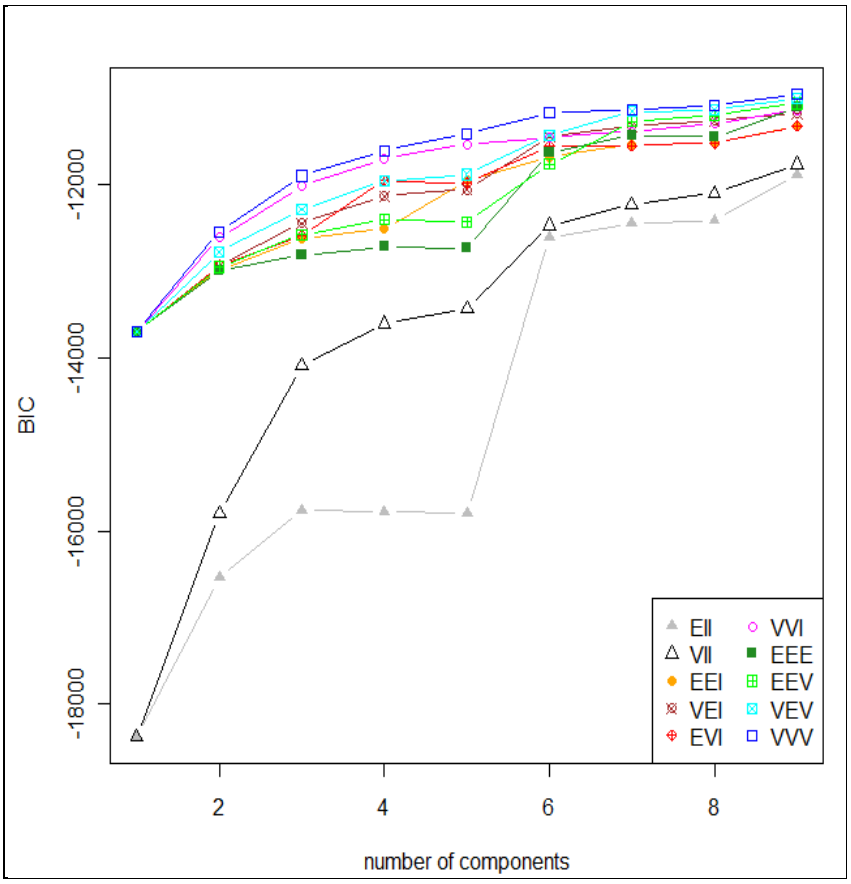


Figure 38: Example of BIC cluster plot produced from Rscript

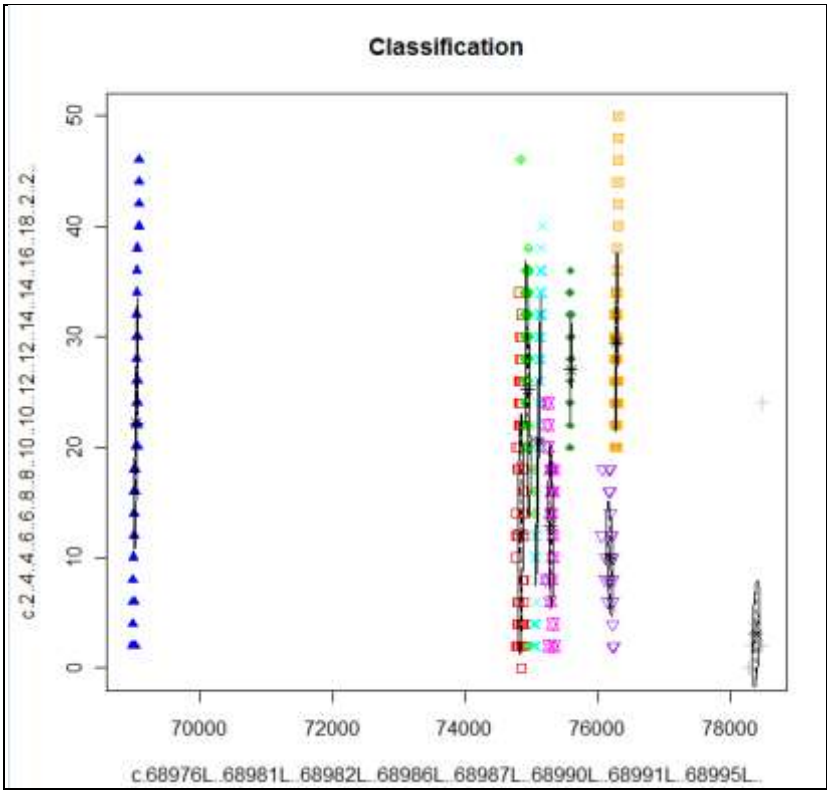


Figure 39: Example of MClust cluster classification produced from Rscript.

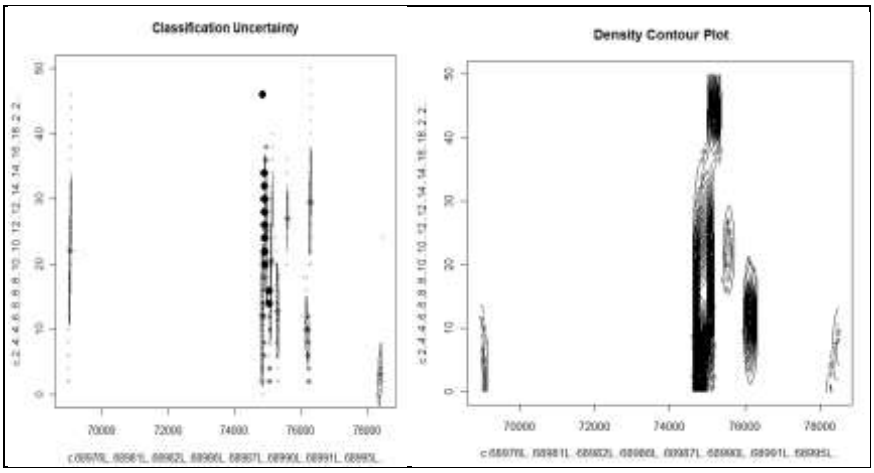


Figure 40: Example of MClust cluster uncertainty plot produced from Rscript and an example of a MClust density contour plot produced from Rscript.

Using the cluster and fpc package, a k-means cluster with 5 clusters from a set of existing collie data source data was created (see Figure-41). A clustered plot against first and 2nd principal components was also created. For this purpose, the elevation and ID are taken into context of computation of clusters.

```
# K-Means Clustering with 5 clusters
fit <- kmeans(collieDSFrame[-3], 5)
# plot against two principal components
library(cluster)
clusplot(collieDSFrame[-3], fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```

Figure 41: R-script

A centroid cluster plot (see Figure 43) against the first and second discriminatory functions was created by using the following Rscript (see Figure-42).

