2007

An Investigation Into the Application of Data Mining Techniques to Characterize Agricultural Soil Profiles

Rowan J. Maddern
Edith Cowan University

Edith Cowan University

Copyright Warning

An investigation into the application of data mining techniques to characterize agricultural soil profiles.

Rowan J. Maddern

Faculty: Computing, Health and Science.
Institution: Edith Cowan University.

Supervisor: Dr L Armstrong

February 2007

This dissertation is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science Communications and Information Technology with Honours.

USE OF THESIS

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

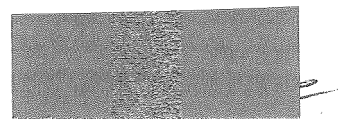
Declaration

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

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

(ii) Contain any material previously published or written by another person except where due reference is made in the text; or

(iii) Contain any defamatory material



Rowan J. Maddern

05/05/2007

Acknowledgments

During the process of undertaking this research I have received the help and contributions of a number of people. Without their help the research would not have been possible. They are too numerous to mention all of them here, but I would like to thank them all. I would particularly like to thank Chris Richardson who personally provided a contact at the DAFWA; Ted Griffin and Dr Leisa Armstrong whose supervision and help went above and beyond; I would also like to thank the staff from the Higher Degrees office at ECU for guidance and resources and Tony Maddern for his on-going help.

Abstract

The advances in computing and information storage have provided vast amounts of data. The challenge has been to extract knowledge from this raw data; this has led to new methods and techniques such as data mining that can bridge the knowledge gap. The research aims to use these new data mining techniques and apply them to a soil science database to establish if meaningful relationships can be found. A data set extracted from the WA Department of Agriculture and Food (DAFWA) soils database has been used to conduct this research. The database contains measurements of soil profile data from various locations throughout the south west agricultural region of Western Australia. The research established that meaningful relationships can be found in the soil profile data at different locations. In addition, comparison was made between current cluster techniques and statistical methods to establish the most effective method.

The research compared two data mining algorithms against a benchmark that was established using standard statistical analysis in use at the DAFWA. The EM and FarthestFirst data mining algorithms were tested in five case studies and it was found that FarthestFirst was more accurate at clustering instances than EM in all cases when tested against actual known clusters groups. The known groups were two traits EC and Clay within two soil types. It was concluded that data mining had a number of advantages over current statistical methods but the methods research can not completely replace them at this stage.

The outcome of the research may have many benefits: to agriculture in general, to soil management and to environmental management. The research has been collaboration between the Edith Cowan University and the DAFWA, with the results and outcomes to be shared between the two organizations.

Table of Contents

List of Figures

List of Tables

1. Introduction

In the current age of technology the collection, storage, analysis and use of data has become critical in understanding complex systems and relationships within those systems. The computerization of data collection has enabled vast amounts of data to be collected and stored; this data has application in marketing, science, engineering, agriculture and many other disciplines.

"The convergence of computing and communication has produced a society that feeds on information." (Witten and Frank, 2005, p. xxii). The interpretation of the data has become as important as the data itself, for without understanding the patterns and trends contained within the data the information becomes less valuable. The amount of information has meant that the analysis of the data has become too onerous for humans and a technique of computer based analysis is required.

This need to extract knowledge from information has resulted in the development of computer systems that can bridge the gap between human understanding and the volume of information. The development of data mining and machine learning techniques has provided the methods to solve this growing problem by allowing large amounts of information to be quickly analyzed. "Data mining is the extraction of implicit, previously unknown, and potentially useful information from data" (Witten and Frank, 2005, p. xxii).

Data mining software applications, using various methodologies, have been developed by both commercial and research centres. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining has been used to analyse large data sets and establish useful classification and patterns in the data sets. "agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods" (Cunningham and Holmes, 2005, p.5).

The research determined whether data mining techniques can also be used to improve pattern recognition and analysis of large soil profile experimental datasets. Furthermore, the research did establish if data mining techniques can be used to assist in the classification methods by determining whether meaningful patterns exist across various soils profiles characterized at various research sites across Western Australia. Various data mining techniques that were used to analyse a large data set of soil properties attributes. The data set has been assembled from soil surveys of Western Australian agricultural areas. The research has utilized existing data collected from ten commonly occurring soil types in order to establish patterns and correlations between a number of soil properties. The soil studies, which have been conducted by the Western Australian Department of Agriculture (DAFWA) researchers over the past 20 years, provide a vast amount of information on the classification of soil profiles and chemical characteristics.

The analysis of these agricultural data sets with various data mining techniques may yield outcomes useful to researchers in the DAFWA. It is envisaged that the information gained from this research will contribute to the improvement and maintenance of soils and the agricultural environment of Western Australia.

1.1 Research question

Can the application of data mining techniques to an agricultural soil profile data set improve the verification of valid patterns and profile clusters when compared to standard statistical analysis techniques?

Sub questions:

a) What current analysis techniques are being used to determine valid patterns and soil profile clusters?

b) What standard data mining techniques, when used on the data set can establish valid patterns and soil profile clusters?

c) Which data mining techniques are the most efficient in determining patterns and clusters when compared to standard statistical analysis techniques?

1.2 Background to the study

The DAFWA conducted a large scale soil mapping project in the south west of the state in the mid-1980s. This soil mapping project was conducted with the support of the National Soil Conservation Program (NSCP), National Landcare Program (LCP) and Natural Heritage Trust. Current classification techniques to analyse the soil survey data have been outlined (Schoknecht, Tille, and Purdie, 2004). Work by the soil scientist, Purdie in the early 1990s led to the standardization of the methods and outputs of the soil-landscape mapping program (Schoknecht, Tille, and Purdie, 2004, p.9). This included the development of a nested hierarchy of soil-landscape mapping units (Please see Appendix 8.1). This new method was advantageous as it allowed varying levels of information to be displayed from varying scales of mapping. In addition the standard classification allowed for possible correlations to be established between different surveys and also enabled computer processing of data on a statewide (and national) level. It also provided a means by which the pre-existing survey could be incorporated into a seamless map across the agricultural districts (Schoknecht, Tille, and Purdie, 2004, p.10).

This Purdie soil classification is the basis of Australian soil classification standards which have subsequently been adopted by DAFWA as the official system (Isbell, 1996). The use of soil classification maps has been shown to play a substantial role in agricultural production, salt control, large scale land management and land improvement. This has allowed a greater understanding of biophysical and environmental management (Schoknecht, 2002). The soil profile data set contains a high level of variability in some survey sites with limited set of experimental data. For example, a reduced number of soil attributes is available for older developed areas, due to testing being carried out at the time the land was cleared.

According to Schoknecht, Tille, Purdie (2004, p. 14) "the soil groups of Western Australia are classified into 60 main groups; this provides a standard way of giving common names to the main soils of the state". Thirteen soil super groups are defined using three primary criteria: texture or permeability profile; coarse fragments and water regime. Sixty soil groups are defined by further divisions of the soil super groups based on one or more of the following secondary and tertiary criteria: calcareous layer (presence of carbonates), colour, depth or horizons/ profile, pH (acidity/alkalinity); structure.

Further, (Schoknecht, 2002, p.5) outlines the collection of data in the field:

> Soil description is best conducted on an exposed profile such as a pit or road cutting, but alternatively using a soil auger or coring device. In the field the soil profile is divided into layers (horizons) based on one or more above properties listed above. The properties, depth and arrangement of the layers are used to assign the soil to a soil super group or soil group. (p. 5)

Figure 1: Images of soil profiles of experimental field sites located in the southwest agricultural region.
(Schoknecht, 2002, p. 177 and p.121).

Soil profile data is collected through the exposure of the site, as shown in Figure 1. The soil profile data set was collected from soil surveys of Western Australia over the last 20 years. A high level of variability was found in the data set. Some surveys sets have limited sets of experimental data, for example a reduced number of soil attributes is available for older developed areas due to testing being carried out at the time of land clearing.

Following the collection of the soil profile data, all data was stored in a central database. Measurements were made based on a visual assessment of the profile, notes on soil location including longitude and latitude and chemical analysis of soil samples taken across the profile site. The data was compiled into a number of different forms within the database with the forms linked by unique keys. The DAFWA – soil profile version 3.5.0 database is an MS Access database that allows the collection and extraction of data via a graphical user interface (GUI). The database allows large amounts of soil data to be entered, stored and accessed quickly; this tool has meant that research into agricultural soil can be conducted on a large scale.

1.3 The significance of the study

The research has a number of potential benefits to the DAFWA and the users of land within the south west land division of Western Australia. The collection and storage of large amounts of data in the DAFWA – Soil Profile Version 3.5.0 database has provided a valuable tool in the study of soils across Western Australia agricultural regions. However, the analysis and interpretation of such a large data set is problematic due to nature and volume of the data. The proposed study will establish if new data mining techniques will improve the effectiveness and accuracy of the analysis of such large data sets. The analysis of such soil data sets is difficult given the complex relationships between large numbers of variables collected for each geographical location. The current process uses standard statistical procedures to interpret the soil profile data sets. The use of standard statistical analysis techniques is both time consuming and expensive. If alternative techniques can be found to improve this process, an improvement in the management of these soil environments may result.

The outcomes of the research could improve the management and systems of soil uses throughout a large number of fields that includes Agriculture, horticulture, environmental and land use management. The application of data mining techniques has never been conducted for Western Australia soil data sets. A comparison of data mining techniques and statistical methods could produce a model for further understanding the data. In addition, the research could remove the constraining factors that have limited soil scientist's effective utilization of the large amounts of data collected in the last 20 years of research. The benefits of a greater understanding of soils could improve productivity in farming, maintain biodiversity, reduce reliance on fertilizers and create a better integrated soil management system for both the private and public sectors.

The research could be extended in the future with the possible inclusion of additional soil variables; these factors could include other location site information such as climatic data. This could result in the effective uses of soil profile data for the improvement of crop agronomy practices (Moore, 2004, p.3). A new method of interpretation of data could improve knowledge and the methods of data collection, with important factors within the data having been identified. The outcomes of the proposed research could be used for the creation of models of soils within the survey areas that could reduce the cost of data collection by reducing the amount of data collection required in the future.

1.4 The purpose of the study / statement of the problem

The purpose of the study is to examine the most effective techniques to extract new knowledge and information from existing soil profile data contained within DAFWA soils database. DAFWA has collected a large amount of information within its database system; however this data has limited meaning. The study will apply data mining methods to a subset of data created by the DAFWA researchers to facilitate an improvement in the interpretation of the soil profile data set.

The Western Australian soil profile data set to be utilized in the proposed investigation has been selected as it is a representative sample of the data sets population that has been collected over the past 20 years. Mr. Ted Griffin, soil scientist for the DAFWA outlined the limitations of time, resources and data complexity to conducting in-depth analysis to date (see appendix 8.2). The data set has allowed each soil type to be compared in a number of geographical locations, for example a loamy gravel soil from Wagin to be compared with loamy gravel from Albany.

It is envisaged that the application of new techniques to the selected data set may overcome the limitations of current soil science research methods. In addition, it will provide a framework of methods that can be applied from the sample population to larger soil databases. The research has overcome a number of problems contained within the data set; one problem was contained was in the data source and was a small numbers of missing values. The data has been collected from a natural source that contains missing values that could affect results of any experiments conducted and require clearing prior to commencement. The problem that the study aims to overcome is the selection of the correct methods to apply within the data mining application. In addition the selection of appropriate data mining techniques is critical in the understanding of the soil profile data. In order for this process to be of some benefit to the understanding of soil characteristics, the findings must be discussed in close consultation with DAFWA statisticians and other soil experts.

Table 1: Definitions of terms or operational definitions

TERM	DESCRIPTION	SOURCE
Ag Data	The Agriculture soil data set used for the proposed research.	Mr. E.A Griffin
Algorithm	A precise rule (or set of rules) specifying how to solve some problem.	(*Online dictionary*, 2005)
Calcium (Ca)	A white metallic element that burns with a brilliant light; the fifth most abundant element in the earth's crust; an important component of most plants and animals.	(Northcote, 1984)
Cation	A positively charged ion.	(Northcote, 1984)
Data mining	Data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases; a way to discover new meaning in data.	(*Online dictionary*, 2005)
Horizon	Layer within the soil profile having morphological characteristics and properties different from the layers which occur below and / or above it.	(Northcote, 1984)
Horizon A Zone of depilation	Master horizons, either consisting of one or more surfaces or mineral horizons with some organic accumulation and darker in colour than underlying horizons or consisting of surface and subsurface horizons that are lighter in colour but have a lower content of clay minerals, iron, and aluminum than underlying horizons.	(Northcote, 1984)
Horizon B zone of collection	Master horizon consisting of one or more mineral soil layers characterized by (a) a concentration of clay, and /or iron, and/or aluminum, and/or translocated organic material; and/or (b) having a structure and/or consistency unlike that of the A horizon above.	(Northcote, 1984)
Machine learning	The ability for a machine to get knowledge by study, experience or being taught.	(Witten and Frank, 2005)
Magnesium (Mg)	A light silver-white ductile bivalent metallic element; in pure form it burns with brilliant white flame; occurs naturally only in combination.	(Northcote, 1984)
Neural network	A network of many simple processors that imitates a biological neural network. Neural networks have some ability to "learn" from experience, and are used in applications such as speech recognition, robotics, medical diagnosis, signal processing, and weather forecasting. Also called artificial neural network.	(*Online dictionary*, 2005)

TERM	DESCRIPTION	SOURCE
Potassium (K)	A light soft silver-white metallic element of compounds. The alkali metal group; oxidizes rapidly in air and reacts violently with water; is abundant in nature in combined forms occurring in sea water and in carnallite and kainite and sylvite.	(Northcote, 1984)
Soil profile	The soil profile is the face of soil exposed in a vertical section. More realistically it is a column or prism of soil of small cross-sectional area and extending from the soil surface to the parent material.	(Northcote, 1984)
Sodium (Na)	A soft silver-white reactive metallic chemical element of which common salt and soda are derived.	(Northcote, 1984)
Soil classification	The type of soil that is found at a location for example: sand or clay. There are many combinations of types of soils.	(Moore, 2004)
WEKA	Waikato Environment for Knowledge Analysis data mining application based on a JAVA platform.	(Frank, Hall, and Trigg, 2005)

2 Review of the literature

2.1 Introduction

The revolution brought about by the information age has provided technology to handle vast amounts of information faster than at any stage in history. The information stored in these databases has been collected from countless sources and with the use of modern techniques and storage devices; this has facilitated easy access to large data sets. The challenge that has arisen with this new information asset has been to acquire knowledge and for this raw information and technology to verify this acquired knowledge. Computer science has provided a number of techniques including artificial intelligence (AI), machine learning, data mining, decision trees, neutral systems and statistical analysis which have aided in the effective searching, validation and interpretation of this raw information.

