Proceedings of the Knowledge Discovery for Rural Systems 2010, Hyderabad, India

Interrogation of water catchment data sets using data mining techniques

A. Sehovic[1], L. J. Armstrong [1], D.A. Diepeveen[2],
[1] School of Computer and Security Science, Edith Cowan University,
Mt Lawley, 6050 Western Australia
Australia
[2] Department of Agriculture and Food, Western Australia,
2 Baron-Hay Court, South Perth 6048,
Western Australia, Australia
a.sehovic@ecu.edu.au, l.armstrong@ecu.edu.au, ddiepeveen@agric.wa.gov.au

Abstract. Current environmental challenges such as increasing dry land salinity, water logging, eutrophication and high nutrient runoff in south western regions of Western Australia (WA) may have both cultural and environmental implications in the near future. Advances in computing through the application of data mining and geographic information services provide the tools to conduct studies that can indicate possible changes in these water catchment areas of WA.

The research examines the existing spatial data mining techniques that can be used to interpret trends in WA water catchment land use. Large GIS data sets of the water catchments on Peel-Harvey region have been collected by the Western Australian government. This paper describes the techniques that will be used to explore the large GIS data sets and provides cluster analysis of a sample subset of the data set as a proof of concept. This research will contribute to the later development of a data mining interrogation tool that measures and validates the effectiveness of different data mining techniques such as: classical statistical methods, cluster analysis and principal component analysis on the sample water catchment data set. The interrogation tool will incorporate some of the geospatial data mining techniques described in this paper to discover meaningful and useful patterns specific to current agricultural problem domain of dry land salinity. This research will contribute towards an understanding of the data mining techniques that can be used in the tool. The tool is expected to be used by government agencies, such as Department of Agriculture and Food, Western Australia researchers and other agricultural industry stakeholders.

Keywords: Data mining, water catchments, geospatial data sets

1 Introduction

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 [1]. This research addresses the current industry problem of dryland salinity in WA. [2] emphasise on current impacts of dry land salinity as being 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 in particular, the effects that pose threats to land use and water catchments of Australia.

Previous studies have been reported on the use of data mining to elucidate water catchment land use patterns. For this purpose, the research by [2] 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 [2]. 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). 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 [2] .

Another study conducted by [3] 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 [3], 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 [3] can simulate effective agricultural socio economic land decisions.

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 [4] Due to the vast land size and intensive agricultural practice, the region has a number of environmental sustainability problems, including increasing salinity, eutrophication water logging, soil acidity and loss of biodiversity [4]. To illustrate, Fig. 1 shows the salinity discharge of the Peel Havey catchment area.

A national land and water resources audit assessment[1] 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 [1] 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, [5].

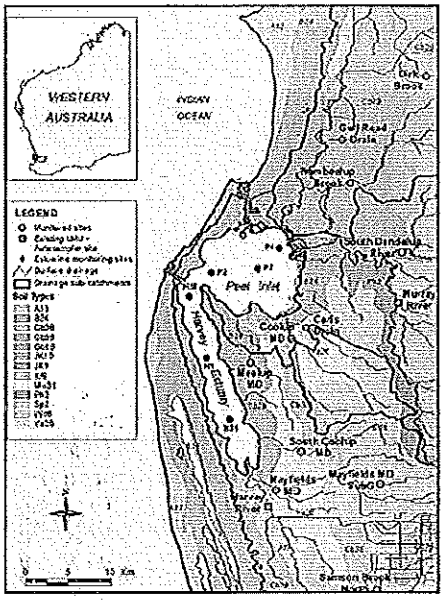

Fig. 6. Peel inlet salinity discharge. [6].

Spatial data mining uses geographic information. To create knowledge 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 [7]. 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. Spatial data mining techniques can comprise various data mining techniques, generally speaking, these may include, classification, associations, clustering and principal component analysis. These techniques can be used for analysis of 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.

There are various classification methods which can be applied to spatial data mining; according to 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 [8]

"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. Studies have shown how association rules can be used for spatial data mining. For example. [9] have 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.

Spatial cluster analysis is also an important data mining technique with an essential role in quantifying geographic variation patterns [10]. [10] 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".

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 [11]. According to [12] spatio temporal data mining is an emerging research area which dedicates development of novel data mining applications. Similarly, it was proven 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 [13].

This paper aims to describe the techniques that could be used to investigate and determine the most feasible spatial data mining techniques for conducting an analysis of water catchment data sets through the interrogation of a subset data set. The results extracted during the analysis stage could be used to assess whether any significant spatial patterns are present and allow predictions to be made in relation to potential changes in climate and land use. The paper also outlines the design of a proposed interrogative data mining tool that could be used to determine most appropriate data mining techniques for the Peel Harvey water catchment area.