```
# Create a centroid plot against the fisrt and second discriminate functions
library(fpc)
plotcluster(collieDSFrame[-3], fit$cluster)
```

Figure 42: R-script

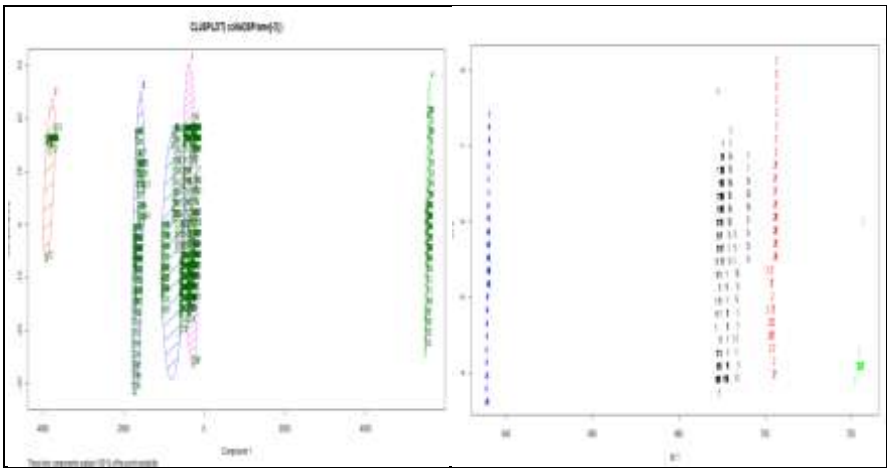


Figure 43: Example of a Clustplot comprised of five clusters produced from an Rscript.

4.2 Transforming the shapefile datasets into relational database tables

At first, the data context currently held under large shapefile datasets under regions of central and and south areas of Western Australia are held in the PostgreSQL database using the Shapefile to PostGIS importer utility.

Name of the raw shape file datasets that were used during the import process:

a. Subsystems_south.shp: 80.0 MB in size
b. Subsystems_centra.shp: 104.4 MB in size

PostGIS ships a free utility for importing geospatial datasets directly into PostgreSQL and creating database indexes and geographical coordinate polygon references (ie shp2psql). This tool, as illustrates in Figure-44 and Figure-45 necessitates the necessary database credentials, such as the username and a password including the PostGIS extension.

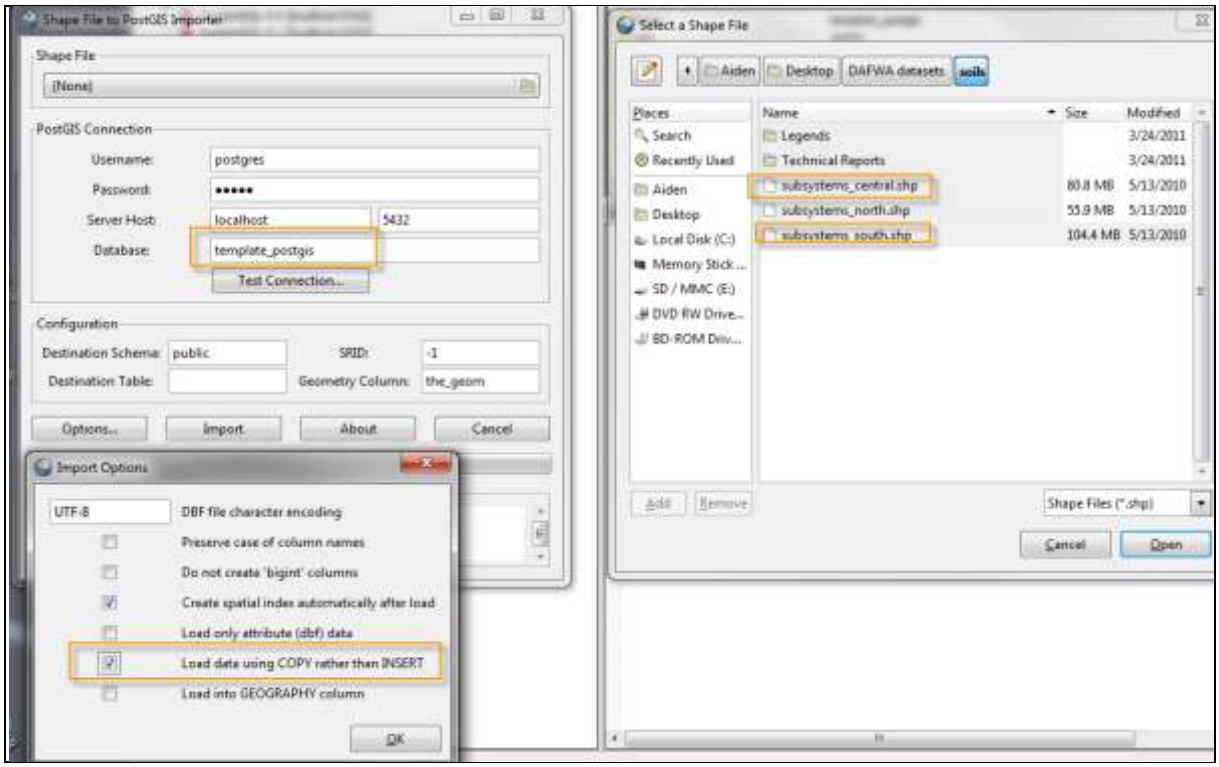


Figure 44: Importing shapefile datasets into PostgreSQL database

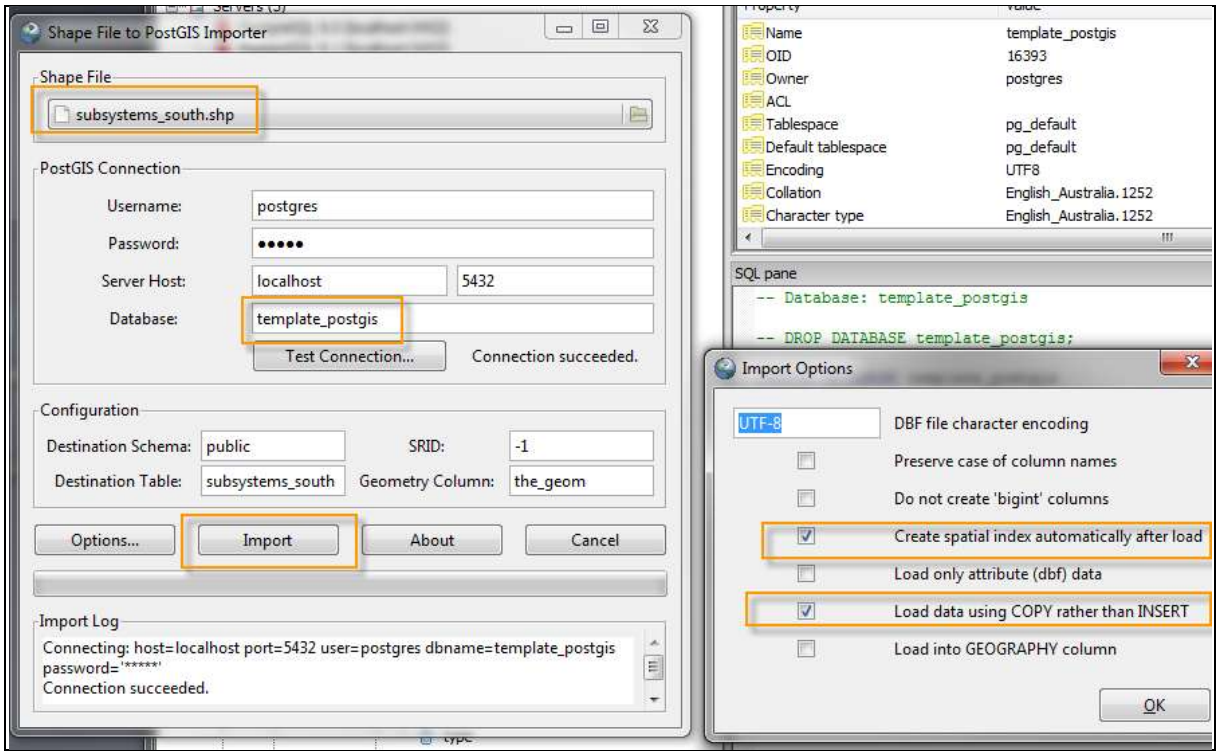


Figure 45: Creating spatial indexes automatically and using COPY function

The transfer of raster shapefiles is processed into a block of tables per dataset. Generally this is achieved by generating table name with the same name as the shapefile. In this instance, subsystems-south.shx and subsystems-north.shx are generated into table names (see Figure-45, Figure-46 and Figure-47).

shp2psql datasource names	Shapefile name	PostgresSQL table name
	subsystems_south.shp	public.subsystems_south
	subsystems_south.shp	public.subsystems_south

Figure 46: Table names using by shp2psql function

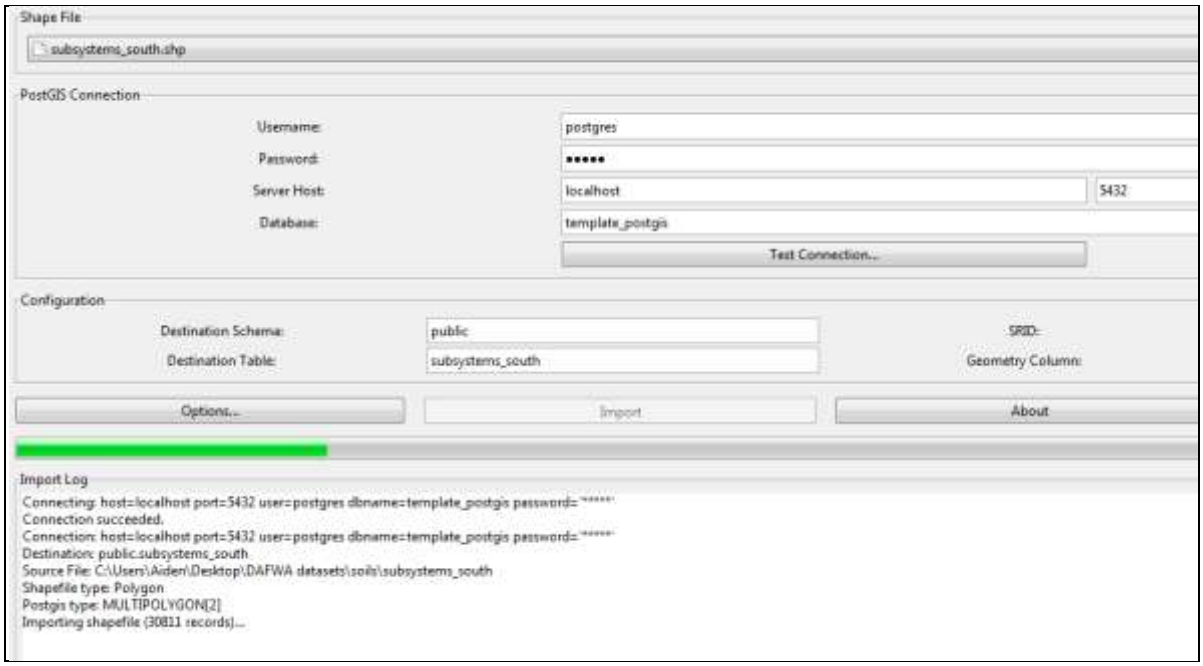


Figure 47: Shapefile Import process

```
Connecting: host=localhost port=5432 user=postgres dbname=template_postgis
password='*****'
Connection succeeded.
Connection: host=localhost port=5432 user=postgres dbname=template_postgis
password='*****'
Destination: public.subsystems_south
Source File: C:\Users\Aiden\Desktop\DAFWA datasets\soils\subsystems_south
Shapefile type: Polygon
Postgis type: MULTIPOLYGON[2]
Importing shapefile (30811 records)...
Creating spatial index...

Shapefile import completed.
```

```
Connection: host=localhost port=5432 user=postgres dbname=template_postgis
password='*****'
Destination: public.subsystems_central
Source File: C:\Users\Aiden\Desktop\DAFWA datasets\soils\subsystems_central
Shapefile type: Polygon
Postgis type: MULTIPOLYGON[2]
Importing shapefile (42850 records)...
Creating spatial index...