Bentley, (1997, p.1) states that "Humans can handle a number of variables when they consider patterns, possibly as low as eight, whereas machines can handle hundreds or thousands of variables". The process of acquiring any new knowledge has a number of steps that require a fundamental understanding of the data first before any methods can be applied. This understanding allows the selection of the correct process, the process involves selection of the data set, cleaning or allowing for functions within the data (missing data, outliers), application of an algorithm, and interpretation of any results.

The application of data mining techniques to a number of large databases, with missing and incomplete data has resulted in valued outcomes for a diverse range of medical, agricultural and commercial applications. Previous research studies have focused on ways to gain new knowledge and improve application of previously collected data to gain improvements in productivity.

The application of this technology was outlined by Witten and Frank, (1999, p.1) by showing that "Technology now allows us to capture and store vast quantities of data. Finding and summarizing the patterns, trends and anomalies in these data sets is one of the grand challenges of the information age".

The need to understand and find new knowledge in vast amounts of raw information has proven to be far outside the capacity of humans without the aid of modern techniques and tools. The techniques and methods of data mining have been possible due to the development of AI and machine learning research developments. Modern methods of data analyses requires a computational based approach to acquire this new knowledge and the algorithms used to conduct data mining depend on the type of information and outputs required. The uses of modern technology to gain new knowledge was outlined by (Bentley, 1997, p.1) "Once again as with any technology the answer lies in the successful application of the new capabilities."

A number of studies have applied data mining techniques to extract meaning from data collected from natural systems research. For example, the collection of data from natural systems is challenging, with most of the data sets incomplete due to the difficulty and methods of data collection. Missing data sets can be problematic and may limit the analysis and extraction of new knowledge. The problem of missing values was analysed by Ragel and Cremilleux (1999, p.1): "To complete missing values a solution is to use relevant associations between the attributes of the data. The problem is that it is not an easy task to discover relations in the data containing missing values."

A number of research groups have established methods to improve data mining techniques. In addition, these groups have developed software applications which provide a user friendly interface to facilitate the application of these data mining techniques to large data sets. A research group from the University of Waikato (New Zealand) is a leader in the field of data mining and is the developer of WEKA. WEKA is a tool which has been used in a number of data mining research studies. These studies have analyzed biological data sets from a wide range of research areas that are as diverse as mushrooms grading, cancer prediction and soil chemical analysis.

The following sections are a review of literature. The review will detail the current research being conducted into data analysis and the application of these techniques to acquire new knowledge from raw data. The review will detail topics which include, data mining techniques, current research trends, statistical analysis, application of data mining to industrial / medical databases, decision support systems, software and algorithms. The review aims to outline the major areas of current research and to provide an understanding of the topic so that the research may be conducted.

2.2 The Acquisition of Knowledge

Data mining is the process of gaining new knowledge from data that was not previously known due to the complex relationship within the data. Access to data has never been greater in the history of human development , (Witten and Frank, 1999, p.10) outlined this when they said "The convergence of computing and communications has produced a society that feeds on information. Yet most of the information is in its raw form." The acquisition of this knowledge has many applications for research, industry and business, by lowering cost, improving productivity, and finding new processes.

Developing understanding and finding patterns within datasets extracted from large databases is a complex task. The process has become so complex that any value that may be gained from this information is lost due to the limitations of human learning and processing ability. The aim of data mining is to overcome these limitations by using computational based systems and software. Data mining system are designed to search through the data and automatically establish patterns and trends within the data set. In addition the software provides outcomes that present new knowledge in a basic format that may be understood. This process is by no means as simple as outlined above; the process of machine learning is complex.

The process was further outlined by Gertosio and Dussauchoy (2003) who provided an analysis of the data mining process:

> "The proliferation of large masses of data has created many opportunities for those working in science, engineering and business. The field of data mining (DM) and knowledge discovery from database (KDD) has emerged as a new discipline in engineering and computer science" (Gertosio and Dussauchoy, 2003, p.1)

The data mining process includes prediction, estimation, classification, and the development of rules based on the data contained within the data set (Brown and Kros, 2003). The process depends on the accuracy of the dataset that is being used; the accuracy is based on the completeness of the data and any missing data may affect the results of the data mining process. For example, in Brown and Kros (2003) the impact of missing data is researched and possible solutions to the problem are outlined.

2.3 Current research trends

The research has followed a number of different paths to obtain new knowledge from data. The research trends show that data mining and intelligence systems are being used more frequently than statistical and other older methods such as geostatistical methods. One of the major problems that all methods of data analysis must overcome is that the number of samples in a given study is limited. This is true for soil research: "We can measure the soil at only a finite number of places and times on small support and any statement concerning the soil at other places or times involves prediction" (Heuvelink and Webster, 2001). Research that is currently being undertaken with data mining tools is aimed at overcoming the gap that statistical analysis cannot handle.

Current research studies focus on the role of algorithms and improvement of accuracy to obtain a higher confidence level for the classification process. The current methods of application require a high level of understanding of the relationships between the variables within the dataset. This research aims to improve the outcomes and to make data mining easier to conduct, so that outcomes are more relevant to the user.

This current research also aims to provide a benchmark and an understanding of the correct selection of an algorithm for a data mining process; this is critical in obtaining the correct results and knowledge. The research conducted by Rajagopalan Krovi (2002) has provided a benchmark of a selection of algorithms to allow a comparison to be made. This study concluded that "successful implementation of data mining efforts requires a careful assessment of various tools and algorithms available" (Rajagopalan and Krovi, 2002, p.1).

2.3.1 Current Statistical analysis techniques

Research conducted in statistical modeling has allowed the mapping and / or prediction of variables for a number of different applications including, soil chemistry, fire prediction and modeling and vegetation. One such study by Little Edwards and Porter (1997) analyzed the impact of coastal development on estuarine ecosystems through standard statistical techniques. The study compared the accuracy of eight different kriging methods in the prediction of water quality variables in established waterways around South Carolina. Statistical and geostatistical techniques still provide a valuable tool in the analysis and interpretation of natural systems and data sets. For example, Little, Edwards and Porter (1997, p.1) outlined the uses of geostatistical methods:

> Geostatistical methods are becoming an essential tool for understanding the spatial distribution of biological and chemical species in estuaries. At the heart of these methods are the spatial prediction / mapping methods known as "kriging"; these can construct statistically optimal predictions for data at unobserved locations using a relatively small, spatially explicit sample.

In addition, similar methods used for the prediction of various soil properties have been described in a study by McBratney, Odeh, Bishop, Dunbar and Shatar, (2000) this study investigated and evaluated the evolution of soil science methodologies over the past 60 years. These techniques listed below show the other methods of statistical analysis that have been used in soil science: Numerical classification, Fussy logic and Fussy sets, Pedodiversity, Geostatistical techniques, Hybrid techniques, Universal kriging, Cokriging, Regression kriging, Regression tree , Kriging with external drift, Factorial kriging (McBratney, Odeh, Bishop, Dunbar, and Shatar, 2000).

The research conducted by McBratney, Odeh, Bishop, Dunbar and Shatar, (2000) outlines the need for careful selection of soil analysis techniques to provide an accurate result for any given research. They state that "Application of each of the pedometric techniques depends on the purpose, resolution and setup of the survey as the ultimate use of soil survey information determines the accuracy required." (McBratney, Odeh, Bishop, Dunbar, and Shatar, 2000, p.30).

The use of statistics to conduct data analysis has a number of benefits that allow research to be undertaken. This includes ease of application, low level of understanding required for application, ease of interoperation of results. Research carried out by Selvanathan, Selvanathan, Kellor, and Warrack (2000) outlines the methods of conducing statistical analysis. They state that "statistics is a body of principles and methods concerned with extracting useful information from a set of numerical data to help managers make a decisions" (Selvanathan, Selvanathan, Kellor, and Warrack, 2000, p.1).

The application of statistics methods to acquire new knowledge out of information has a long history and application is relatively simple with small sets of data. As the data complexity and volume increase so do the time and skills required to extract useful knowledge (Selvanathan, Selvanathan, Kellor, and Warrack, 2000).

2.3.2 Data mining

Investigations to improve data mining techniques and provide effective tools for data mining activities have been undertaken by a number of university groups including the University of Waikato and Monash. These groups have undertaken research into the development of algorithms for the classification of large data sets. The development process has been combined with a computer application to allow users to conduct data mining activities without the need for a fundamental understanding of the underlying process (Witten and Frank, 2005). Data mining techniques are used to develop models of the schematic nature of the data sets in order to develop predictions or classifications Cunningham and Holmes (2005, p.1) discussed the application of these models:

> The 'mined' information is typically represented as a model for the semantic structure of the dataset, where models may be used on new data for prediction or classification. Alternatively, human domain experts may choose to manually examine the model, in search of portions that explain previously misunderstood or known characteristics of the domain under study.

The selection of the correct data mining method has become very important to extracting new knowledge for data.

> With the development and penetration of data mining within different fields and industries, many data mining algorithms have emerged. The selection of a good data mining algorithm to obtain the best results on a particular data set has become very important. What works well for a particular data set may not work on another (Ibrahim, 1999).

Several types of analysis methods are available but generally four types of relationships are sought, these are outlined by Palace (1996, p.3):

> **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
>
> **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
>
> **Associations:** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
>
> **Sequential patterns:** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

The relationships are applied according to the required results of the research being conducted, with a combination of some or all used depending on the situation. The application of these techniques will be detailed in section 2.6.

2.4 Decision support systems

Decision support systems (DSS) are common tool used to solve problems where an action or outcome is required from the input of data. DSS have many applications in natural systems and with the advancement of information technology can be used to solve biological problems on a large scale. For example research carried at the University College Dublin investigated the use of DSS to improve milk collection practices. Butler, Herlihy and Keenan (2004, p.1) stated that "the geographic information system (GIS) based DSS allows a scheduler to interact with optimization algorithms to plan milk collection routes." The developed systems collected real time data from each producer and then correlated it allowing the information management systems to improve the collection and distribution process. The system allowed the collection of the milk at the most critical time and reduced collection cost and running cost for the producers.

Decision support systems have also been used in Agriculture to improve the long term productivity and management of cropping systems. Research undertaken by Canillas and Salokhe, (2001) analyzed three soil types, clay, silty clay loam and silty loam to quantify the effects of type variables and soils variables to develop compaction models for assessment. An application of DSS was investigated in order to improve soil compaction assessment in agricultural soils.

> The study conducted by Canillas and Salokhe (2001, p.1) reported that: "The soil compaction models were found to provide good prediction of bulk density and core index. Using the compaction models and other secondary data, the decision support system was developed to access the compaction status of the soil in relation to crop yields".

Based on the data collected a predictive model was produced to map and understand the relationship between the variables and provide a greater understanding of the effects of certain factors.

DSS have been created for a large number of applications to help take the guess work out of difficult decisions based on limited amount of data. For example, in a study by Goddard, Harms, Reichenbach, Tadesse and Waltman, (2003, p.1), an examination was made of how a DSS could be used to overcome problems related to global-drought-risk-management. It was proposed that the research could improve drought risk management and reduce cost of annual drought-related losses which been by the Federal Emergency Management Agency in the US at US $6 – 8 billion dollars. The project to develop a National Agricultural Decision Support System (NADSS) was carried out by the University of Nebraska-Lincoln and a number of US government departments. It was concluded that the key to the system was to obtain data about environmental events over a period of time, with the key to the warning system the understanding of past historical events and the probability of drought in time and space. The data for the NADSS has been collected from a number of different sources that include weather stations, various

geospatial databases to create drought indices, risk assessment and exposure analysis (Goddard, Harms, Reichenbach, Tadesse, and Waltman, 2003, p.2). The researchers developed two new algorithms that mapped the relationships between climate and oceanic parameters to create a DSS system to allow farmers to assess their own risk via a web interface. The project has allowed farmers to improve planting, tillage and soil management practices to reduce the loss for drought.

2.5 Data mining software

The complex problem of data mining has triggered the creation of a number of software systems to facilitate the efficient application of data mining techniques. A large number of data mining applications are available and include WEKA and YALE (Rudi Alberts et al., 2005).

These software systems provide platform and methods for knowledge discovery processes and their application to data mining techniques requires understanding of the outcomes required and the limitations of the systems. The developments of software systems are ongoing and changing with research and development. The YALE software is an environment for machine learning experiments and data mining. Applications of YALE cover both research and real-world data mining tasks (Rudi Alberts et al., 2005, p.1). This software provides the means to conduct experiments with a large number of arbitrarily nestable operators. The operator's setup is described by using XML files that can easily be created with a graphical user interface.

The computer science department at the University of Waikato has undertaken one of the major research studies into the application of these data mining techniques. The major researchers involved are Frank and Witten and other members of the university have developed a software package called Waikato Environment for Knowledge Analysis (WEKA).