2 Research Tools and Methods

The software tools used in this study include computer development tools, spatial database management tools , graphic GIS tools and Visual Simulation tools. These tools are outlined in Table 1.

Table 4. Listing of software tools used in data preparation and analysis.

Computer development	Description
Eclipse, Jee Galileo, version SR2:	Eclipse is an integrated development environment comprising various tools for java developers to create enterprise and online applications.
Project R,	Project R will serve as a primary statistical and data mining

version 2.1.1.0:	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.
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.
Spatial database	
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.
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.
Graphical GIS	Both 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.
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.
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.
Visual simulation	
Processing expert, version 1.0.9	open source programming language and environment that allows individuals to developer artistic images, construct computer simulated animations and allow user interactions.

3. Proposed Data mining Software Tool

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 mining, data set, visual and data mining and a visual data mining component.

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

Visual Context: 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 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 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 and data mining: 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 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.

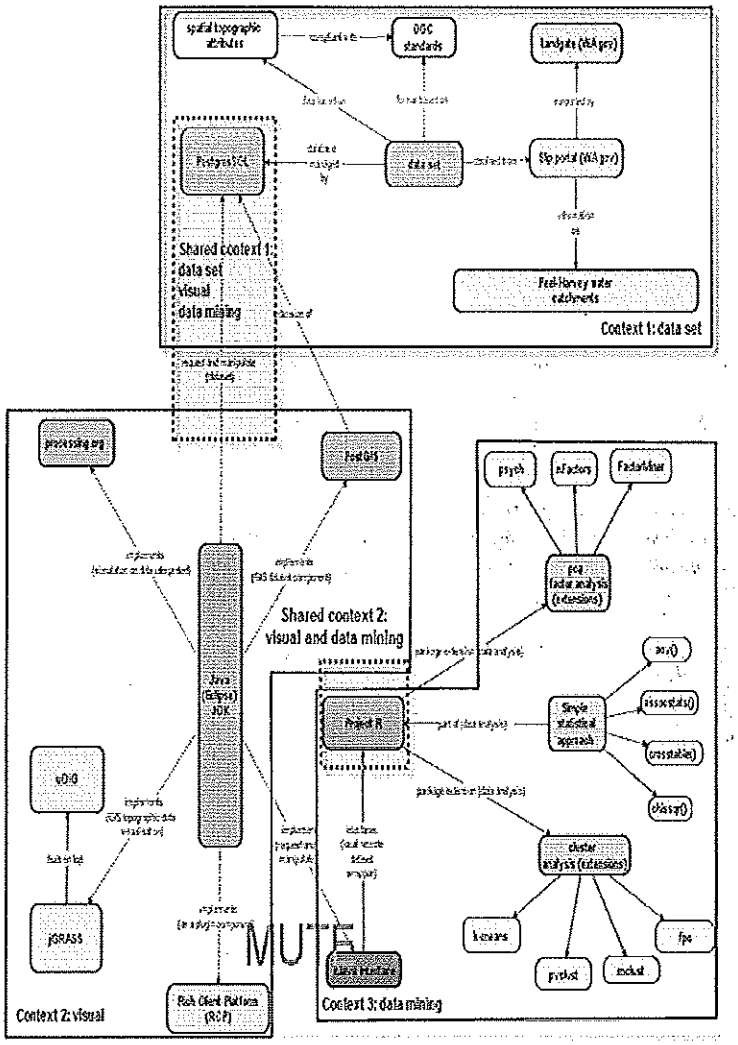

Fig. 2. Proposed data mining software tool design.

4. Case Study Activities

The interrogation of the Peel Harvey catchment data was undertaken to demonstrate processed involved to prepare, preprocess and perform some characteristic cluster analyses on the sample geospatial data sets. 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 file with 2:250 000 resolution. The file was 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. While the Pinjara2 layer is placed on top of the parent layer and represented in the yellow colour. This is displayed in Fig. 3. Shape file are imported into the uDig software.

4.1 Selecting Specific Regions

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 select a region on the map. Alternatively, using the border region selection function from a toolbar section, we can select a boundary (see Fig. 4a, b). 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.