Shapefile import completed.
```

Figure 48: Output from shp2psql

4.3 Interfacing with Project R and PostgresSQL database

Applying RPostgresSQL package to the Projcet R package repositories enables R to interface to the PostgresSQL database system. And as such, this enables the user to construct a database question on the existing already imported shapefile dataset tables named “subsystems_south” and “subsystems_central” (see Figure-49).

```
con <- dbConnect(PostgreSQL(), user= "postgres", password="admin",
dbname="template_postgis")
rs <- dbSendQuery(con,"select *, AsGML(the_geom) from subsystems_south
WHERE the_geom IS NOT NULL LIMIT 1000 OFFSET 1001")
out <- dbApply(rs, INDEX = "the_geom",
FUN = function(x, grp) quantile(x$DATA, names=FALSE))
```

Figure 49: R-script

Using the postgresql database, the results from the import can be seen in the following outputs (see Figure-50, Figure-51 and Figure-52)

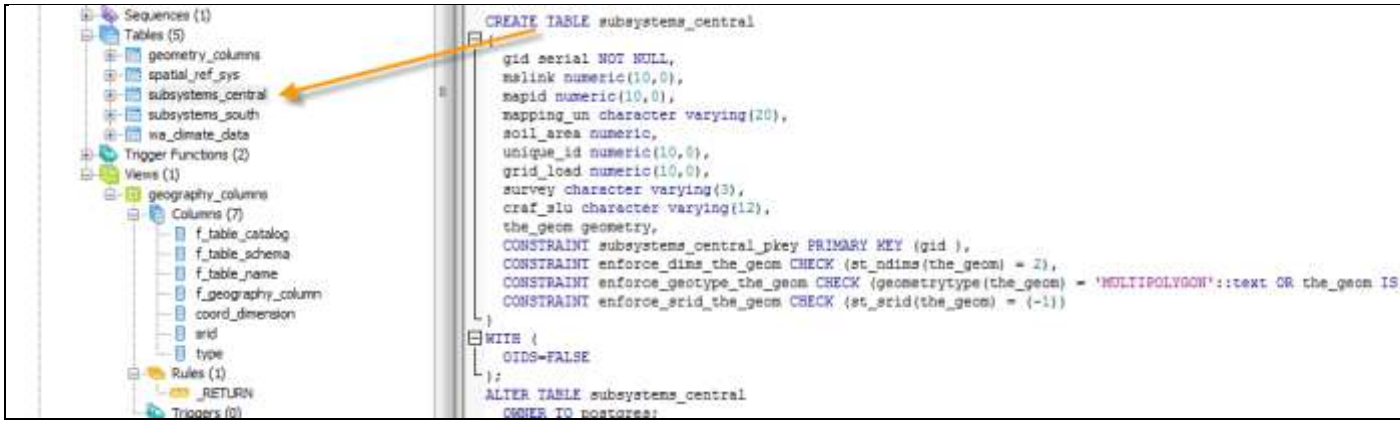
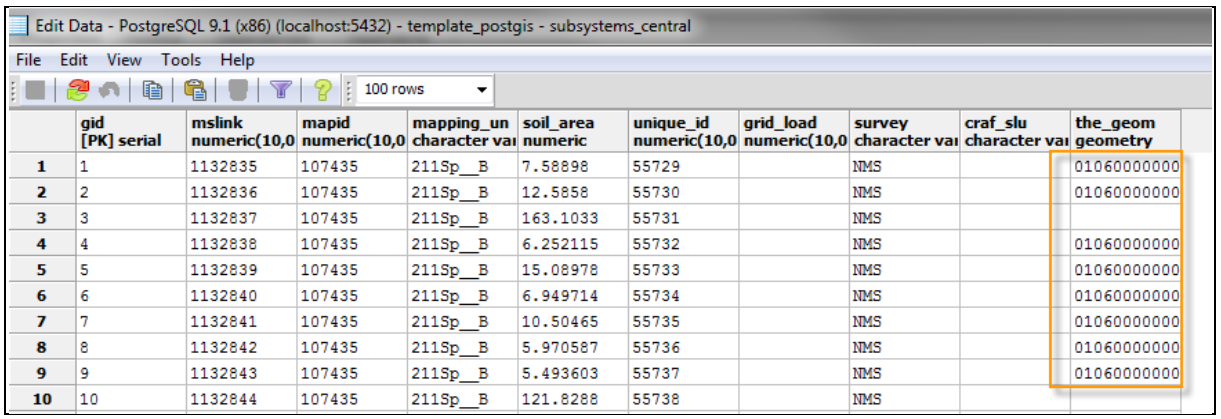


Figure 50: Output from Postgresql (1)



	gid [PK] serial	mslink numeric(10,0	mapid numeric(10,0	mapping_un character var	soil_area numeric	unique_id numeric(10,0	grid_load numeric(10,0	survey character var	craf_slu character var	the_geom geometry
1	1	1132835	107435	211Sp__B	7.58898	55729		NMS		01060000000
2	2	1132836	107435	211Sp__B	12.5858	55730		NMS		01060000000
3	3	1132837	107435	211Sp__B	163.1033	55731		NMS		
4	4	1132838	107435	211Sp__B	6.252115	55732		NMS		01060000000
5	5	1132839	107435	211Sp__B	15.08978	55733		NMS		01060000000
6	6	1132840	107435	211Sp__B	6.949714	55734		NMS		01060000000
7	7	1132841	107435	211Sp__B	10.50465	55735		NMS		01060000000
8	8	1132842	107435	211Sp__B	5.970587	55736		NMS		01060000000
9	9	1132843	107435	211Sp__B	5.493603	55737		NMS		01060000000
10	10	1132844	107435	211Sp__B	121.8288	55738		NMS		

Figure 51: Output from Postgresql (2)

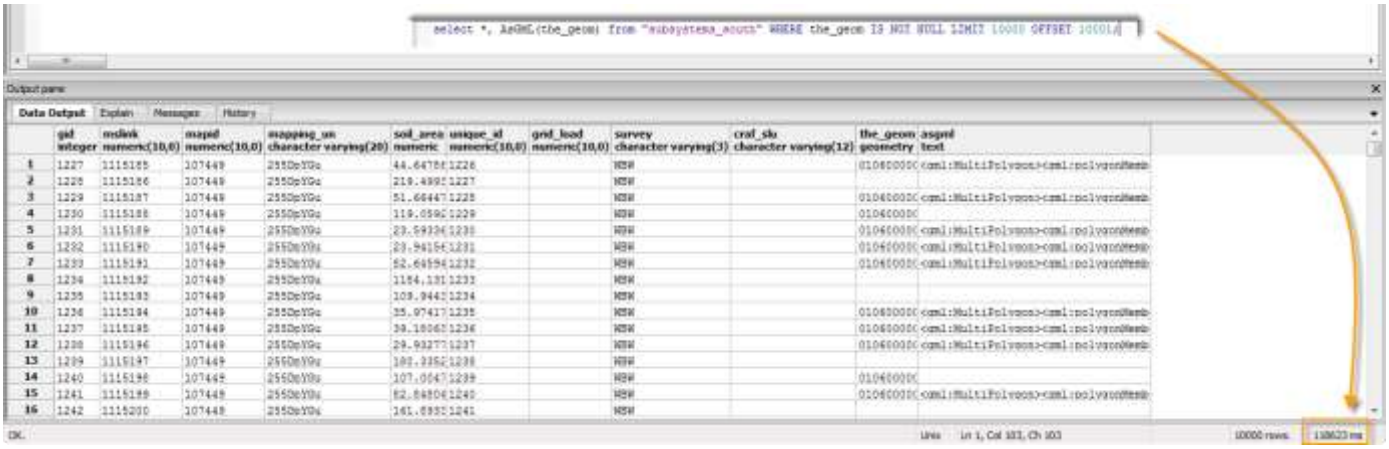


Figure 52: Output from Postgresql (3)

We had to reduced the database query to 10,000 records in order to achieve a reasonable database response. This being, due to the complexity and high load of data currently stored in the datawarehouse comprising of 30811 thousand of records it has taken approximately 2 minutes to fetch 10,000 records from the subsystems_south table with the following query:

```
select *, AsGML(the_geom) from "subsystems_south" WHERE the_geom IS NOT
NULL LIMIT 10000 OFFSET 10001;
```

Figure 53: SQL query with R-script (see Figure-49)

Similarly, the same query conducted against the "subsystems_central" database was 4.5 times quicker, although the magnitude of the data being so large had to be exported into .dat and .csv files in order for the effective data mining analysis to be carried out, especially since the data that was exported was substantially large. For

example, the subsystems_central.dat file measured 43,285KB in size while subsystems_south.dat measured 4.5 times more, 176,174KB in size.

4.4 Creating raster shapefile data-layers and interface with PostgreSQL database using uDIG

A set of shape files and database connectable dataset stores was established using uDIG tool (see Figure-54, Figure-55, and Figure-56). In doing so, we were able to dissect a bounding box region of our interest and project visual analysis and access to metadata for further Project R data analysis.

The data set was exported as a resource shape file. Other data extraction formats are also possible included image files. The data export will result in the production of prj file, wld files and shp files for each layer file.

The data set was exported as a resource shape file. Other data extraction formats are also possible included image files. Data export will result in the production of a prj file, wld files and shp files for each layer file.

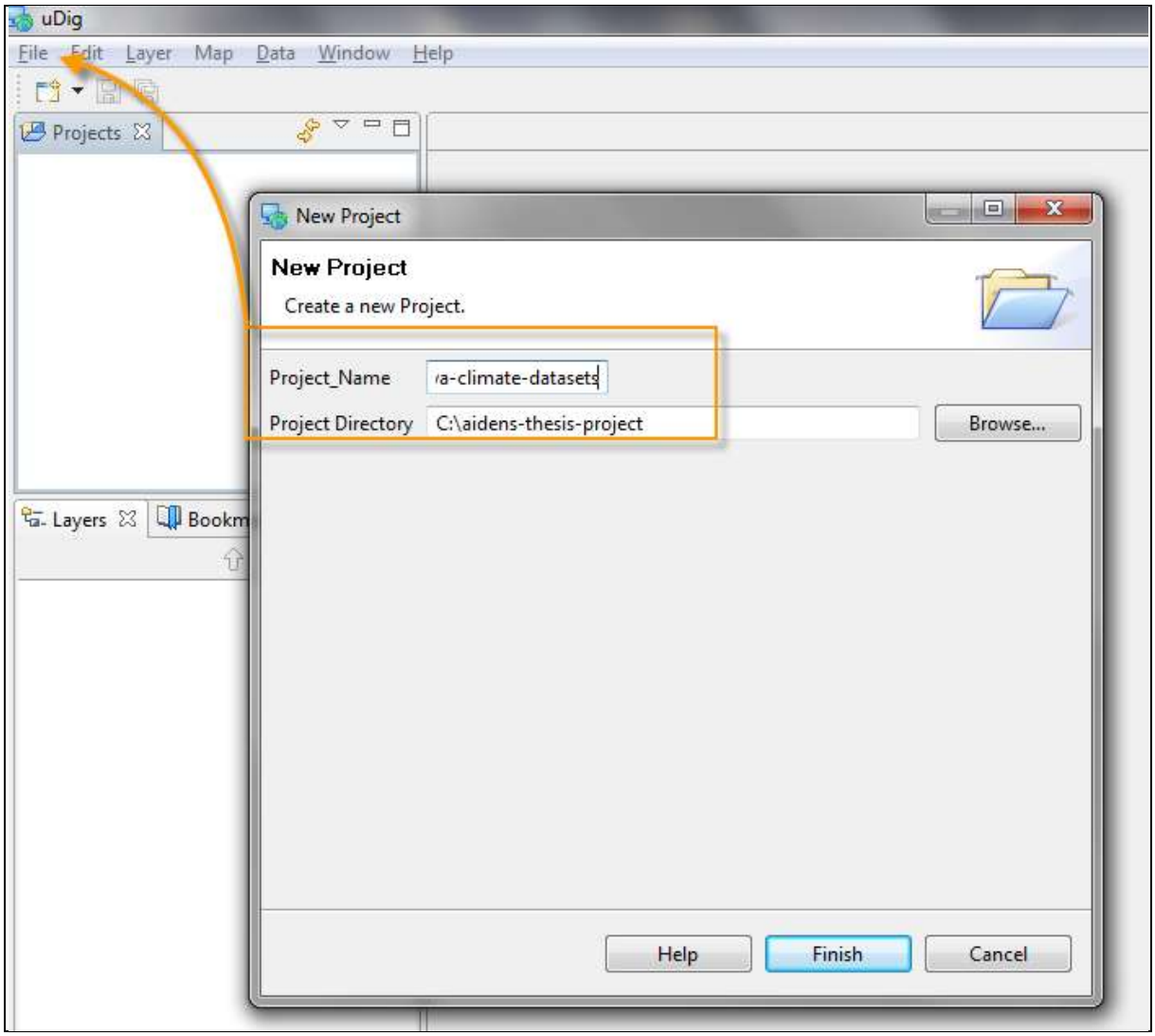


Figure 54: Creating a fresh uDig project for wa_climate_datasets

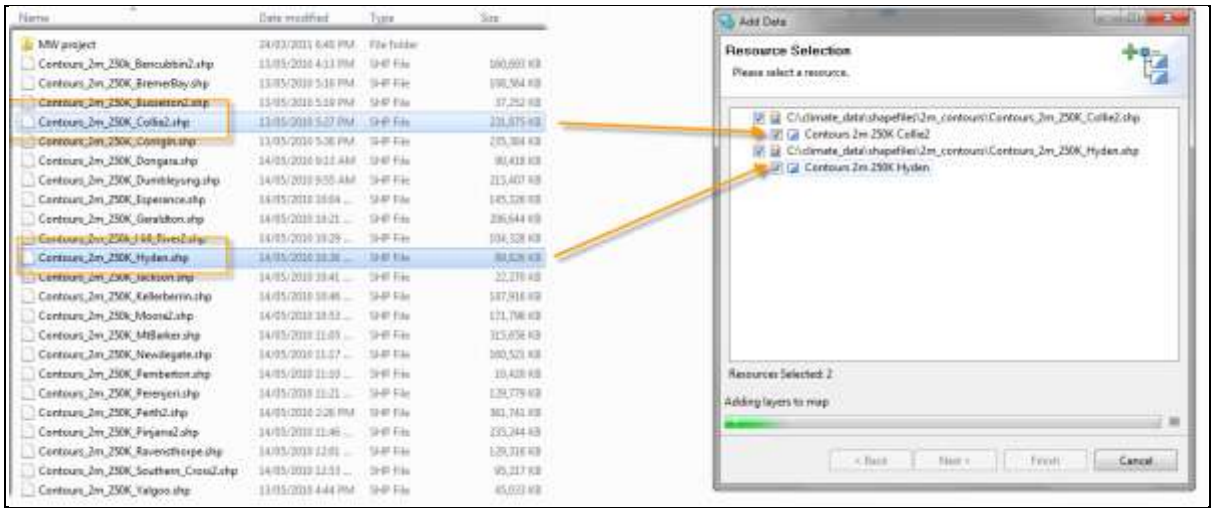


Figure 55: Adding contour shapefile resouces into map layers

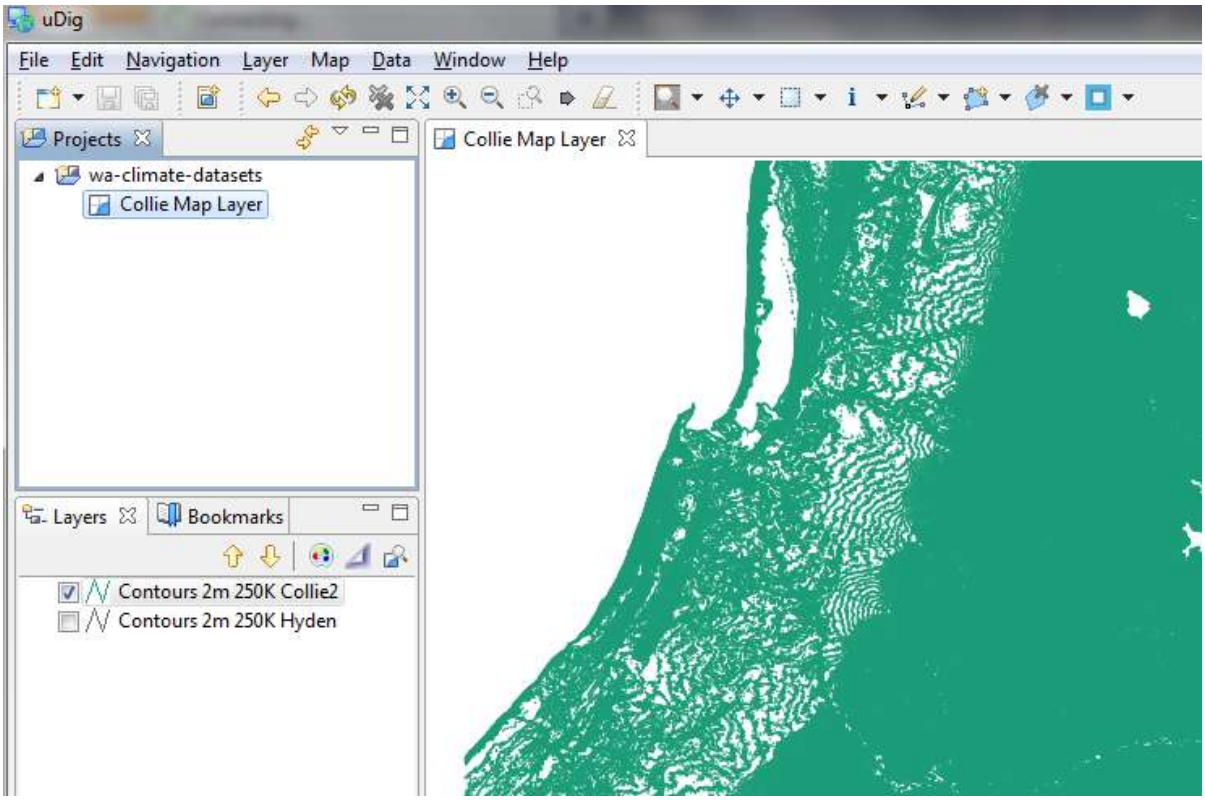


Figure 56: WA uDIG caption

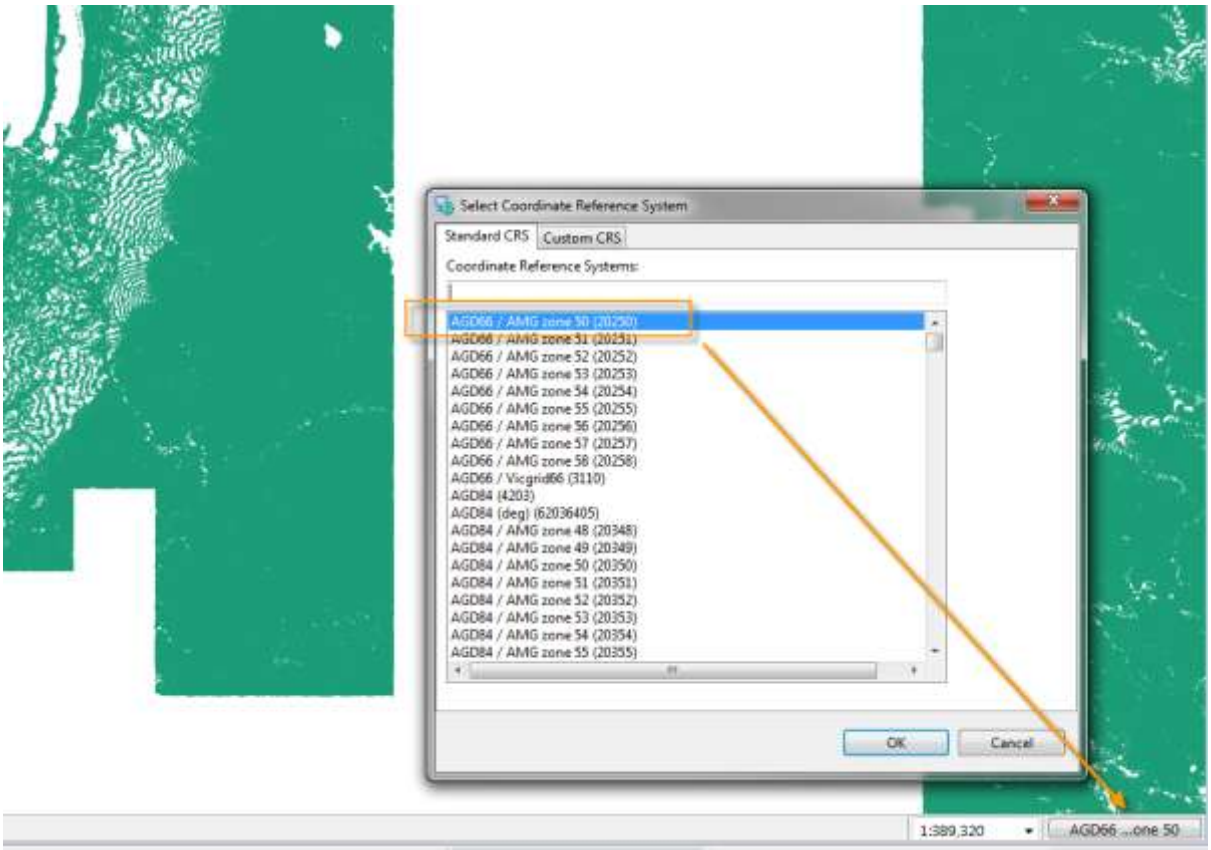


Figure 57: WA Climate Datasets uDig project

In order to correctly establish spatial references on to our spatial maps of Australian nature we need to select the “Zone 50”, also known as AGD66 coordinate system (see Figure-57).

4.4.1 Pre-processing the WA Climate Rainfall data into PostgreSQL