Frank and Witten (2005) stated that "The Weka machine learning workbench provides a general-purpose environment for automatic classification, regression, clustering and feature selection." p 366. The machine learning function in WEKA allows for decision tree to be constructed, this type of analysis method may be required in the data mining of the Agdata set.

The algorithms that have been developed for data mining are integrated in a GUI software package that allows the program to be applied to different instances. The creation of a data mining benchmark tool such as WEKA provides a platform for which other developments can be made to suit individual data sets. The WEKA platform was developed using Java and is open source programming code so new algorithms may be created. Ibrahim outlined the benefits of using WEKA:

> Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written using java and runs on almost any platform. The algorithms can either be applied directly to a data set or called from Java code. WEKA is also well suited for developing new machine learning schemes. WEKA is open source software issued under public license. Implemented schemes for classification include decision tree inducers, rule learners, Naïve Bayes decision tables, locally weighted regression, support vector machines, instance-based learners, logistic regression and voted perceptions (Ibrahim, 1999, p.45).

Farrand, (2002) concluded that the data mining process can be tailored to the desired outcomes. For example, the "WekaMetal is a Meta-Learning extension to the data-mining package Weka has been used by a number of researchers. It provides additional algorithm selection, based on the expected accuracy and time performance" (Farrand, 2002, p.1) The tools are created so that they can be tailored to meet the need of each data mining project; this is critical to obtain correct results.

2.6 Data mining Techniques

Data mining requires the use of a number of techniques to allow the process of machine learning to be performed.

The analysis of any data requires an understanding of the outcomes required and the methods of analysis. There are a number of different systems that can be applied in the knowledge discovery process they include decision trees, clustering and neural networks.

2.6.1 Decision trees

Decision trees display all the possible outcomes of an expression they typically use Boolean expression. Zhang, Valentine and Kemp, (2004, p.1) argue that "decision trees, one of the data mining methods, has been widely used as a modeling approach and has shown better predictive ability than traditional approaches (e.g. regression)".

Zhang, Valentine and Kemp, (2004) used a decision tree model to map the productivity of naturalised pasture in the north island of New Zealand. The study developed a number of models for the prediction of natural hill top grass growth. The dataset was collected from research conducted over a number of years; this data included pasture production, soil properties (bulk density test, pH), fertilizer management, and topography. The research concluded that the model was more accurate in a number of ways but does not include much discussion on the methods of creation of the models. It was found that the decision tree model had the smaller average squared error (ASE) when compared with regression models. In addition, predictions based on the decision tree were twice as accurate as these based on the regression models and were 90 per cent accurate. Other research has found that decision trees maybe inaccurate and limited in their application. Iverson and Prasad (1998, p.10) outlined the limitations:

> Naturally, decision tree also has its limitations: it requires a relatively large amount of training data; it can not express linear relationships in a simple and concise way like regression does; it cannot produce a continuous output due to its binary nature; and it has no unique solution that is there is no best solution.

2.6.2 Clustering

Clustering is the process of grouping similar instances together so that analysis and models can be created. As a tool, it allows data to be grouped and analyzed. It has become an important part of the pre data mining process. In the study conducted by Berry and Linoff (1997, p.12), the methods and application of simple clustering was discussed: "Making sense of complex issues is naturally approached by breaking the subject into smaller segments that can be each explained more simply. Clustering aims at finding smaller and more homogeneous groups from a large heterogeneous collection of items." Clustering techniques are usually applied during the early part of the knowledge discovery process; clustering will show if the dataset has potential for the application of data mining techniques.

Clustering approaches have been applied to the analysis of medical databases. The research that has been undertaken over the last 15 years into the fields of conceptual clustering and AI unsupervised learning paradigms has provided a greater understanding of data mining techniques for medical databases (Veiga, 1996). Research studies have identified a number of functions that any clustering algorithm should fulfill when conducting biomedical research. Clustering has also been applied to a number of different applications in data mining. During a study conducted by Ryu and Eick in 2004 the methodology and tools to apply clustering to databases were researched. Ryu and Eick, (2004) stated that

> Clustering is a popular data analysis and data mining technique. However applying traditional clustering algorithms directly to a database is not straightforward due to the fact that a database usually consists of structured and related data; moreover, there might be several objects views of the database to be clustered, depending on a data analyst's particular interest. (Ryu and Eick, 2004, p.1)

The aims of the study were to identify those discrepancies and to show the impact of clustering techniques on databases, with an interest in ways clustering can be applied to real world databases. The application of clustering to a database was conducted in steps. These seven steps were outlined by Ryu and Eick, (2004, p.3-4).

1. Define object-view
2. Select relevant attributes
3. Generate suitable input format for clustering tool
4. Define similarity measure
5. Select parameter settings for the chosen algorithm
6. Run clustering algorithm
7 characterize the computed clusters

The research outlines the process involved in different clustering applications and the methods that have been improved from ordinary clustering to database clustering. Figure 2 shows an example of the clustering process as defined by Ryu and Eick (2004). The process shows the steps required when undertaking a cluster analysis on an ordinary database. The process in figure 2 shows how knowledge can be produced by the clustering on raw data and in this example outlines the groups of people and when they are likely to eat during a giving data. The process produced three clusters with young people coming in at midnight, retired people coming at lunch time and white collar coming at for dinner. This knowledge allows the restaurant to improve there menu and service to take advantage of these times and groups.

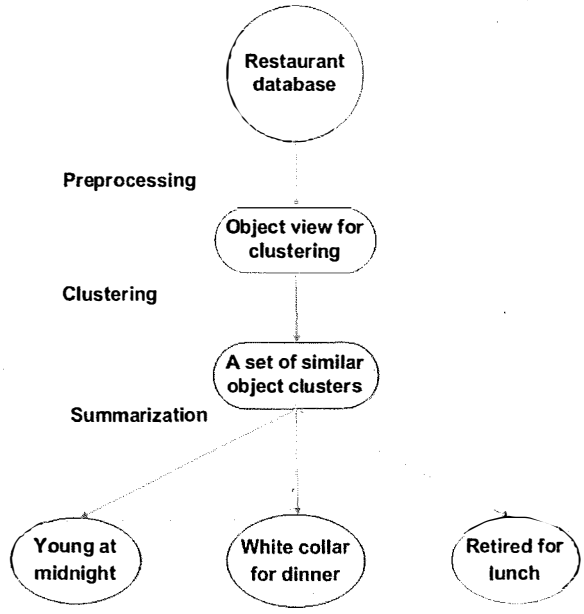


Figure 2: Example of database clustering.
(Ryu and Eick, 2004, p.3)

Any clustering process requires the selection of the correct algorithm to conduct the research and to ensure that the variables are grouped correctly. The techniques for the selection of a cluster algorithm may be very complex and require an in depth understanding of the clustering process. The increase in popularity of clustering in the past 10 years has resulted in an increase in the number of algorithms and their complexity. The research conducted by Grabmeier and Rudolph (2001, p.2) outlined the techniques for the selection of the correct data mining algorithm is outlined. They provided a general outline of the main problems and stated that "One main problem is that measuring a quality of resulting clusters depends heavily on the original application problem." The processes outlined in the paper were highly complex and outside the scope of this research.

2.6.3 Artificial neural networks

Artificial neural networks (ANN) are rule based learning systems that are conducted to provide an "answer" to a question by the creation of rules and paths. This method of analysis is modeled on human neural pathways. A study conducted by Kaul, Hill and Walthall (2004, p. 1) researched the prediction of yield for corn and soybean using a modeled with the aid of a neural network. The objectives of the study were to:

(1) Investigate if ANN could effectively predict Maryland's corn and soybean yields for typical climatic conditions

(2) Compare the prediction capabilities of models at state, regional and local levels

(3) Evaluate ANN models performance relative to variations of developmental parameters

(4) Compare the effectiveness of multiple linear regression models to ANN models

The research developed crop growth models based on a number of variables including soil, climate and crop factors. The development of the model used a feed-forward back-propagating ANN structure as shown below:

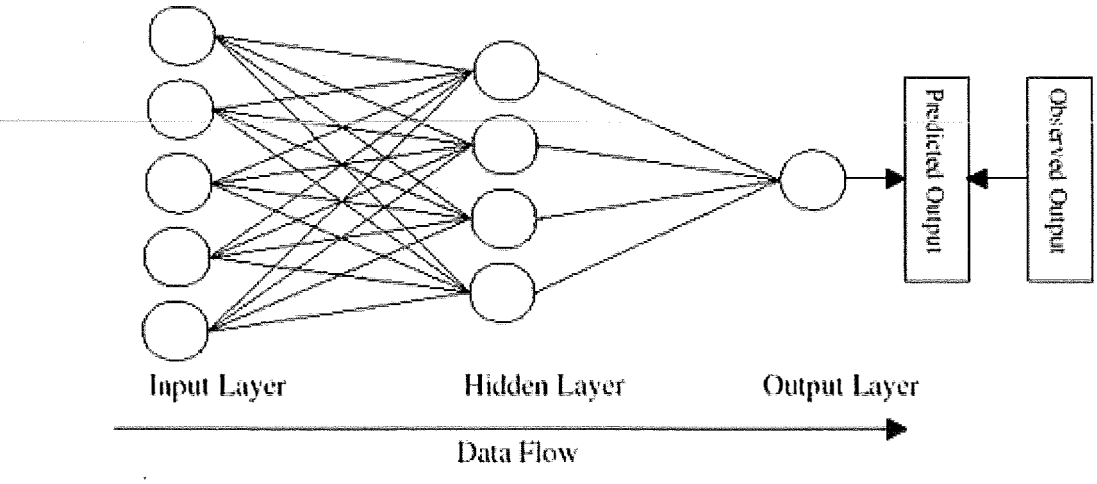


Figure 3: Layers and connections of a feed-forward back propagating artificial neural network.

(Kaul, Hill, and Walthall, 2004, p.3)

The system was applied to a training set of data, with weekly and monthly rainfall data entered into the system. The training and test data was randomized before the development of the model with training data containing observations for more than 60 corn and soybean varieties. The total data set held more than 100 corn and soybean varieties with the rest of the data set used for bench mark testing Kaul, Hill and Walthall, (2004). This study produced a model in parallel with the neural network and used linear regression to produce a model with the same dataset. The linear model was selected because it allowed a direct comparison of the outcome and accuracy with the same inputs. Kaul, Hill and Walthall (2004, p.16) outlined the benefits of the study:

> These ANN models have the potential to be useful as a component of nutrient management planning within Maryland given further development and validation. ANN modeling with additional locations will increase the variability of soil types and should broaden the usefulness, and possibly increase the predictive capabilities of ANN-based yield prediction.

ANN may be applied to solve a number of problems. Researchers at the University of Oklahoma, USA used ANNs to facilitate rainfall estimation (Trafalis, Richman, White & Santosa, 2002). The aim of their research was to use ground truth rainfall data to create models of rainfall patterns. This was possible with the collection of terabytes of data from ground weather stations. A number of different data analysis techniques were used including; back propagation, pattern recognition, clustering and ANNs. The research had mixed outcomes and it was concluded that further research was required.

2.7 Medical and industrial uses of data mining techniques

The uses of data mining techniques can be applied to a large number of applications within the medical and industrial fields of science and technology. For example, research conducted at Oklahoma State University predicting the survivability of breast cancer used three data mining method Delen, Walker and Kadam (2004, p.1) stated that "the prediction of breast cancer survivability has been a challenging research problem for many researchers." The purpose of the study was to improve screening techniques and improve treatments. The study used two data mining algorithms to develop the prediction models to analyse more than 200,000 cases. The process used artificial neural networks and decision trees along with statistical logistic regression to develop the system. The methods of data analysis were tested on 10-fold cross-validation methods to measure the greatest unbiased of the three prediction models for performance comparison purposes.

The data analyzed in the study by Delen, Walker and Kadam was collected from the SEER Cancer Incidence Public-Use Database from the years 1973 to 2000. The database held records that were assigned a case number to identify the patient within the system; each record contained 72 variables related to a specific incidence of cancer. A considerable amount of effort was needed in the research study to clean and prepare the data for modeling, approximately 80 per cent of the total time spent on the project. The results of the research showed the C5 decision tree was the most accurate with 93.6 per cent on the sample data, the second was artificial neural network with 91.2 per cent accuracy. The study concluded that both methods of data mining proved to be more accurate in the predication of breast cancer than conventional statistical methods with only 89.2 per cent accuracy.

The application of data mining techniques to medical databases has a number of precedents and has increased predictability and reduced the time expenditure of data analysis. The study conducted by Veiga (1996, p.1) outlined the application of clustering to a medical dataset. Veiga states that:

> Clustering is an important data analysis tool for discovering structure in data sets. Although research on conceptual clustering has produced algorithms showing significant advantages over earlier numerical ones, existing methods still present some limitations regarding applicability to biomedical domains.

Data mining techniques have been used on industrial data sets in an attempt to improve business function and productivity and reduce costs. The uses of advanced data mining methods have proved that its application can improve understanding of complex systems where knowledge is required from large amounts of information quickly to confirm a form of action and possibly improve or save a life. Gertosio and Dussauchoy (2003, p.3) conclude that the use of data mining and knowledge discovery for industrial engineering is beneficial for the development of new concepts and techniques in the field.

> The field of data mining (DM) and knowledge discovery from database (KDD) has emerged as a new discipline in engineering and computer science. In the modern sense of DM and KDD the focus tends to be extracting information characterized as "Knowledge" from data that can be very complex and in large quantities. Industrial engineering, with the diverse area it comprises, present unique opportunities for the application of DM and KDD, and for the development of new concepts and techniques in the field.