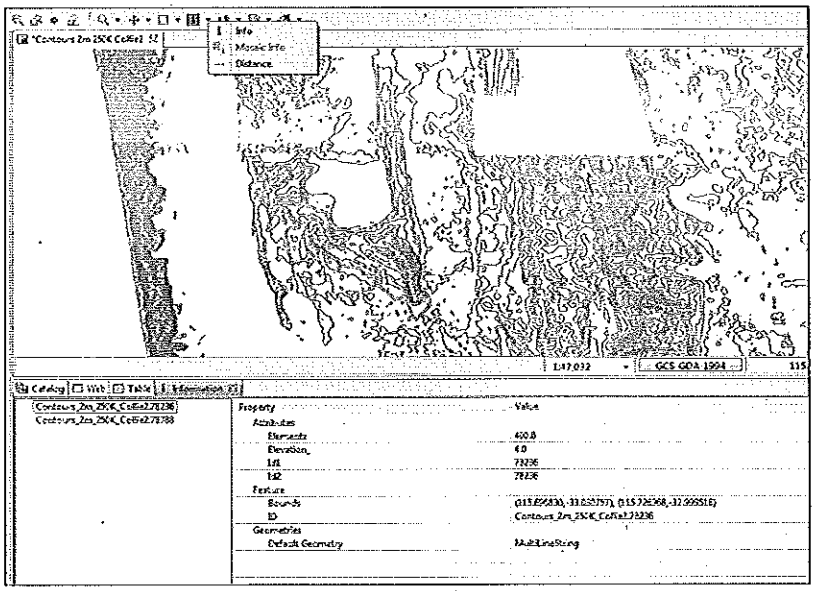

Fig. 3. Map information identifier feature using uDig software.

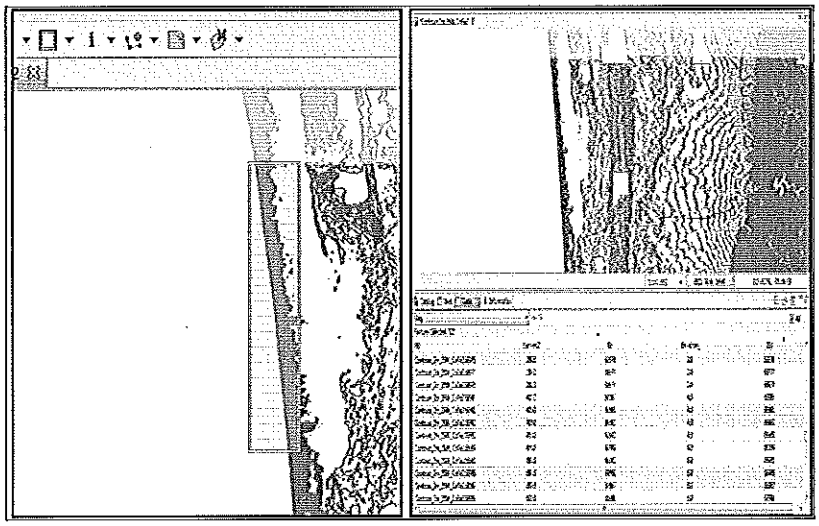

Fig. 4a Map selection tool bar feature using uDig software. b) Automated data selection for region selection.

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 prj file, wld files and shp files for each layer file.

4.2 Analysis of Water Catchment Data.

To demonstrate the possible data analysis that can be used to interrogate the water catchment data, a series of cluster analyses were carried out using 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:-

Step 1: Read the shape file and assign the collie dataset to an object for further manipulation. 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("C:/Users/Setsuna/uDig/SelectedRegions/Contours_2m_250K_Collie2")
```

Step 2: 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.

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

Step 3: 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.

```
# 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'))
```

Using the hclust and stat package, a hierarchical agglomerative graph was created using the Euclidean distance matrix representation.

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

MClustering Example

A model based clustering was created using the mclust R package library as described following.

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

As a result of this, the following four diagrams are displayed, **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.

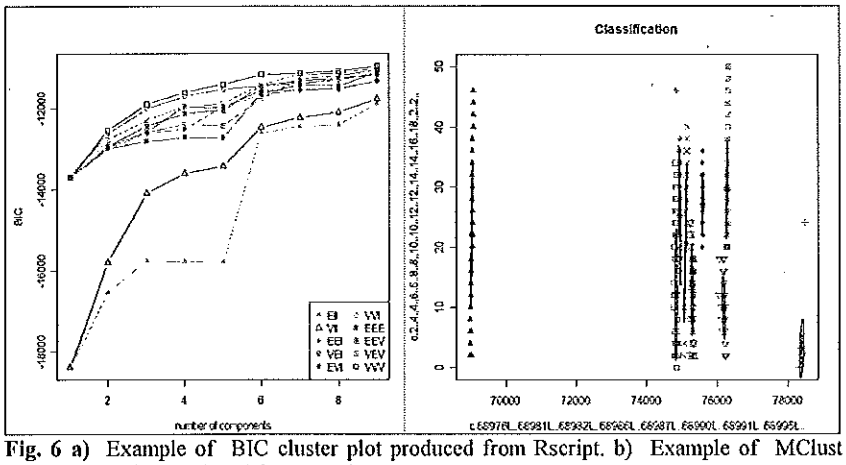

Fig. 6 a) Example of BIC cluster plot produced from Rscript. b) Example of MClust cluster classification produced from Rscript.

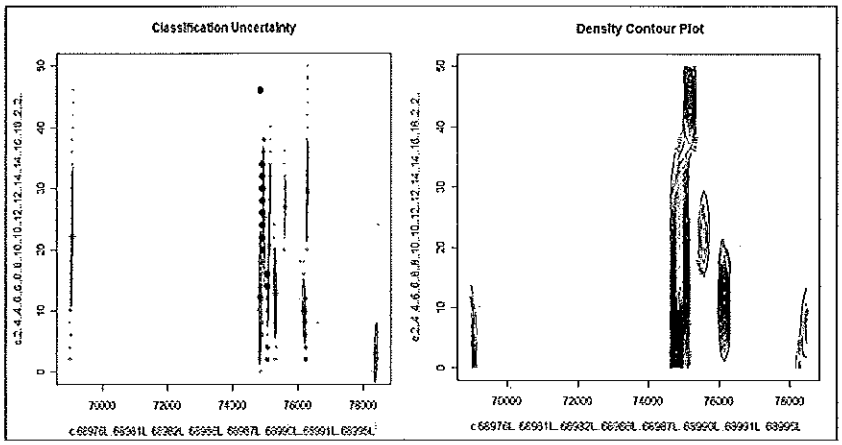

Fig. 7 a) Example of MClust cluster uncertainty plot produced from Rscript. b) Example of MClust density contour plot produced from Rscript.

Using the cluster and fpc package, a k-means cluster with 5 clusters from a set of existing collie data source data was created. 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)
```