In order to study the climate water catchment datasets we will need to de-associate a batch of four years of longitudinal CSV datasets. To commence the process of visualising the problem of WA climate data and proceeding with the Project R Cluster Analysis, we will need to firstly import all the climate data into PostgreSQL database using automated COPY scripts. These being said, a partition of four different csv dataset folders as illustrated in the Figure-59, represent 12 months of for the the given years, such as 1990, 1980, 2009 and year 2000.

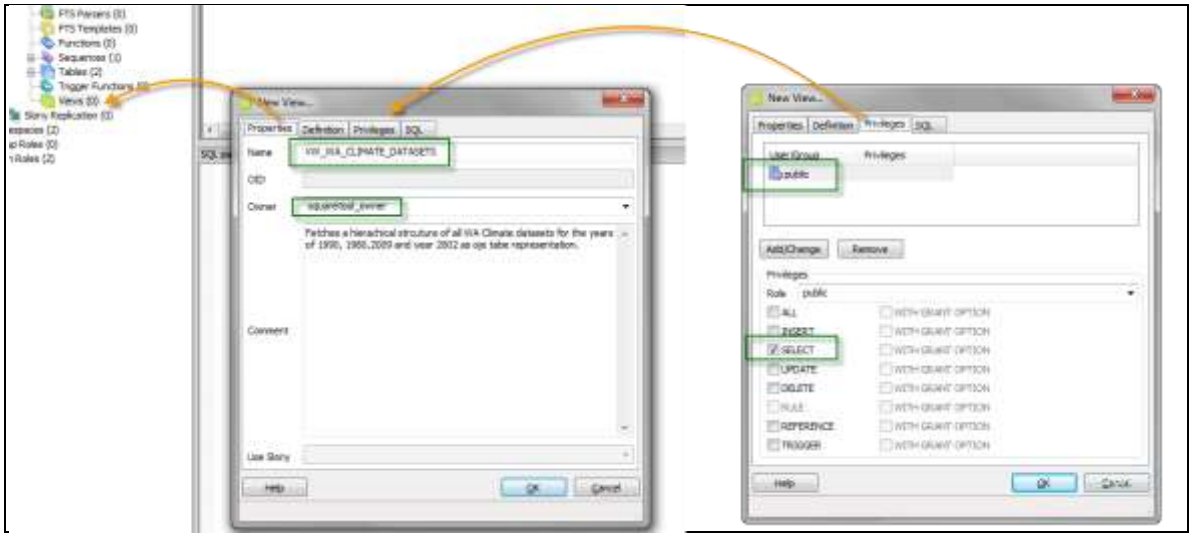


Figure 58: WA Climate Dataset View

It is necessary to create a single database view; in this instance we will call this view “VW_WA_CLIMATE_DATASETS” (see Figure-58). The view in its own form will facilitate this study by providing a tabular representation for all 48 CSV files; in a denormalised manner (see Figure-59)

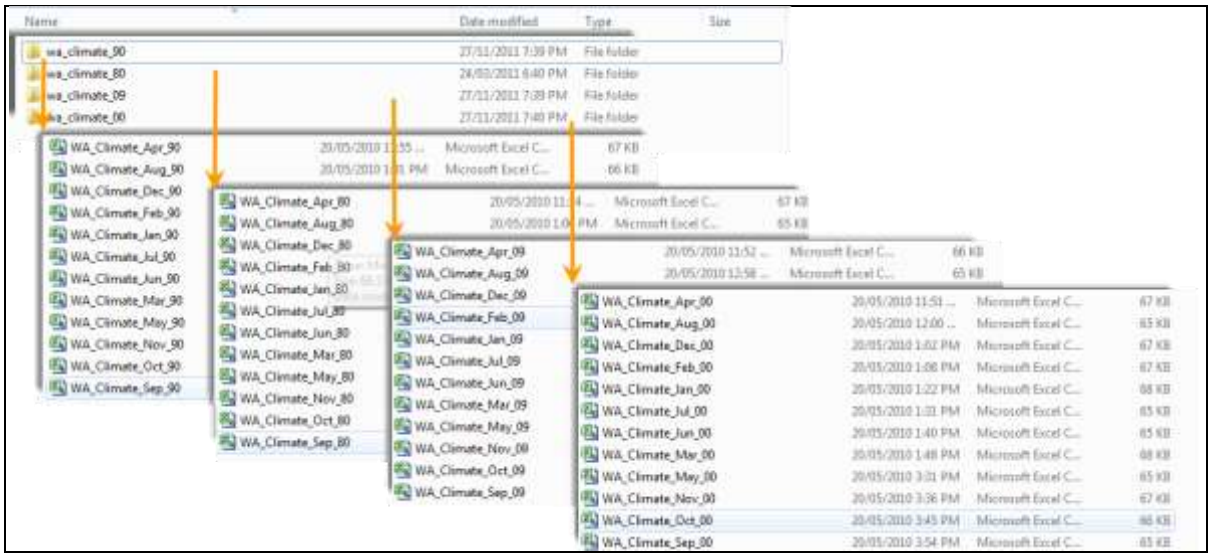


Figure 59: Overview of aggregated CSV files

In the same manner, as running analysis on water catchments we can also augment the same process and apply it to the geospatially referenced WA Climate shape files for a total of four years, grouped into its own year, such as: 00, 90, 80 99. For this purpose, in order to aggregate the data into a singular tabular form we need to create a flat table schema definition in order for the parse CSV files (see Figure-59). The WA_Climate_Data postgresql table holds the same columns as the CSV files, except that we had to create an additional two columns and populate each CSV file. These beings the year_id and month_id. The reason for creating a year_id and month_id and prefixing these columns to the CSV files was so that we could establish a longitudinal and yearly association between various months and years.

4.4.2 Creating a denormilized WA_Climate_Data table for CSV datasets

This script (see Figure-60) creates the WA_Climate_Data table into the postgresql database for us which we than use to perform the analysis in the same manner as demonstrated with the Project R. We also, create a transform a geospatial longitude and latitude columns into a geometrical “the_geom” column type as part of PostGIS which we can use for converting and manipulating data with the GML data format. And finally, we assign a geographical coordinate referenced to an ESRI ID of 4326. Although, we can assign any other reference id, however for this purpose we have used the 4326 reference.