In the research conducted by Gertosio and Dussauchoy (2003), a French truck manufacturing company applied data mining to the analysis of its engine production line to decrease manufacturing time and to improve testing. The study was conducted on tests carried out on different capacity engines. The engines were checked at three phases: running in, stabilization and control for any signs of leaks, noises and tolerances outside the normal operation range of the engines. The recording of more than 30,000 measurements was made during the tests with the recording of 20 to 30 variables depending on the type of engine. The variables included speed, torque, temperatures, oil and water pressures. Due to the low occurrence of detected problems during the test, to obtain a sample group of errors a large number of tests were carried out to produce a

valid sized sample group. It was determined that a linear regression model provided the best results when conducting real time tests. Linear regression was used to create a model that reduced the performance of the test and reduced the overall test time by the real time analysis. The variables in engine faults were compared with the relationship between the variables and problems identified.

2.8 Data validity

The results of any data mining exercise depend on a number of factors that must be addressed before analysis can be undertaken. The presences of missing or null values in any data set can affect the outcomes of the data mining process depending on the type of data and the algorithm used to mine that data. The inconsistency in data can arise from a number of different reasons which can include procedural factors, refusal of response, inapplicable response, change in collection methods, missing data and information availability at time of collection (Brown and Kros, 2003). The type of data that is missing can also affect the process: "It is important for analyst to understand the different types of missing data before they can address the issue." (Brown and Kros, 2003, p.3). The common types of missing data can include: Data missing at random data missing completely at random non-ignorable missing data and outliers treated as missing data.

An explanation of these types of missing data and their effects on possible results and possible methods to counteract the effect is outlined by Brown and Kros (2003). This research highlighted the methods to address missing data. These methods include the following categories: use of complete data only, deleting selected cases or variables, data imputation and model-based approaches.

"These categories are based on randomness of the missing data and how the missing data is estimated and used for replacement" (Brown and Kros, 2003, p.3).

If the problem of missing data is not addressed it may affect the algorithms being used to mine that data. Brown and Kros have researched the effect of a number of algorithms commonly used. These include the K-nearest neighbor algorithm, Decision trees, Association rules and natural networks.

The problem that arises from the analysis of real world or natural data is that there are always missing or null values in any large data set. The problem can be overcome with the pre-data mining process by the replacement of the missing data with a value so that weight of missing data does not skew the outcome of the analysis. For example the data mining application WEKA allows missing information to be replaced with a "?" and so prevents it from affecting the outcome of the process. (Witten and Frank, 2005)

2.9 Similar studies to current research

A study by Ibrahim (1999) for the RMIT University investigated the uses of data mining techniques on agricultural data. This study applied six classification algorithms to 59 data sets and then six clustering algorithms were subsequently applied to the data generated. The results were studied and the patterns and properties of the clusters were formed to provide a base for the research. The research provided a comparison of performance for the 6 classification algorithms set to their default parameter settings. It was found that Kernel Density, C4.5 and Naïve Bayes followed by rule learner, IBK and OneR were the most accurate. The study utilized the WEKA data mining benchmark program.

The main objectives of the research conducted by Ibrahim (1999) were to:
(1) Build a file of data set names and the characteristics and performance of a number of algorithms on each data set; and
(2) Apply unsupervised clustering to the file built in step 1 to analyze the generated clusters and determine whether there are any significant patterns.
Ibrahim (1999, p.2) outlined a number of findings:

> It was discovered that number of instances was not useful in clustering the data sets, as it was the only significant variables in clustering the data sets before it was excluded from the generated data set. This prevented analysis based on other variables including the variables that contain values for the accuracy of each classification algorithm.

The research conducted by Ibrahim (1999) has provided a platform from which further work in this field might be undertaken. The scope of the research was limited and the investigation revealed a number of interesting clusters in machine learning performance data. It points to the fact that a larger investigation is required using more data sets and data set characteristics.

In another study WEKA was used to develop a classification system for the sorting and grading of mushrooms (Cunningham and Holmes, 2005). The system developed a classification system that could sort mushrooms into grades and attained a level of accuracy equal to or greater than the human inspectors. The process involved the pre-processing of the data, not just cleaning the data, but also creating a test dataset in conjunction with agricultural researchers. The attributes used to create the set included both objective and subjective measurement. The total dataset used a total of 282 mushroom types, criteria and attributes. The objective attributes were weight, firmness and percentage of cap opening. The subjective attributes were used to estimate the degree of dirt, stalk damage brushing, shrivel and bacterial blotch. The above data was collected and then compared with the grading of the three human inspectors and allocated a grade 1st, 2nd or 3rd.

The data, a total of 68 attributes including photo images, was used by the j4.8 algorithm classifier within WEKA to create a model for the human inspectors and the automated system. The model created using the human rules showed that each inspector used different combinations of attributes when assigning grades to mushrooms (Cunningham and Holmes, 2005). The application of data mining techniques provided within the WEKA software application created a model that analyzed all attributes and created a model that was faster and more accurate than the human system.

The decision tree analysis method has been used in the prediction of natural datasets in Agriculture and was found to be useful in prediction of soil depth for a dataset. In Mckenzie and Ryan (1999) the uses of slope angle, elevation, temperature and other factors were analyzed and models created for prediction of soil depth across a sample area. The model was tested through the use of random data sets. "at each level, trees with increasing numbers of terminal nodes were fitted 20 times with 5% of the data randomly selected and withheld to provide a test of the predictive strength of the model" (Mckenzie and Ryan, 1999). This process is outlined in Figure 4.

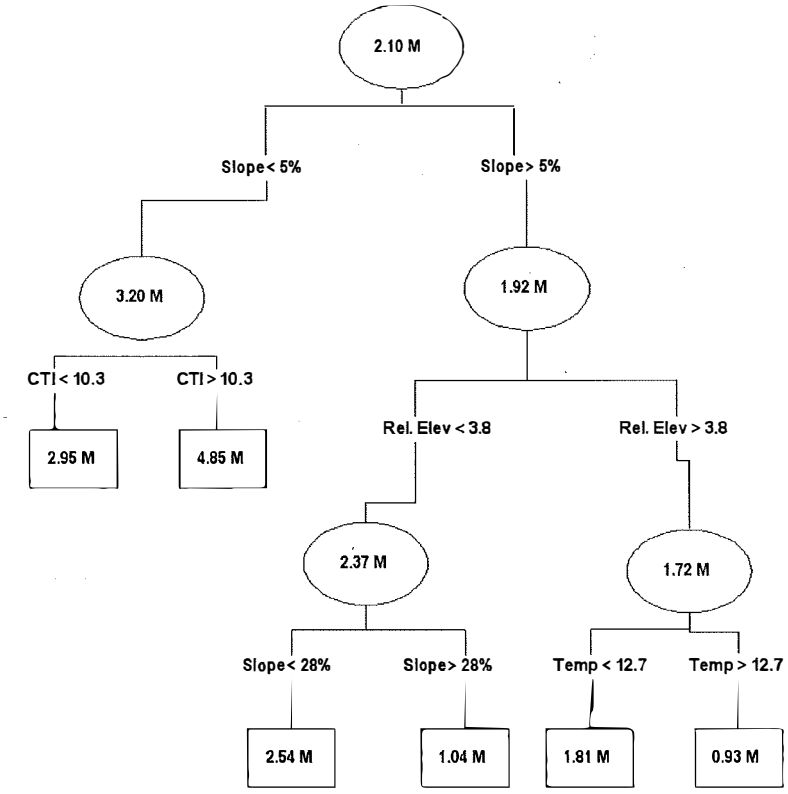


Figure 4: Regression tree.
(Mckenzie and Ryan, 1999, p.83)

2.10 Future trends

The analysis of data has a history as long as data has been collected and stored. The process of data analysis has mirrored the understanding, improvement and collection of the data, thus the process has been an evaluation not a revaluation.

In the study conducted by Shim, Warkentin, Courtney, Power, Sharda and Carlsson (2002, p.1) the evaluation of modern data analysis is outlined "Since the early 1970s, decision support systems (DSS) technology and applications have evolved significantly. Many technological and organizational developments have exerted an impact on this evolution'. The future of knowledge computing was outlined by Shim et al. (2002, p.13) when they identified the major objectives of DSS over the next 10 years:

DSS researchers and developers should; 1. Identify areas where tools are needed to transform uncertain and incomplete data, along with qualitative insights into useful knowledge. 2. be more prescriptive about effective decision making by using intelligent systems and methods. 3. Exploit advancing software tools to improve productivity of working and decision making time. 4. Assist and guide DSS practitioners in improving their core knowledge of effective decision support.

Research trends indicate that improvement in algorithms and methods will increase. The current data mining methods require an understanding of the fundamental techniques involved in the process, even with the aid of GUI software. The future of data mining and knowledge discovery will overcome the current short comings in process and provide improvements in the process so that the user obtains more knowledge.

3 Materials and methods

This section outlines the research approach and the equipment and methods required. The section provides a general outline of the steps required to carryout the research described in this thesis. The process required the data to be analyzed using statistical and data mining methods, with a comparison made of the two to answer the research question.

3.1 Overview

The research methodology required to conduct the research is a form of action research. This methodology pursues action or change and research understanding at the same time. The action research methodology is conducted in a spiral process which alternates between action and critical reflection (Clayden, 2005). Research is refined and a greater understanding of the subject is gained until the research is completed with the best possible result.

The process of applying data mining techniques to a natural database requires consideration of a number of factors. These factors are dataset size, format of data, complexity of the data and the range of factors within the dataset. The dataset used in the study was provided by Mr. Ted Griffin of the DAFWA. It contained ten soil types with a relatively complete range of values to be analyzed. The dataset was cleaned to remove any values outside the normal range, and null values were changed so that the data results are not affected. The ten soil types will be analyzed individually by both the statistical and data mining methods. The two techniques were then compared to determine the most efficient method for conducting analysis.

The research was conducted by establishing a benchmark method using Excel software to represent the current method of analysis. Data mining software was then used to produce comparative results. The selection of the correct statistical methods was critical and was made in conjunction with experts at Curtin and Murdoch universities. The comparison will allow the researcher to compare the two methods and improve the analysis techniques that are currently in place at the DAFWA.

The dataset was collected as part of a survey by Schoknecht, Tille and Purdie (2004), and included a large amount of information from different sites within the target area of Western Australia. This information was collected from various locations where a pit was dug and samples taken. The samples were then sent for chemical and physical analysis at the DAFWA laboratories in South Perth.

The data that was then stored in a database with the following information points and site data: "Site Description, soil profile description, soil classification, soil profile chemical properties, soil profile physical properties" (Schoknecht, Tille and Purdie, 2004, p.10).

The total number of sites analyzed was over 7000 with varying amounts of information obtained for each site. The amount of detail in the database about a given location varies in relation to the period in which it was taken. More in depth information was collected as sampling methods improved. The database is linked to other databases, a map unit database, a soil photos database and map unit polygons. The system uses Oracle and Microsoft Access as the platforms to run and maintain the data.

The classification of the soils is critical to the study because the soil typing must be the same in all locations across the study area for the results to be accurate. The soils were classified according to work by Schoknecht, (2002, p.5); that outlined the technique for the grouping of soil types. They are:

1. Soil super groups: Thirteen soil super groups are defined using three primary criteria: Texture or permeability profile, coarse fragments *(presence and nature)* and Water regime.
2. Soil groups: Sixty soil groups are defined by further divisions of the soil super groups based on one or more of the following secondary and tertiary criteria:

Calcareous layer (presence of carbonates): colour, depth or horizons/profile, Ph (acidity/alkalinity) and structure.

Further, Schoknecht (2002, p.5) provides the following useful description of soil in the field:

> Soil description is best conducted on an exposed profile such as a pit or road cutting, but alternatively using a soil auger or coring device. In the field the soil profile is divided into layers (horizons) based on one or more of the properties listed above. The properties, depth and arrangement of the layers are used to assign the soil to a soil super group or soil group.

The creation of soil data set was conducted over a number of years, with the first survey undertaken in the 1930s of the area around Salmon Gums by Burvill and Teakle for the CSIRO. Since then a number of surveys have been conducted by the CSIRO and DAFWA for a number of locations within Western Australia. The problems that have arisen from these surveys are the scale and the amount of chemical analysis conducted for each. Until a standardized method was introduced analysis methods were not uniform due to the large volume of samples taken. Not all chemical testing was conducted for all locations.

The methods of collection have generated gaps in the data set and not all samples for all locations contain all possible values.

3.2 Design

The research followed quantitative, positivistic methods for scientific research, (Clayden, 2005). The raw data for the experimentation was supplied by the DAFWA. The results of all experiments and recommendations will be supplied to DAFWA researchers upon completion and submission of the project. The dataset was selected for completeness as minimum number of null values existed within the set. The research was designed with the aid of a statistician from DAFWA and with feedback from Mr Ted Griffin.

The initial data analysis of the dataset was used to establish if the data is valid and that the profile data for the entire sample in a single location was consistent down the soil profile. In cases where there were two soil types e.g. sand over duplex and differences were observed the highest sample was taken. Based on this knowledge we used the full data set with each of the horizon points to conduct the analysis.

The equipment required to complete the research included both software and hardware. As the data had already been collected, the research only needed the platform for the basics of application of statistical and data mining techniques.

Software

Data formatting and storage software:

- Text based editor
- Excel 2003
- Access 2003

Statistical Excel software:

- Microsoft Excel 2003, with statistical add on package.

Data mining software:

- WEKA version: weka-3-5-3jre

Hardware

The hardware equipment required to complete the research includes the use of a Windows based platform that can operate the software required. This hardware is available to the researcher and has the following specifications:

- *CPU Speed:* 2.4 GHz
- *RAM:* 512 MB
- *Hard Drive:* 40 Gbytes
- *Input / output:* USB Thumb drive, CD burner, and Internet capacity.

As the hardware is above the capacity required to undertake the research it has allowed the data mining application to be run quickly and any analysis may be undertaken efficiently.