A centroid cluster plot against the first and second discriminatory functions was created by using the following Rscript.

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

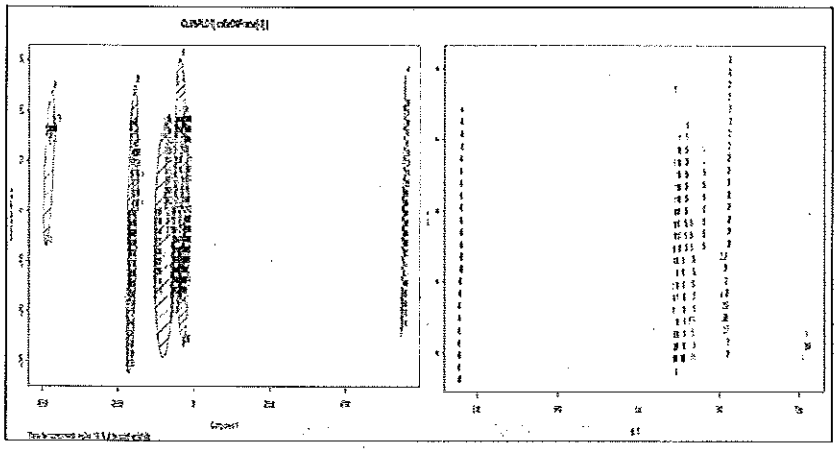

Fig. 8 a) Example of Clustplot of five clusters produced from Rscript. b) Example of Centroid cluster classification produced from Rscript

5 Discussion

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 will 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 the changes in land use with changes in climate and other agricultural factors.

References

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

2. Dunstan N., Armstrong L. and Diepeveen D.: 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 (2009).
3. Ekasingh, B., Ngamsomsuke, K., Letcher, R., & Spate, J.: 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 (2005).
4. Rivers, M.R.: Overview of the Peel Inlet and Harvey Estuary – genesis to water quality. Proceedings of the 7th International River Symposium. Brisbane, Queensland (2004).
5. Rose, T.: Catchment water management planning. Retrieved February 6, 2010, from http://www.ecohydrology.uwa.edu.au/research/cwmp (2003).
6. Western Australian Department of Environment and Conversation: 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 (2003)
7. Gahegan, D. G.: Spatial ordering and encoding for geographic data mining and visualization. Journal of Intelligent Information Systems, 27(3), 243 (2006).
8. Deren, L. I., Kaichang, D. I., & Deyi, L. I.: 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 (2000).
9. Ding, Q., Ding, Q., & Perrizo, W. PARM--an efficient algorithm to mine association rules from spatial data. IEEE Transactions on Systems, Man, and Cybernetics--Part B: Cybernetics, 38(6), 1513. (2008).
10. Jacquez, G. M.: Spatial Cluster Analysis. Blackwell Publishing, pages 395-416 (n.d.).
11. Zeilhofer, P., Lima, E. B. N. R., & Lima, G. A. R.: 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 (2006).
12. 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
13. Wachowicz, M: Uncovering Spatio-Temporal Patterns in Environmental Data. Water Resources Management, 16(6), 469-487. doi:10.1023/A:1022259531710 (2002).