```
-- DROP TABLE wa_climate_data;

CREATE TABLE wa_climate_data
```

```
(
  id serial NOT NULL,
  year_id character varying(50),
  month_id character varying(50),
  site_id character varying(50),
  longitude character varying(50),
  latitude character varying(50),
  monthly_rainfall character varying(50),
  avge_daily_max_temp character varying(50),
  avge_daily_min_temp character varying(50),
  avge_daily_evap character varying(50),
  avge_daily_rad character varying(50),
  the_geom geometry,
  CONSTRAINT enforce_dims_the_geom CHECK (st_ndims(the_geom) = 2),
  CONSTRAINT enforce_geotype_the_geom CHECK (geometrytype(the_geom) = 'POINT'::text OR
the_geom IS NULL),
  CONSTRAINT enforce_srid_the_geom CHECK (st_srid(the_geom) = 4326)
)
WITH (
  OIDS=FALSE
);
ALTER TABLE wa_climate_data
  OWNER TO postgres;
```

Figure 60: SQL script

The import process will be carried over with the PostgreSQL COPY function. The COPY function is an internal database import low-level function which is comprised with various parameters inputs that instructs the PostgreSQL server to directly read from a designed CSV file and parse the data into database repository by creating a designed table name in a performance efficient manner. In our instance, we will instruct the PostgreSQL to read from the CSV file. .

“The file must be accessible to the server and the name must be specified from the viewpoint of the server. When STDIN or STDOUT is specified, data is transmitted via the connection between the client and the server as explained under PostgresSQL documentations (PostgreSQL, 2011).

After creating the table, we create a safe ground executing the postgresql COPY functionality. Similarly, this function is performed in the same manner using shp2psql tool. However, in this instance we demonstrate this process by hand (see Figure-61, Figure-62, Figure-63, Figure-64).

```
-- COPY function for all the months in 80
-- parses all the csv comma deliminated csv meta-data and creates a geospatial referential dataset

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jan_80.csv' delimiters ',';
```

```
copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_feb_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_mar_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_apr_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_may_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jun_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jul_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_aug_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_sep_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_oct_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_nov_80.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_dec_80.csv' delimiters ',';
```

Figure 61: SQL script for years "80"

```
-- copy function for all the months in 90
-- parses all the csv comma deliminated csv meta-data and creates a geospatial referential dataset


copy wa_climate_data
```

```
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jan_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_feb_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_mar_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_apr_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_may_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jun_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jul_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_aug_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_sep_90.csv' delimiters ',';
```

```
copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_oct_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_nov_90.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_dec_90.csv' delimiters ',';
```

Figure 62: SQL script for year "90"

```
-- copy function for all the months in 00
-- parses all the csv comma deliminated csv meta-data and creates a
geospatial referential dataset

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jan_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_feb_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_mar_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_apr_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
```

```
'c:/climate_data/wa_climate_may_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jun_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_jul_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_aug_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_sep_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_oct_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_nov_00.csv' delimiters ',';

copy wa_climate_data
(year_id,month_id,site_id,longitude,latitude,monthly_rainfall,avge_daily_ma
x_temp,avge_daily_min_temp,avge_daily_evap,avge_daily_rad) from
'c:/climate_data/wa_climate_dec_00.csv' delimiters ',';
```

Figure 63: SQL script for year "00"

```
-- COPY function for all the months in 09
-- Parses all the CSV comma deliminated csv meta-data and creates a
geospatial referential dataset

COPY WA_CLIMATE_DATA
```

```
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Jan_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Feb_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Mar_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Apr_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_May_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Jun_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Jul_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Aug_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Sep_09.csv' DELIMITERS ',';
```

```
COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Oct_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Nov_09.csv' DELIMITERS ',';