3.2.1 Statistical

Statistical processes were used to establish a benchmark for the analysis of the dataset from against which the effectiveness of data mining could be tested against. The process was conducted with a cluster analysis in Microsoft Excel software. The following steps are a general to guide to the application of the techniques for the experiments:

A) Data collection cleaning and checking

Relevant data was selected from a subset of the DAFWA soil science database.

B) Data formatting

The data was formatted into Microsoft Excel format from the MS Access database, based on ten soil types and relevant related fields. The data was then copied into a single Excel spread sheet. The Excel spread sheet (ESS) then underwent initial formatting to replace any null or missing values in the dataset to allow coding for the file in the next phase. See figure 5, step 2

C) Data coding

The dataset was then converted into individual datasets to allow analysis of the data. The statistical analysis was broken up into five stages; each stage provided analysis on a part of the dataset or a different technique.

Stage 1: Full Agdata set (all soils) – normal data

The analysis plotted the 20 traits for each location against its latitude and longitude. The data for each trait was placed into a new Excel spread sheet with its latitude and longitude and any location's missing values were removed by sorting the data. The latitude and longitude data was rounded to a single decimal place to overcome a data mapping problem. The data was then selected and placed in a pivot table with the latitude placed on the X direction and longitude placed on the Y direction. The trait data was then placed into the middle section of the pivot table and the whole table copied to allow graphing to be undertaken. The 20 individual traits sets were then converted to 3D surface graphs to allow them to be compared against each other and relationships and trends to be established. The 3D surface maps were named and are available in the appendix section (8.3.1).

Stage 2: Full Dataset (all soils) – standardized data

The second stage was conducted by standardizing the full set of 20 traits and repeating the processes in stage 1 on the new dataset. The pivot tables have a problem with the creation of the surface of the map. The problem was overcome by establishing a baseline and adding each value in the pivot table to the lower values of the data range. In addition, the lowest value was negative so a formula was required to correct this problem, into a positive integer value. The 3D surface maps were named and are available in the appendix section (8.3.2).

Stage 3: Full Dataset (all soils) –correlation table

Stage 3 was conducted to outline the relationships between each of the soil traits to all other traits. This was conducted by the creation of a correlation table that outlines the statistical relationship between each trait for the full soil dataset. The analysis was conducted using an add-on statistical package for Microsoft Excel 2003. The table outlines the traits with a strong correlation with, for example, CEC – CEC having a correlation of 1.00 and any relationship greater than 0.47 (+/-) having significant affect of the two traits values. The correlation table is available in the appendix section (8.3.3).

Stage 4: 3 Main soils - normal data

Stage 4 is based on analysis of stages 1-3 and it was found by DAFWA researchers that further analysis of the three main soil types in the dataset was needed to establish clusters on the main soil types. The three main soils are grey deep sandy duplex, loamy gravel and pale deep sand and the first three stages of analyses process repeated on the new data sets. The data was divided into traits and combined with rounded latitude and longitude data for each location. The data was then used to create a pivot table and used to create a 3D surface graph that was used for analysis. The 3D surface maps were named and are available in the appendix section (8.3.4).

Stage 5: 3 Main soils - Standardized data

The process conducted in stage five used the same guidelines as the stage 2 analysis. The 3D surface maps were named and are available in the appendix section (8.3.5).

D) Analysis and review of outcomes

The methods of comparison and review of the experiments was conducted using the methods referred to in figure 5, 6 and 7.

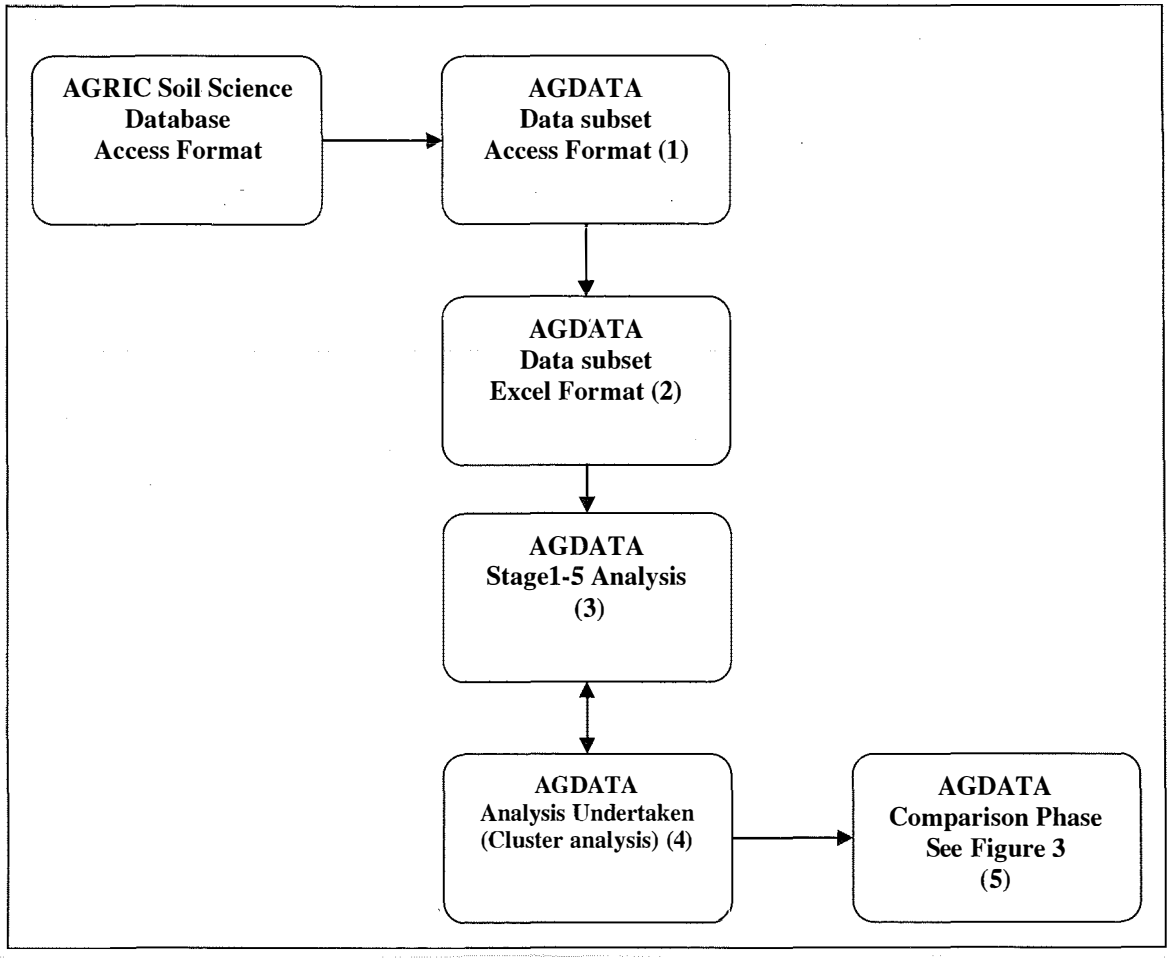


Figure 5: Experimental technique: Statistical processes.

3.2.2 Data mining

The data mining process was conducted in accordance with the results of the statistical analysis. The following steps are a general outline of the procedure that allowed a cluster analysis to be conducted on the dataset:

A) Data collection cleaning and checking

Relevant data was selected from a subset of the DAFWA soil science database.

B) Data formatting

The data was formatted into an Excel format from the Access database, based on the ten soil types and relevant related fields. The data was then copied into a single Excel spread sheet. The Excel spread sheet (ESS) was then formatted to replace any null or missing values in the agdata set to allow coding for the file in the next phase. See figure 6, step 1.

C) Data coding

The agdata set was then converted into a comma delimited (CSV) format file for the ESS. This file was then saved and opened using a text editor. The text editor was used to format and code the data into the type that will allow the data mining techniques and programs to be applied to it. The coding was formatted so that the input will recognize names of the attributes, the type of value of each attribute and the range of all attributes. Coding was then conducted to allow the machine learning algorithms to be applied to the agdata set to provide relevant outcomes that are required of the project See figure 6, step 2. The data coding attributes were named in line with the data table (see table 2, section 8.2, p, 80).

D) Stage 6 - five case studies

The Agdata was then broken down to five profiles;

1. One soil – one trait.
2. One soil – two traits.
3. Two soils – one trait.
4. Two soils – two traits.
5. All soils – all traits.

Grey deep sandy duplex and loamy gravel were used as these soils and their traits contained the maximum number of values within the data set. The traits that contained the highest number of values were clay and EC and these were used in the first four stages with clay used in single trait instances. The sub data set were then applied to the expectation-maximization (EM) algorithm and FarthestFirst algorithm. The clustering data was then collected including means and standard deviation to determine algorithm accuracy against actual values. See figure 6, step 3.

E) Analysis and review of outcomes

Comparison and review of the experiments were conducted according to the methods referred to in figures 5 and 7. See figure 6, step 4, below.

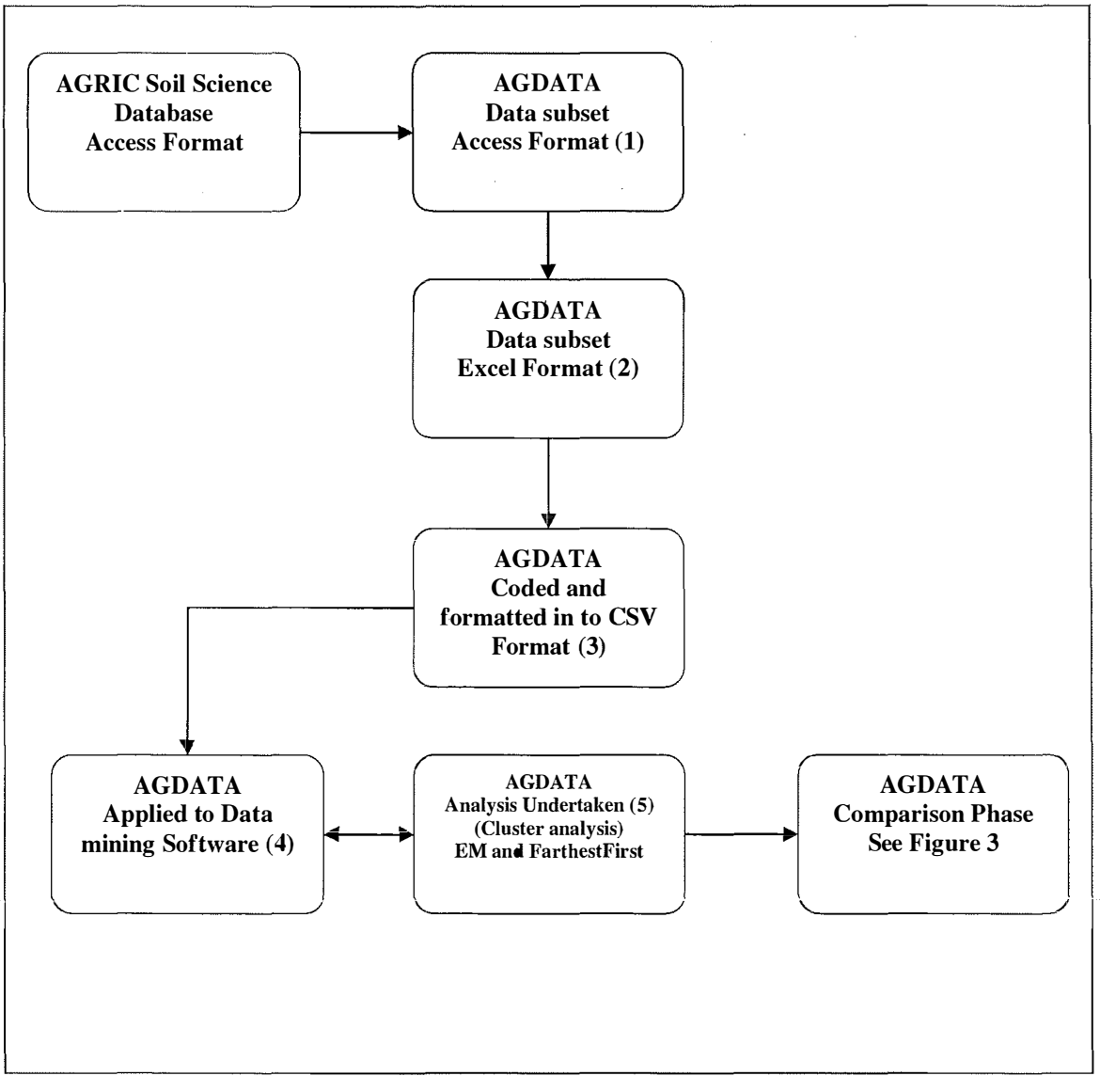


Figure 6: Experimental technique: Data mining

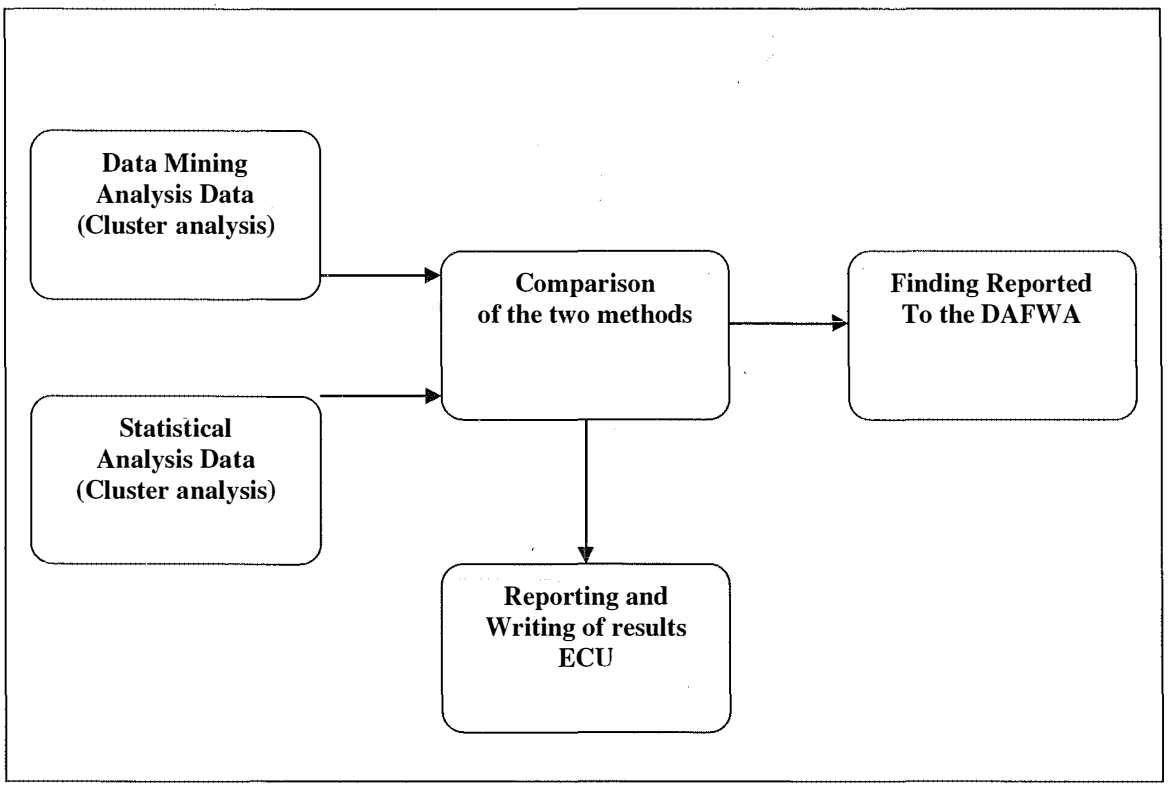


Figure 7: Data mining vs. traditional statistical methods.