COPY WA_CLIMATE_DATA
(YEAR_ID,MONTH_ID,SITE_ID,LONGITUDE,LATITUDE,MONTHLY_RAINFALL,AVGE_DAILY_MA
X_TEMP,AVGE_DAILY_MIN_TEMP,AVGE_DAILY_EVAP,AVGE_DAILY_RAD) FROM
'C:/climate_data/WA_Climate_Dec_09.csv' DELIMITERS ',';
```

Figure 64: SQl script for year "09"

After the copy process has been finalised (see Figure-65), we can clearly see that over 40,000 files have been ported across into a single table name. With this in mind, this enables us to run some data mining techniques as illustrates with the Pinjara datasets, except that we no longer read directly from shapefile but instead we parse the data directly into tabular Project R tables and perform data mining techniques. This paper has demonstrates various way of how the pre-processing can be carried out, both using the shapefile and csv file and augmenting the geospatial longitude and latitude columns into geometrical polygon areas for later visual representation.

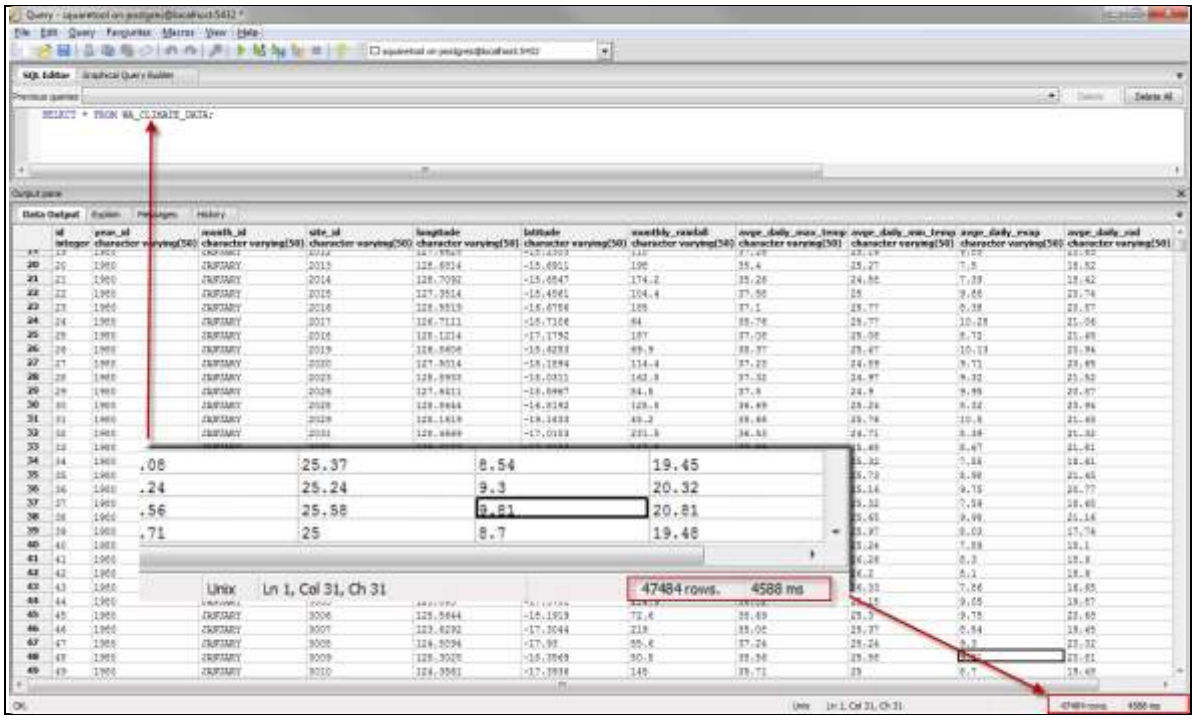


Figure 65: Aggregated table query

5 DISCUSSION AND CONCLUSIONS

There is a need to find better approaches to predict possible land use changes in the South Western Australia agricultural areas. The increasing degradation of agricultural lands from soil salinity, waterlogging, nutrient runoff and eutrophication could have devastating consequences for future food production in Western Australia. The use of data mining provides a means to interrogate the geospatial data sets of land use and soils in this region. A number of data mining techniques could be used to achieve this interrogation. This research has demonstrated the techniques that could be used to preprocess and analyze the data sets. The research has used opensource software tools to demonstrate the process of importing processing and displaying spatial datasets. The study has focused on the Peel Harvey region of Western Australia which is a representative region of the agricultural production areas of South West of Western Australia.

This study has also outlined the design of a proof of concept component based software tool. The techniques described in this paper can be used to integrate into the data mining context of the software tool. It is proposed that this software tool will be used by stakeholders, such as land planners and agricultural scientists to interrogate individual catchment areas or regional areas for land usage. The software tool may provide a means to work through climate and land use scenarios to make predictions of land use with changes in climate and other agricultural factors.

Though, several limitations have been experienced during this project, these being loading of shapefile data larger than 30,000 megabytes renders it impossible on the current desktop system specifications to process data without some form of high-cluster networking infrastructure. As a result the project experienced the need of high-powered computing to facilitate in processing geographically spatial datasets which can proceed in gigabytes worth of geographical data once exported into postgreSQL database. Likewise, the Project R is confined only to limited amount of desktop power using 32-bit infrastructures due to current computer specification. Although, this project has the potential to expand into a vast framework software platform that requires much more dedicated and reasonable timeframes in order to establish effective data mining techniques on spatial data.

6 APPENDICES

6.1 APPENDIX A. WEBSITE LINKS OF OPEN SOURCE TOOLS EMPLOYED

Name	Source
Eclipse, Jee Galileo, version SR2	http://www.eclipse.org/downloads/
Project R, version 2.1.1.0	http://cran.ms.unimelb.edu.au/bin/windows/
PostgresSQL, version 8.2	http://www.postgresql.org/download/
PostGIS, version 8.2	http://postgis.refractions.net/download/windows/
uDIG, version 1.2	http://udig.refractions.net/download/
jGRASS, version 2.0.20060730	http://sourceforge.net/projects/jgrass/
Processing, version 1.0.9	http://processing.org/download/processing-1.1-expert.zip

7 REFERENCES

Andrienko, G., Malerba, D., May, M., & Teisseire, M. (2006). Mining spatio-temporal data. Journal of Intelligent Information Systems, 27(3), 187-190. doi:10.1007/s10844-006-9949-3

Answers. (2010a). answers.com search. [hydrogeology]. [online]. Available WWW: http://www.answers.com/topic/hydrogeology [May 6 2010].

Answers. (2010b). answers.com search. [method]. [online]. Available WWW: http://www.answers.com/topic/ method [May 6 2010].

Answers. (2010c). answers.com search. [model]. [online].Available WWW: http://www.answers.com/topic/ model [May 6 2010].

Answers. (2010d). answers.com search. [precision agriculture]. [online]. Available WWW: http://www.answers.com/topic/ precisionagriculture [May 6 2010].

Armenakis, C. (2008). Spatial data infrastructures and clearinghouses. Remote Sensing and Spatial Information Sciences: 2008 Isprs Congress Book

Australian National Resource Australian [ANRA]. (2000). Dryland salinity assessment 2000 Western Australia. Retrieved February 16, 2010 from http://www.anra.gov.au/topics/salinity/pubs/national/salinity_wa.html

Australian National Resource Australian [ANRAa]. (2000) Kamarooka Case Study Catchment, Victoria. Retrieved 22nd February, 2010 from http://www.anra.gov.au/topics/salinity/pubs/national/salinity_case_kamarooka.html

Australian National Resource Australian [ANRAb]. (2000). Waterway and estuarine Australian Natural Resources Atlas

Basili, V. (1992). The experimental paradigm in software engineering. Experimental Software Engineering Issues: Critical Assessment and Future Directions, 1–12.

Bates, D. M. (1999). Programming with Data: A Guide to the S Language. Technometrics, 41(3), 266.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures.

Bowyer J, Marie S. (2007). Natural Resource Management intermediaries as potential next users. Retrieved 25th April, 2010 from http://www.esri.com/library/brochures/pdfs/open-gis-platform.pdf

BusinessDictionary (2010a). businessdictionary.com search. [data collection].[on-line]. Available WWW: http://www.businessdictionary.com/definition/data-collection.html. [May 06 2010].

BusinessDictionary (2010b). businessdictionary.com search. [data mining]. [on-line]. Available WWW: http://www.businessdictionary.com/definition/data-mining.html. [May 06 2010].

BusinessDictionary (2010c). businessdictionary.com search. [machine learning]. [on-line]. Available WWW: http://www.businessdictionary.com/definition/machine-learning.html [May 06 2010].

BusinessDictionary (2010d). businessdictionary.com search. [methodology]. [on-line]. Available WWW: http://www.businessdictionary.com/definition/methodology.html [May 06 2010].

BusinessDictionary (2010e). businessdictionary.com search. [topology]. [on-line]. Available WWW: http://www.businessdictionary.com/definition/topology.html [May 06 2010].

Chaudhuri, S., & Dayal, U. (1997). Data warehousing and OLAP for decision support. *ACM Sigmod Record*, *26*(2), 507–508.

H. Kargupta, A. Joshi, K. Sivakumar and Y. Yesha. (2006). Data Mining-Next Generation Challenges and Future Directions. Biometrics, 62(1), 312-312. doi:10.1111/j.1541-0420.2006.00540_16.x

Department of Environment and Conversations [DEC], (2003) guiding document with strategies for establishing a monitoring network capable of accurately measuring nutrient loads. Retrieved April 16, 2010 from http://www.epa.wa.gov.au/docs/WQIP/AppendixD.pdf

Deren, L. I., Kaichang, D. I., & Deyi, L. I. (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, 238–245.

Dees, T. (2002). Understanding GIS. *Law & Order*, *50*(8), 42.

Dictionary(2010a). dictionary.reference.com search. [geospatial]. [online]. Available WWW: http://dictionary.reference.com/browse/geospatial [May 6 2010].

Ding, Q., Ding, Q., & Perrizo, W. (2008). PARM--an efficient algorithm to mine association rules from spatial data. IEEE Transactions on Systems, Man, and Cybernetics--Part B: Cybernetics, 38(6), 1513.

Dunstan, N., & Qiang, X. (2007).Finding Solutions with Land Cover and Change Models, Proceedings of the IASK International Conference on E-Activity and Leading Technologies, Oporto, December 2007.

Dunstan N., Armstrong L. and Diepeveen D. (2009). Selecting Areas for Land Use Change in a Catchment. Proceedings of the 4th India International conference on Artificial Intelligence, 16-18 December 2009. Tumkur India.

Ekasingh, B., Ngamsomsuke, K., Letcher, R., & Spate, J. (2005). A data mining approach to simulating farmers' crop choices for integrated water resources management. Journal of Environmental Management, 77(4), 315-325. doi:10.1016/j.jenvman.2005.06.015

Environmental Systems Research Institute. [ESRI]. (2003) The Open GIS Platform. Retrieved April 02, 2010 from http://www.esri.com/library/brochures/pdfs/open-gis-platform.pdf

Environmental Systems Research Institute. [ESRIa]. (1998) ESRI Shapefile Technical Description. Retrieved April 03, 2010 from http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Farmer, D., Cattlin, T., Stanton, D., & Coles, N. (2004). 2004. A revised criteria for runoff management in south-west western Australia. Proc. Of 13th Int. Soil Con. Org. Conf.: Conserving Soil and Water for Society: Sharing Solutions ISCO 2004 Bris. July 2004

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & others. (1996a), A, From data mining to knowledge discovery in databases. Communications of the ACM, 39(11), 24–26.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & others. (1996b). B, Knowledge discovery and data mining: Towards a unifying framework. In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR (pp. 82–88).

Foody, G. M. (2003). Uncertainty, knowledge discovery and data mining in GIS. Progress in Physical Geography, 27(1), 113.

Forbus, K. D., Usher, J., & Chapman, V. (2004). Qualitative Spatial Reasoning about Sketch Maps. AI Magazine, 25(3), 61.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479.

Gahegan, D. G. (2006). Spatial ordering and encoding for geographic data mining and visualization. Journal of Intelligent Information Systems, 27(3), 243.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In Proceedings of the New Zealand computer science research students conference (pp. 57–64).

Gould, M., & Hecht Jr, L. (2001). OGC: A Framework for Geospatial and Statistical Information Integration. Joint UNECE/Eurostat Work Session on Methodological Issues. Tallinn, Estonia. Reichardt, M. (2004, October 12). The havoc of non-interoperability.

Han, J., & Kamber, M. (2001). Data mining: concept and techniques. Morgan Kaufmann.

Han, J., & Kamber, M. (2006). Data mining : concepts and techniques. Morgan Kaufmann.