The research has a number of limitations that could impact on the results achieved. These were outlined by Palace (1996) who stated that the applications of data mining techniques improve as the data set size increases. The agdata set has over 2800 sets of data with an average 12 measurements per geographical location. The size and type of the dataset was a major limitation, this was predicted by Mr Ted Griffin. The accuracy of any analysis technique increases with the amount of data contained in the dataset. This is due to the patterns that are contained with the data. More patterns point help define a stronger relationship, thus more reliable results. The size of the data set analysed was limited because of the short time available for the research and the ability of the human interpretation of the outcomes because of the complexity of the dataset used.

3.4 Data Analysis

The research adopted action research methodology, where improvement and changes may have to be undertaken to provide the DAFWA with outcomes that meets their required specifications for the project.

The research used Excel software to conduct qualitative analyses and to create a benchmark for the analysis of the dataset. The benchmark allowed current statistical methods for the dataset to be established and any limitations to be identified. The dataset was then analyzed using a clustering process within the data mining software. The results were then compared against the benchmark for a number of factors that included ease of application, speed, time and accuracy of results to determine if data mining was superior to current methods. The results of statistical and data mining experiments may still require expert analysis to be understood and used.

4 Results

The analysis and interpretation of patterns is a time consuming process that requires a deep understanding of statistics. The process requires a large amount of time to complete and expert analysis to examine any patterns and relationships within the data.

4.1 Statistical results

The research activities involve a process to establish if patterns can be found in the data. These processes involve the statistical manipulation of the data set in Excel. The aim of the research was to determine if a relationship or correlation can be established with soil trait data. The process involved the creation of analysis tools and charting the data so that longitude and latitude data and trait data is displayed and experts can interpret the findings.

The initial statistical data analyses involved four processes:

1. Raw data traits plotted against longitude and latitude for each sample location using a 3D surface map.
2. Standardized traits plotted against longitude and latitude with the data leveled using the minimum trait value to level the data for plotting in the 3D surface map.
3. Correlation table analysis.
4. Regression correlation analysis.

The process of plotting data required expert analysis for a relationship to be established. Such analysis was conducted in conjunction with Mr. E.A Griffin, Soil Scientist for the Department of Agriculture and Food, Western Australia.

The dataset was constructed from the DAFWA soil science data was designed to collect repetitive samples of the data contained in the south western agricultural region. The total data set contained 493 sites with an average of 5 samples taken for each location with a total of 2841 sample sets taken. The samples were analysed for a possible 41 traits but very few sets were complete for all data. The total number of data points possible was 116,481 but due to missing values the total number of data points considered was 34881.

Stage 1: Initial raw data analysis

1. Collect the raw data into a single Excel spread sheet, the addition of the elevation trait for each sample.
2. The creation of a new sheet for each of the soil traits to allow plotting against longitude and latitude.
3. For each sample location that did not have longitude / latitude data was removed from the data set and the soil classification plotted on to a chart (see appendix – section 8.3.1).
4. The creation of a 3D surface map requiring the latitude and longitude to be rounded to a single decimal place prior to the data insertion into a pivot table. The data was rounded by using the following formula:
    a. **(=Round (Round number, number of decimal places))**
5. The rounded longitude and latitude was then inserted into a new spread sheet with the individual traits e.g. (Lo-Lat-CaC03) for CaC03 and the samples with missing longitude and latitude data were removed.
6. The longitude – latitude – trait data was then selected and a pivot point table created in a new sheet e.g. (Lo-La-Cac03 (PP-G)) where the '*PP*' stands for pivot point table and '*G*' stands for graph.

Example:

Longitude	114.2	114.3	114.4	114.5
Latitude		Trait	Trait	Trait
-35	1.333334	Trait	Trait	Trait
-34.9	Trait	Trait	Trait	Trait
-34.8	Trait	Trait	Trait	Trait

7. The pivot point table was then formatted, by removing the column and row totals and then the count was changed into an average so that the fields were representative of the data.
8. The pivot point table was then copied and all data was formatted to two decimal places to allow ease of analysis.

Example:

Longitude	114.2	114.3	114.4	114.5
Latitude		Trait	Trait	Trait
-35	1.33	Trait	Trait	Trait
-34.9	Trait	Trait	Trait	Trait
-34.8	Trait	Trait	Trait	Trait

9. The formatted data was then graphed using a 3D surface map with longitude (X), latitude (Y) and trait values (Z), an example of this is shown below in figure 8.
10. The process was repeated for traits (CaC03 – ExKP).

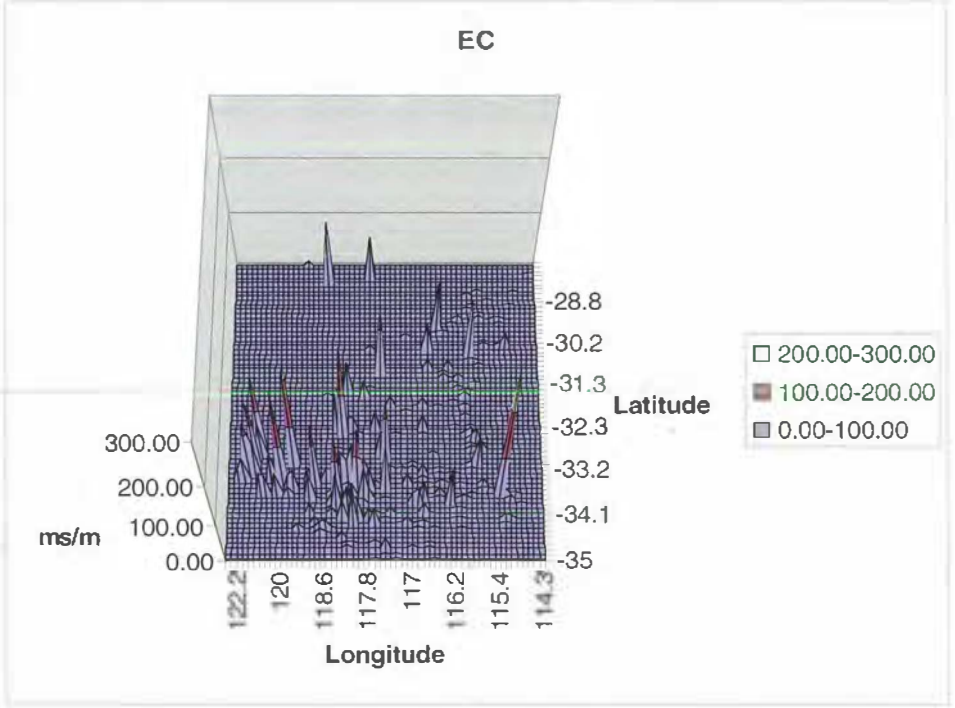


Figure 8: Normal data, EC all soils

Figure 8 displays the levels of Electrical Conductivity for the soils in the south west; the graph shows that there are high levels in the south eastern corner of the state.

Stage 2: Standardized data

1. Creation of a new Excel spread sheet (Stage 2 Research.xls).
2. Copy the new initial data spread sheet (Full dataset.xls) into next stage of analysis process.
3. Copy longitude and latitude for Stage 2 research file into full dataset file.
4. Round longitude and latitude to 1 decimal place, using (=Round (Row number, number of decimal places)).

   Standardize all trait data using the following formula

   =IF(Data.V2="";" ";
   STANDARDIZE(Data.V2;AVERAGE(Data.V$2:V$2842);STDEV(Data.V$2:V$2842)))
5. The data was then isolated as per stage 1 and place into a worksheet for each trait, sheet name given by trait name e.g. CaC03
6. Creation of a second set of pivot point tables for each trait, longitude and latitude and placed into a new worksheet (e.g. CaC03 (PP-G)).
7. The pivot point table then had the row and column totals removed and the fields changed from count to average.
8. The pivot point table was then copied to obtain the raw data and the table formatted to two decimal places.
9. The minimum data value is obtained using the following formula =MIN(B74:BH136)
10. The minimum value is used to establish a baseline, with the lowest value of the trait added to trait dataset, using the following formula =A138+B74
11. Longitude and latitude data was then set out in a new table as per the other tables, with the above formula filling the trait section of the table.
12. The data was then graphed using the 3D surface charts, latitude (X), longitude (Y) and trait value (Z). The process of adding the individual standardized trait values to the minimum value displayed the data with a plan of minimum values with peaks on both sides of the plan. An example of this is shown below in figure 9.

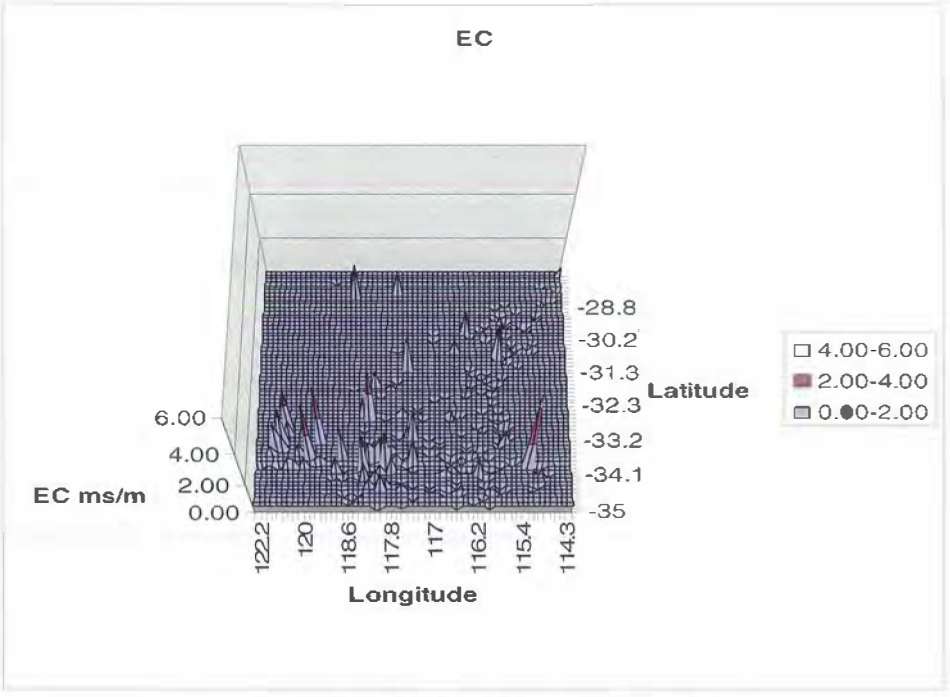


Figure 9: Standardized data, EC all soils

Figure 9 displays the level of Electrical Conductivity throughout the research area. The data shows that there is a high concentration on the coast, located in the Margaret River region of the state. The data shows that there is a comparison between the normal (figure 8) and standardized data graphs and that there could be correlation.

Stage 7: Correlation table

The next stage in the research involved the creation of a correlation table from the basic dataset. The correlation table outlines the relative relationships between all traits in the dataset. The correlation table showed that there were a number of strong relationships between the traits; these results can be seen in appendix 8.3.3.

The current method was then applied to a subset of the agdata set; the subset contained the three main soil types that had the greatest number of geographical locations in the survey area. The three main soil types were: grey deep sand duplex, loamy gravel and pale deep sand. The analysis process was repeated because of a request from researchers at the DAFWA. The process was repeated with the exception of the creation of the correlation table, as that was not required with the limited amount of data. The overall process was very time-consuming and repetitive, with the research requiring seven days to complete the whole process with a large amount of help and feedback from the statistician. The research, in addition to being time consuming and complex, required a large amount of human input and interaction to complete the process. The process was designed with the aid of an agricultural statistician from the DAFWA to ensure that the analysis is true to the process presently used at the DAFWA.