Han, J., & Kamber, M. (2006). Data mining: concepts and techniques. Morgan Kaufmann.

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT Press.

Hernández, D. (1994). Qualitative representation of spatial knowledge. Springer.

Howe, D. (2010a). The Free On-line Dictionary of Computing. [algorithm]. [on-line]. Available WWW: http://foldoc.org/algorithm. [May 06 2010].

Howe, D. (2010b). The Free On-line Dictionary of Computing. [architecture]. [on-line]. Available WWW: http://foldoc.org/architecture. [May 06 2010].

Howe, D. (2010c). The Free On-line Dictionary of Computing. [attribute]. [on-line]. Available WWW: http://foldoc.org/attribute. [May 06 2010].

Hussein A. Abbass, Ruhul A. Sarker, and Charles Newton, Data mining: a heauristic approach (Idea Group Inc (IGI), 2002).

Jacquez, G. M. (n.d.). 22 Spatial Cluster Analysis. Blackwell Publishing, pages 395-416.

Kamil, S., & Yalçın, Y. (2001) Free GIS software and their usage in the river basin management, 244-257. Retrieved April 05, 2010 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.3635&rep=rep1&type=pdf

Keim, D. A. (2002). Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1), 1-8.doi:10.1109/2945.981847

Khosla, R., & Dillon, T. S. (1997). Engineering Intelligent Hybrid Multi-agent Systems. Springer. Retrieved April 01, 2010 from http://books.google.com/books?id=iT98_yh_bBgC&printsec=frontcover&dq=Engineering+Intelligent+Hybrid+Multi-agent+Systems&source=bl&ots=yK3LV60-38&sig=7ZM0cDJFkek2xQ7otbx1Exr5VbE&hl=en&ei=KVYCTO31DszXcZ-l3NUB&sa=X&oi=book_result&ct=result&resnum=1&ved=0CBsQ6AEwAA

Kiran, M., Murali, K., & Venugopal, R. (n.d.). Data Mining Of Geospatial Database For Agriculture Related Application. Retrieved Feb 25, 2010, from http://www.gisdevelopment.net/proceedings/mapindia/2006/agriculture/mi06agri_124.htm

Kiran, M., Murali, K., & Venugopal, R. (n.d.). Data Mining Of Geospatial Database For Agriculture Related Application. Retrieved Feb 25,, 2010, from http://www.gisdevelopment.net/proceedings/mapindia/2006/agriculture/mi06agri_124.htm

Kitamoto, A. (2002). Spatio-Temporal Data Mining for Typhoon Image Collection. Journal of Intelligent Information Systems, 19(1), 25.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In Advances in spatial databases (pp. 47–66).

Kowalski, S. (2000). Some misconceptions about data mining. Catalog Age, 17(4), 97.

Kravchenko, A. N., & Bullock, D. G. (2002). Spatial variability of soybean quality data as a function of field topography: I. Spatial data analysis. Crop Science, 42(3), 804.

Cios, K. J., Pedrycz, W., Świniarski, R., & Swiniarski, R. (1998). *Data mining methods for knowledge discovery*. Springer.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & others. (1996). From data mining to knowledge discovery in databases. *Communications of the ACM*, *39*(11), 24–26.

Landgate (2008), Geospatial Data Dictionary; Western Australian Land Information Authority. Available Online https://www2.landgate.wa.gov.au/slip/portal/services/files/GeospatialDataDictionary.pdf

Lee, C., & Percivall, G. (2008). Standards-Based Computing Capabilities for Distributed Geospatial Applications. Computer, 41(11), 50-57.

Norton, M. J. (1999). Knowledge discovery in databases. *Library Trends*, *48*(1), 9–21.

M. Yuan et al., “Geospatial data mining and knowledge discovery,” A research agenda for geographic information science (2004): 365–388.

Maimon, O. Z., & Rokach, L. (2005). Data mining and knowledge discovery handbook. Springer Science & Business.

Masumoto, S., Raghavan, V., Yonezawa, G., Nemoto, T., & Shiono, K. (2004). Construction and visualization of a three dimensional geologic model using GRASS GIS. Transactions in GIS, 8(2), 211–223.

Tjoa, A. M., & Trujillo, J. (2006). *Data warehousing and knowledge discovery*. Springer. Retrieved March 13, 2010 from http://books.google.com.au/books?id=2MXz3RXWV5cC

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). *Data mining in agriculture*. Springer Verlag. Retrieved March 02, 2010 from http://books.google.com.au/books?id=I85EsrHmdM4C

Neteler, M., & Mitasova, H. (2008). *Open source GIS: a GRASS GIS approach*. Springer.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In Proceedings of the International Conference on Very Large Data Bases (pp. 144–144).

Open GIS Consortium [OGC]. (2003). Data Models and Interoperability. Retrieved 24th February, 2010 from http://portal.opengeospatial.org/files/?artifact_id=3805

Ordance Survey. (2008). A guide to coordinate system in Great Britain, 1-46 Retrieved April 04, 2010 from http://www.ordnancesurvey.co.uk/gps/docs/A_Guide_to_Coordinate_Systems_in_Great_Britain.pdf

Project R [R]. (2009a). Retrieved April 04, 2010 from http://bm2.genes.nig.ac.jp/RGM2/R_current/library/cluster/man/clusplot.default.html

Project R [R]. (2009b). Retrieved April 04, 2010 from http://finzi.psych.upenn.edu/R/library/mclust/html/Mclust.html

Pujari, A. K. (2001). *Data Mining Techniques*. Orient Blackswan.

Renz, J. (2002). Qualitative spatial reasoning with topological information.

Ripley, B. D. (2001). The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network, 1(1), 23–25.

Sehovic, A. , Armstrong L.J. , Diepeveen D.A. (2010) Interrogation of water catchment data sets using data mining techniques. in Editors: L.J. Armstrong and J. Clayden "Proceedings of the Knowledge Discovery for Rural Systems Workshop 2010" at the 14th Pacific-Asia

Conference on Knowledge Discovery and Data Mining., IIIT Hyderabad, Hyderabad India , 21-24 June, 2010 - Hyderabad, India.

Santos, M. Y., & Amaral, L. A. (2004). spatial data mining in the analysis of a demographic database. Soft Computing, 9(5), 374-384. doi:10.1007/s00500-004-0417-0 Shared Land Information Platform [SLIP]. (2010a). Topography Data Dictionary Hydrography. Retrieved April 28, 2010 from https://www2.landgate.wa.gov.au/slip/servlet/portal/serve/Library/SLIP/services/files/Topo-Data-Dictionary-hydrography.pdf

Sehovic, A. Armstrong L.J. , Diepeveen D.A. (2010) Interrogation of water catchment data sets using data mining techniques. in Editors: L.J. Armstrong and J. Clayden "Proceedings of the Knowledge Discovery for Rural Systems Workshop 2010" at the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining., IIIT Hyderabad, Hyderabad India , 21-24 June, 2010 - Hyderabad, India.

Shared Land Information Platform [SLIP]. (2010b). Topography Data Dictionary Ground Surface. Retrieved April 28, 2010 from https://www2.landgate.wa.gov.au/slip/servlet/portal/serve/Library/SLIP/services/files/Topo-Data-Dictionary-groundsurface.pdf

Shared Land Information Platform [SLIP]. (2010c). Subscription Datasets. Retrieved April 28, 2010 from https://www2.landgate.wa.gov.au/slip/servlet/portal/serve/Library/SLIP/services/datasets_subscription.html

Shared Land Information Platform [SLIP]. (2010d). Geological Map of WA (Polygon)-2.5M (DMP-011) in GML format. Retrieved April 25, 2010 from https://www2.landgate.wa.gov.au/datadownloads/DMP011/DMP_011_4283_20090325145626_GML.zip

Shared Land Information Platform [SLIP]. (2010e). Geological Map of WA (Polygon)-2.5M (DMP-011) in SHP format. Retrieved April 25, 2010 from https://www2.landgate.wa.gov.au/datadownloads/DMP011/DMP_011_4283_20090325145626_SHP.zip

Shared Land Information Platform [SLIP]. (2010f). Dola Topographic Series 1:25 000. Retrieved April 25, 2010 from https://www2.landgate.wa.gov.au/datadownloads/DMP-011/CI_DLI_topographic+25000_200304.pdf

SharpMap (2008). Geospatial Application Framework for the CLR Available at: http://www.codeplex.com/SharpMap.

Shneiderman, B. (2002). Inventing discovery tools: combining information visualization with data mining. Information Visualization, 1(1), 5.

SLIP (2008). Shared Land Information Platform, Data Consumer Handbook Soman, K. P., Diwakar, S., & Ajay, V. (2006). Insight into Data Mining: Theory and Practice. PHI Learning Pvt. Ltd.

Sun, C., Bandi, N., Agrawal, D., & El Abbadi, A. (2006). Exploring spatial datasets with histograms. Distributed and Parallel Databases, 20(1), 57-88.The value of standards: a delphi study. (2003, June). . Delphi group.

Tjoa, A. M., & Trujillo, J. (2006). *Data warehousing and knowledge discovery*. Springer. Retrieved March 13, 2010 from http://books.google.com.au/books?id=2MXz3RXWV5cC

Travinin Bliss, N., & Kepner, J. (2007). pMatlab Parallel Matlab Library. The International Journal of High Performance Computing Applications, 21(3), 336-359. doi:10.1177/1094342007078446

PostgreSQL, 2011). http://www.postgresql.org/docs/8.1/static/sql-copy.html

Ultsch, A., & Mörchen, F. (2005). ESOM-Maps: tools for clustering, visualization and classification with Emergent SOM.

Ecohydrology. (2010).University Of Western Australia, Catchment water management planning. Retrieved April 12, 2010 from http://www.ecohydrology.uwa.edu.au/research/cwmp.

Wachowicz, M. (2002). Uncovering Spatio-Temporal Patterns in Environmental Data. Water Resources Management, 16(6), 469-487. doi:10.1023/A:1022259531710

W. H Inmon, “What is a Data Warehouse?,” Prism Tech Topic 1, no. 1 (1995).

W. H Inmon, W. H. (1996). “The data warehouse and data mining,” Communications of the ACM 39, no. 11 (1996): 49–50.

Webopedia(2010a). wbopedia.com search. [gis]. [online]. Available WWW: http://www.webopedia.com/TERM/G/GIS.html [May 6 2010].

Webopedia(2010b). wbopedia.com search. [ide]. [online]. Available WWW:http://www.webopedia.com/TERM/I/integrated_development_environment.html [May 6 2010].

Weisstein, E. W. (2003). CRC concise encyclopedia of mathematics. CRC Press

Wilcox, J. (2001). A precision Ag plan. Successful Farming, 99(11), S1.

Witten, I. H., & Frank, E. (2005). Data mining: practical machine learning tools and technique. Morgan Kaufmann.

Woolf, A., Cramer, R., Gutierrez, M., van Dam, K. K., Kondapalli, S., Latham, S., Lawrence, B., et al. (2005). Standards-based data interoperability in the climate sciences. Meteorological Applications, 12(01), 9-22. doi:10.1017/S1350482705001556

Wong, W. (2003). Integrated development environments: The ties that bind. *Electronic Design*, *51*(18), 51.

Zeilhofer, P., Lima, E. B. N. R., & Lima, G. A. R. (2006). Spatial Patterns of Water Quality in the Cuiabá River Basin, Central Brazil. Environmental Monitoring and Assessment, 123(1-3), 41-62. doi:10.1007/s10661-005-9114-4

Zhang, X., Shi, L., Jia, X., Seielstad, G., & Helgason, C. (2009).Zone mapping application for precision-farming: a decision support tool for variable rate application. Precision Agriculture, 11(2), 103-114. doi:10.1007/s11119-009-9130-4

Ziarko, W., & Yao, Y. (2001). Rough sets and current trends in computing. Springer. Retrieved March 21, 2010 from http://books.google.com.au/books?id=YTYQaGnEdxIC