4.2 Data mining results

The benchmark having been established, the data analysis was then replicated using WEKA data mining software to determine if any advantage could be gained in both time saving and interpretation of the agdata set. The application of the data to WEKA required that some preprocessing be undertaken. The dataset produced in Excel for the statistical processes were copied and then converted to .CSV file format to allow them to be applied to WEKA. The .CSV file extension allowed initial analysis to be conducted, with later conversion to be taken in to an ARFF WEKA data file for the experimental outcome to be saved.

The data mining platform allows a number of data interpretations including classify, cluster, and associate routines to be conducted after the preprocessing stage. The agdata

set did not require any filtering because of the limited amount of missing values and the outcomes required by the researchers. The initial screen provided a set of information that is required by the researchers and took a large amount of time to complete with the current statistical methods.

The full agdata set was applied to the EM-1 100-N-1-5 100-M 1.0E6 and FarthestFirst clustering algorithm to see if any patterns could be established with the model being constructed using a training model to build the associations in the data. The clustering algorithms outlined above perform their operations in two different ways, and the differences between the two will be used to determine the accuracy when compared with each other. The expectation-maximization (EM) algorithm was outlined by the WEKA data mining software and provides a basic outline of the algorithms operation. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify how many clusters to generate.

The cross validation performed to determine the number of clusters is done in the following steps:

1. The number of clusters is set to 1
2. The training set is split randomly into 10 folds.
3. EM is performed 10 times using the 10 folds the usual CV way.
4. The log likelihood is averaged over all 10 results.
5. If log likelihood has increased the number of clusters is increased by 1 and the program continues at step 2.

The number of folds is fixed to 10, as long as the number of instances in the training set is not smaller 10. If this is the case the number of folds is set equal to the number of instances. The EM algorithm required that some of the parameters be changed to ensure it produced the same amount of clusters as the FarthestFirst to allow analysis to be undertaken. This was simply done by outlining the number of outputs in the options before applying EM to the individual dataset.

The WEKA data mining software outlines the operation of the FarthestFirst algorithm and clusters using the farthest first traversal algorithm. It works as a fast simple approximate cluster and is modeled after SimpleKMeans. The stages of analysis were conducted using both algorithms and the data collected and formatted below. The comparison of the accuracy of each clustering method was determined by analysis of the grouping numbers, means and standard deviation.

The results of the experiments are shown in the section below with each case study having two tables. The first table displays the algorithm name in the first column and outlines the number of clusters created in the second column, with the number of occurrences. The second table displays the clusters with their means and StdDev created by both of the clustering methods. Each case study also includes data about the actual number of data points in the data set; this was included to determine the accuracy of each algorithm. The first four case studies were designed to limit the amount of data input so that the results were simple to understand. The last case study included all data from the original agdata set for completeness.

The data from each of the two algorithms varied, with EM producing more data than FarthestFirst, and so the last case study contains four large tables of means and StdDev because of the seven clusters created. The analysis of the output involved looking at the cluster percentage and the number allocated to each cluster in comparison with the actual dataset. Case study five, because of the large amount of information produced by the experiment, will not be analysed and the results and discussion will focus of the first four case studies for comparison to the current statistical methods.

The results of the data mining experiments have shown that FarthestFirst algorithm was equal to or more accurate at clustering compared to the EM algorithm in all four tests. The results showed that the FarthestFirst results for case study 2 and 3 grouped the data correctly with a minimum of error. The accuracy of the EM algorithm much less when compared with the actual data points and only grouped the data correctly in case study 3, with the other three case studies clustering much less accurately.

Case study 1: (One soil type – 1 trait - grey deep sand duplex - Clay)

		Trait 1
EM	0	268 (50%)
	1	268 (50%)
FarthestFirst	**0**	**417 (78%)**
	1	**119 (22%)**

	Trait 1
EM	0 Mean = 3.9336 StdDev = 1.745
	1 Mean = 34.6761 StdDev = 15.4472
FarthestFirst	0 centroid = 1.0
	1 centroid = 75.0

Clay = 536 instances

Case study one compared the clay trait for all of grey deep sand duplex soil and it was found that the EM algorithm grouped the two clusters in half. The FarthestFirst algorithm weighted the groups more to cluster 0 and that reflected the actual data with more accuracy with a single group of 536 instances.

Case study 2: (One soil type – 2 traits - Grey deep sand duplex – Clay and EC)

		Trait 1
EM	0	610 (60%)
	1	411 (40%)
FarthestFirst	**0**	**541 (53%)**
	1	**480 (47%)**

	Trait 1
EM	0 Mean = 4.052 StdDev = 2.417
	1 Mean = 37.223 StdDev = 30.9041
FarthestFirst	0 centroid = 35.7 Clay
	1 centroid = 350.0 EC

Clay = 536 instances

EC = 480 instances

Case study two compared grey deep sand duplex with traits clay and EC and it was found that EM grouped the two clusters more accurately than case study one. EM grouped the instances more in line with the actual data but it was found that FarthestFirst still was more accurate. FarthestFirst grouped cluster 1 correctly with EC having 480 instances, but there was still some error with clay instances grouped in cluster 0 with 541 instances and actual data having only 536.

Case study 3: (Two soil types – one trait - Grey deep sand duplex, Loamy gravel (CLAY)

		Trait 1
EM	**0**	**536 (47%)**
	1	**612 (53%)**
FarthestFirst	**0**	**612 (53%)**
	1	**536 (47%)**

	Trait 1
EM	0 Mean = 19.7913 StdDev = 18.9894
	1 Mean = 21.7818 StdDev = 15.3254
FarthestFirst	0 centroid = 47.35 Loamy gravel clay
	1 centroid = 0.0 Grey deep sandy duplex clay

Grey deep sandy duplex (Clay = 536 instances)

Loamy gravel (Clay = 612 instances)

Case study three compared two soil types (grey deep sand duplex and loamy gravel) with one trait (clay). Both the algorithms group the instances correctly with grey deep sand duplex and having 536 instances of clay, and loamy gravel having 612 instances of clay.

Case study 4: (Two soils types; Grey deep sand duplex, Loamy gravel – two traits – (Clay, EC))

		Trait 1
EM	0	34 (2%)
	1	582 (27%)
	2	991 (46%)
	3	562 (26%)
FarthestFirst	**0**	**536 (25%)**
	1	**480 (22%)**
	2	**541 (25%)**
	3	**612 (28%)**

	Trait 1
EM	0 Mean = 101.6457 StdDev = 58.0094
	1 Mean = 34.9632 StdDev = 14.236
	2 Mean = 2.916 StdDev = 1.744
	3 Mean = 10.3491 StdDev = 4.6035
FarthestFirst	0 centroid = 52.0 Grey deep sandy duplex clay
	1 centroid = 350.0 Grey deep sandy duplex EC
	2 centroid = 146.0 Loamy gravel EC
	3 centroid = 0.3 Loamy gravel clay

Grey deep sandy duplex (Clay = 536 instances)

Loamy gravel (Clay = 611 instances)

Grey deep sandy duplex (EC = 479 instances)

Loamy gravel (EC = 540 instances)

Case study four compared two soil types (grey deep sand duplex and loamy gravel) and two traits (Clay and EC). The results show that EM grouped the instances weighted towards cluster 2 at 991 instances and weighted cluster 0 less, having only 34 instances. FarthestFirst was more accurate and only mis-classified one instance out in clusters 1, 2 and 3.

Case study 5: (ALL – DATA)

		Trait 1
EM	0	342 (12%)
	1	340 (12%)
	2	95 (3%)
	3	420 (15%)
	4	464 (17%)
	5	323 (12%)
	6	320 (11%)
	7	491 (18%)
FarthestFirst	0	1154 (41%)
	1	126 (5%)
	2	779 (28%)
	3	337 (12%)
	4	261 (9%)
	5	69 (2%)
	6	69 (2%)

The data was too complex to conduct analysis in the limited time available and was done for completeness of the research on advice from the statistician. Note that table two outlines all the seven clusters created by the two algorithms and the instance allocation for each. The full set of means and StdDev are located in appendix 8.4.2 for case study five.

4.3 Results overview

Two analysis processes both provided a method by which soils data analysis can take place. The results of the experiment showed that clustering may be an effective tool for the comparison of soil types, traits and locations within the study area of the south west agricultural region of Western Australia. The application of statistical methods required a large amount of time to complete and to produce a usable outcome for soil researchers. Data mining required less time and knowledge to complete an analysis process, compared with current statistical processes, but the interpretation of the data still requires expert human analysis from the DAFWA.

5 Discussion

The collection of information and data has increased with the advent of new computing technology, but establishing patterns within this data has become more difficult and requires new approaches and tools if it is to be undertaken. The advent of this problem has provided an opportunity from which data analysis has started to take over from current methods. The advent of this technology has reduced the time taken to undertake data analysis and has increased automation of the process. The research undertaken showed that data mining has advantages and can be easily applied to the soil data set to establish patterns in the data. The application of the WEKA data mining platform provided an easy and quick method for the cluster analysis. The platform provides a number of clustering algorithms that can be used for different tasks. The experiments conducted used two clustering algorithms, EM and FarthestFirst to determine the most accurate when compared with actual results.

The integrity of the data is critical to ensure that results are not affected by outliers and null values in the data set, or other adverse factors. The establishment of clusters in the data required a large amount of human time and input time when using current methods. The current methods still required some post Excel analysis because the platform is limited in the interpretation of the graphs generated. The application of the same clustering techniques using the data mining software reduced the time taken to process the data sets, with the amount of time reduced to one day, and also allowed a greater amount of knowledge to be gained from the data.

5.1 Evaluation of statistical methods

Current statistical methods provided a platform from which analysis of agricultural soil profiles can be undertaken. The application of these methods has proven to provide accurate analysis of clusters and patterns when used on soil science databases. The current technique involves an in-depth understanding of statistics and requires a large amount of human input and time to complete. The current statistical methods that are being used to determine valid patterns and soil profiles clusters are 3D surface mapping and basic statistical methods including correlation tables and distribution analysis.

Increases in the amount of data collected from field experiments have meant that the time and complexity has increased to the point that it has become difficult to obtain new knowledge. The experiments undertaken during the research required expert input and provided a large amount of graphical data; see appendix 8.3 statistical analysis. The processes used in the experiments provided a benchmark for the comparison of the two data mining algorithms and are still useful in conducting basic soil analysis. The three statistical methods that provided 3D surface maps came with a formatting problem. The graphs where also produced in reverse to the actual physical location due the reverse of longitude and latitude data. The problem was that not all graphics where produced in the same scale and the analysis techniques required the graphs to be printed for physical comparison. This problem was a major draw back when submitted to DAFWA researchers and they required extra time and effort to compare traits. The creation of a correlation table for full normalized data sets outlined the significant relations that exist between the traits. This method of analysis provided a quick over view of the data and allowed DAFWA researchers to conduct further research into the relationships.

5.2 Evaluation of data mining

The application of data mining techniques has proven to be almost as accurate as standard statistical analysis techniques and with the increase in the number of instances this is projected to increase in accuracy. The WEKA data mining software provided a simple platform from which to undertake the research and comparison of the data set. The research investigated the accuracy of two methods that were provided as standard WEKA data mining software and compared their clustering accuracy. The input of the data into data mining applications proved to be simple with the conversion of an Excel spread sheet into a CSV file and then an ARRF file. The two algorithms used were also compared for ease of use and time taken to complete the analysis process for each of the five case studies. The EM algorithm required a larger amount of input to setup for each clustering operation and required that the number of clusters set. Processing time taken to complete each case study was also significant when compared with FarthestFirst. The analysis of the results showed that EM only correctly verified the dataset on a single instance. The two soils - two traits was the case study that EM was as accurate as FarthestFirst, both grouped the two clusters the same with cluster 0 at 536 instances and cluster 1 at 612 instances. The reason for this is unknown and both were one hundred per cent accurate on the actual data. The reason behind this requires further investigation but owing to time constraints this will not be conducted. Case study one was the major outlier with a high rate of mis-classification for both instances, with EM splitting the instances into even groups of 268 for cluster 0 and 1. FarthestFirst grouped the instances more accurately than EM but still mis-classified 199 instances or 22 per cent of the dataset.

The FarthestFirst algorithm provided a much more accurate tool for the verification of valid patterns and profile clusters when tested against the benchmark, with most cluster groups within four instances of the actual data.

The FarthestFirst algorithm proved to be much simpler to use and required less processing time to complete each case study. The FarthestFirst algorithm, when applied to data sets, can establish valid patterns and soil profile clusters. The FarthestFirst

algorithm was the most efficient technique in determining patterns and clusters when compared to standard statistical analysis techniques.

The data mining application also provided a number of functions, such as visualize, where all traits were charted against each other and allowed for a quick analysis. This function was validated with the charting of the soil locations with the longitude and latitude data reflecting the initial data analysis graph created: see appendix 8.3.1.

5.3 Comparison between methods

The two methods of soil analysis had advantages and disadvantages, with both providing accurate clustering of the experimental dataset. The accuracy of the data mining clusters method on the agricultural soils was dependant on the selection of the correct algorithm, and was shown to have a wide grouping within the two algorithms researched. The two methods researched showed that data mining can equal the verification of valid patterns when compared to standard analysis techniques. Analysis and classification of soil traits under the current system is very subjective with groups open to human interpretation. This human input means that clustering of similar soil types and traits can become less accurate and this can have an affect of the accuracy of analysis and knowledge gathering. The advantage found in the application of data mining techniques is that human interpretation is reduced and the data is clustered based on the actual information without bias. Although data mining has a number of advantages over the current statistical methods, the WEKA software and process still has a number of problems. The research encountered a number of disadvantages that included selection of the current algorithm and the graph output being only in 2D with no provision for placing a third set of data on the visual display. The application of both methods still requires knowledge of the results required to allow selection of the correct techniques to provide new knowledge.

The literature reviewed outlined that data mining accuracy increased with the amount of data contained in dataset. This was a factor with the dataset used in the research by Berry and Linoff (1997) stated that making sense of complex issues is naturally approached by breaking the subject into smaller segments that can be explained more simply. This was

the focus of the research where complex relationships were broken down to simple segments that could be understood. The design of the first four case studies utilized only a small portion of the data but tested the accuracy of the two clustering algorithms.

Comparison of the two methods has shown that data mining is still not one hundred per cent accurate on all applications, when compared with standard statistical methods, but has shown to have greater benefits. The benefits of data mining include speed and increased levels of automation, but still do not provide all the analytical tools required for analysis of an agricultural soil database.

5.4 Issues related to research

During the course of conducting this research project there where a number of problems that had to be overcome. These problems included the application of data mining techniques, quantity of data, tools including WEKA and Excel, skills required to undertake the research, limited time and interpretation of results.

The application of data mining techniques required a deep understanding of the process involved and required that a large amount of background research be undertaken before commencement of the research. The quantity of data is a problem when it is applied to data mining and statistical research with the data set size having a direct correlation with accuracy of outcomes. The problem of a small dataset was overcome by creating small subsets of known values to remove data size as a variable in the research process and to focus on the methods applied to that data. The tools used in the research included WEKA and Excel and were very effective for conducting this research. The tools were very complex to use at a higher level and this problem was overcome with the help of experts in the field and by background research.

The interpretation of the research results was a complex process and did not focus on the relationships between the soils traits, but on the establishment of this relationship and their accuracy. The accuracy of each method and each algorithm was the corner stone of the research project, with the methods used to determine these being critical to the research outcomes. The problem of establishing the accuracy of method was overcome

by analysis of the grouping and numbers of instances in clusters. The comparison was then made to the actual number of instances in the clusters and the level of mis-classification. The analysis of the relationships within the data traits, as it applies to soil science, is outside the scope of this project will be undertaken by DAFWA researchers.

Due to the limited amount of time available to conduct the research only limited experiments were conducted in the effectiveness of clustering algorithms. The limited amount of data did not allow use of the full range of tools available, within the WEKA software, to be tested. Future research could be conducted to build on this research and analyse more of the functions and algorithms available in the WEKA data mining software. All the problems encountered during the research were overcome with the aid of experts in given fields and with perseverance.

6 Conclusion

The experiments conducted analysed a small number of traits contained within the dataset to determine their effectiveness when compared with standard statistical techniques. The agriculture soil profiles that were used in this research were selected for completeness and for ease of application to data mining. The soil original dataset was almost compete but still contained some missing values that had to be removed in a text editor because of the affect on the clustering process.

Standard statistical analysis was used to establish a benchmark with pivot tables created using normal and standardized data. Pivot tables were then used to create a 3D surface map, charting traits against longitude and latitude. Statistical techniques take soil types that have been classified that are the same to produce means and estimates to determine the properties of that soil type. Data mining clustering methods do not use that same assumption, but, rather than using these predefined classifications, assigns instances based on their values to provide an objective method of classification.

The five case studies were designed to test the concept and methodology of data mining and to establish the accuracy of EM and FarthestFirst clustering algorithms. The two algorithms were selected because of these different methods of grouping the data into clusters. This process was done to allow another level of comparison in the research. The research outcome found that FarthestFirst algorithm grouped instances more accurately that the EM algorithm, when compared with the statistical benchmark. The results showed that FarthestFirst had a lower mis-classification rate, and classified case study two correctly, with limited error in case studies three and four.

The accuracy of data mining depends on the amount of data used to create clusters, with the literature indicating that an increase in dataset size improves accuracy. Further research would look at increased dataset size to determine if this would increase the instance classification. This would create more focus on data mining and less on current statistical methods. There were a number of areas not explored by the research due to time limitations, such as the differences between the soil profile horizons within the same excavation site being of particular interest to DAFWA researchers.

The recommendations arising from this research are: That data mining techniques may be applied in the field of soil research in the future as they will provide research tools for the comparison of large amounts of data. Data mining techniques, when applied to an agricultural soil profile, may improve the verification of valid patterns and profile clusters when compared to standard statistical analysis techniques. The results of this research were passed on to the DAFWA researchers so they can determine if the application of data mining techniques may aid in their current and future soils research.

7 References

Bentley, T. (1997). Mining of information. *Management Accounting*, 75(6), 56.

Berry, M. J. A., and Linoff, G. (1997). *Data Mining Techniques—For Marketing, Sales and Customer Support.* New York: John Wiley and Sons.

Brown, M., L, and Kros, J., F. (2003). Data mining and the impact of missing data. *Industrial Management Data Systems*, 103(8/9), 611.

Butler, M., Herlihy, P., and Keenan, P. B. (2004). Integrating information technology and operational research in the management of milk collection. *Journal of Food Engineering*, 70, 341–349.

Canillas, E., and Salokhe, V. (2001). A decision support system for compaction assessment in agricultural soils. *Soil and Tillage Research*, 64, 221 -230.

Clayden, J. (2005). *Lecture notes: action research.* (Available from Edith Cowan University, 2 Bradford Street, Mount Lawley WA 6050). Edith Cowan University.

Cunningham, S. J., and Holmes, G. (2005). *Developing innovative applications in agriculture using data mining.*

Delen, D., Walker, G., and Kadam, A. (2004). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34, 113-127.

High- Tech Dictionary. (2005). *Neural Network.* Retrieved May 10, 2006, from http://www.computeruser.com/resources/dictionary/definition.html?lookup=3356

Farrand, J. (2002). *WekaMetal.* Retrieved May 25, 2006, from http://www.cs.bris.ac.uk/~farrand/wekametal/

Frank, E., Hall, M., and Trigg, L. (2005). *Weka (Data mining software) (Version Weka 3-4-5) [Java].* Waikato, New Zealand.

Gertosio, C., and Dussauchoy, A. (2003). Knowledge discovery from industrial database. *Journal of Intelligent Manufacturing* (15), 29-37.

Goddard, S., Harms, S., Reichenbach, S., Tadesse, T., and Waltman, W. (2003). Geospatial decision support for drought risk management. *Association for Computing Machinery*, 46(1), 35 - 38.

Grabmeier, J., and Rudoloph, A. (2001). Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery,* 6, 303–360.

Griffin, M. T. (2005). *Data field descriptions.* Perth: Department of Agricultural and Food Western Australia.

Heuvelink, G., and Webster. (2001). Modeling soil variation; past present and future. *Geoderma*, 10, 269-301.

Ibrahim, R. S. (1999). *Data Mining of Machine Learning Performance Data.* Unpublished Master of Applied Science (Information Technology), RMIT University.

Isbell, R. F. (1996). *The Australian Soil Classification. Australian soil and land survey handbook.* (Vol. 4). Collingwood, Victoria, Australia: CSIRO Publishing.

Iverson, L. R., and Prasad, A. M. (1998). Predicting abundance of 80 tree species following climate change in Eastern United States. *Ecology Monograph*, 68, 465-485.

Kaul, M., Hill, R., and Walthall, C. (2004). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems* (85), 1-18.

Little, L. S., Edwards, D., and Porter, D. E. (1997). Kriging in estuaries: as the crow flies, or as the fish swims?. *Journal of Experimental Marine Biology and Ecology*, 213(1), 1-11.

McBratney, A., Odeh, I., Bishop, T., Dunbar, M., and Shatar, T. (2000). An overview of pedometric techniques of use in soil survey. *Geoderma*, 97(3-4), 293-327.

Mckenzie, N., and Ryan, P. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89(1-2), 67-94.

Moore, G. (2004). *Soil Guide (a handbook for understanding and managing agricultural soils).* Perth: Department of Agricultural and Food Western Australia.

Northcote, K. H. (1984). *A factual key for the recognition of Australian Soils.* Adelaide: Rellim Technical Publications.

Online dictionary. (2005). Retrieved May 10, 2005, from http://www.onelook.com/

Palace, B. (1996). *Data Mining: What is Data Mining?* Retrieved Aug 30, 2005, from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

Ragel, A., and Cre´milleux, B. (1999). MVC—a preprocessing method to deal with missing values. *Knowledge-Based Systems*, 12, 285–291.

Rajagopalan, B., and Krovi, R. (2002). Benchmarking data mining algorithms. *Journal of Database Management*, 13(1), 25-35.

Rudi Alberts, Euler, T., Fischer, S., Heinle, E., Hakenjos, D., Homburg, H., et al. (2005). *YALE - Yet another learning environment.* Retrieved 04/03/2006, 2006, from http://www-ai.cs.unidortmund.de/SOFTWARE/YALE/index.html

Ryu, T.-W., and Eick, C. F. (2004). A database clustering methodology and tool. *Information Sciences*, 171, 29-59.

Schoknecht, N. (2002). *Soil Groups of Western Australia (Technical Report).* Perth: Department of Agriculture.

Schoknecht, N., Tille, P., and Purdie, B. (2004). *Soil-Landscape mapping in south-western Australia (Technical Report).* Perth: Department of Agricultural.

Selvanathan, A., Selvanathan, S., Kellor, G., and Warrack, B. (2000). *Australian Business Statistics* (2nd ed.). Melbourne: Nelson Thomson Learning.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, 33, 111 –126.

Trafalis, T. B., Richman, M. B., White, A., and Santosa, B. (2002). Data mining techniques for improved WSR-88D rainfall estimation. *Computers and Industrial Engineering*, 43, 775–786.

Veiga, F. A. D. (1996). Structure discovery in medical databases: a conceptual clustering approach. *Artificial Intelligence in Medicine*, 8, 473-491.

Witten, I. H., and Frank, E. (1999). *Data Mining, practical machine learning tools and techniques with Java implementations*. Sydney: Morgan Kaufmann.

Witten, I. H., and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd. ed.). Sydney: Morgan Kaufmann Publishers.

Zhang, B., Valentine, I., and Kemp, P. (2004). Modelling the productivity of naturalised pasture in the North Island, New Zealand: a decision tree approach. *Ecological Modelling*, 186, 299-311.

8 Appendices

8.1 Soil zone map of research area.

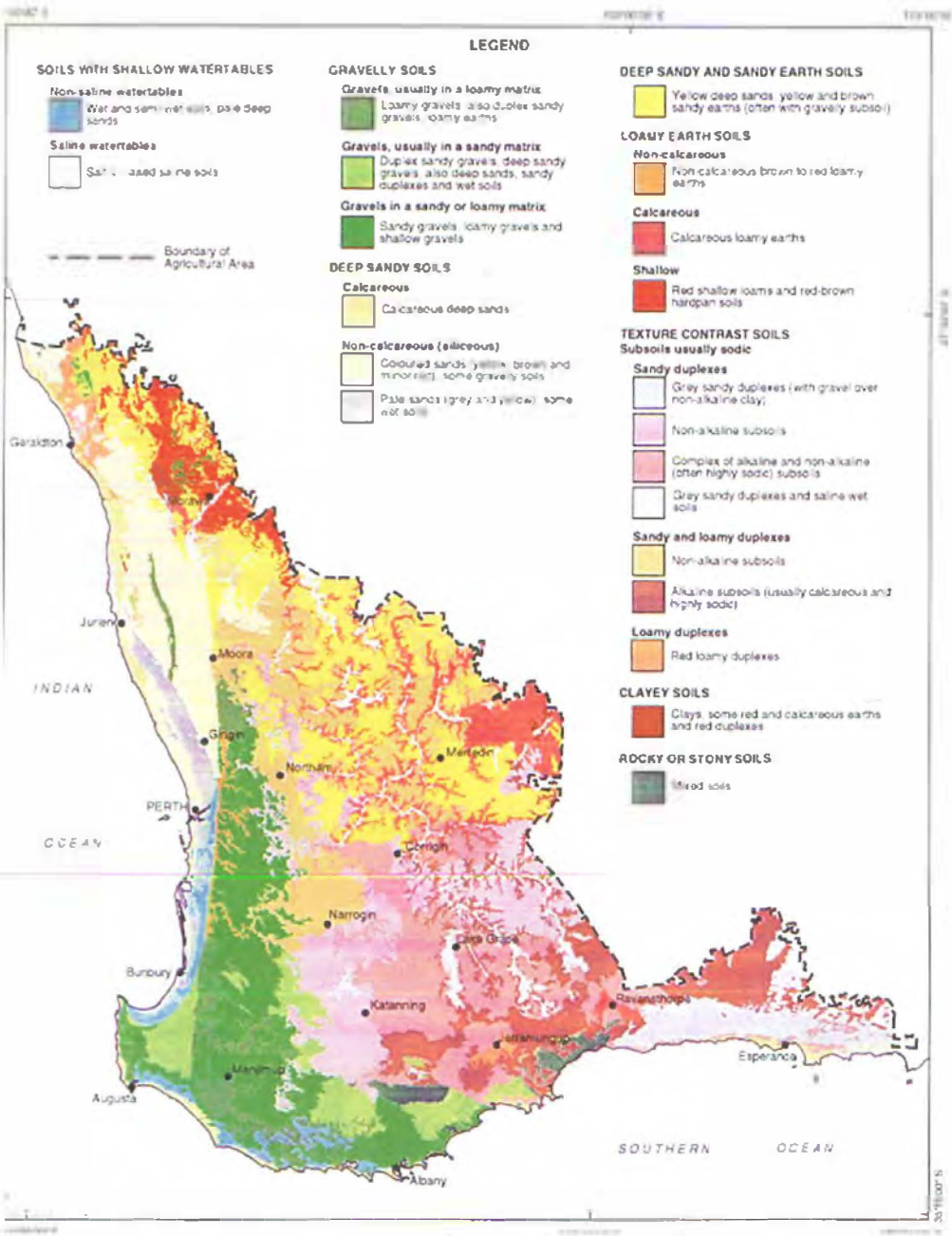


Figure 10: Characteristic of soils of south-western Australia. (Schoknecht, 2002, p.91)

8.2 Data field description

Table 2: Data field descriptions.

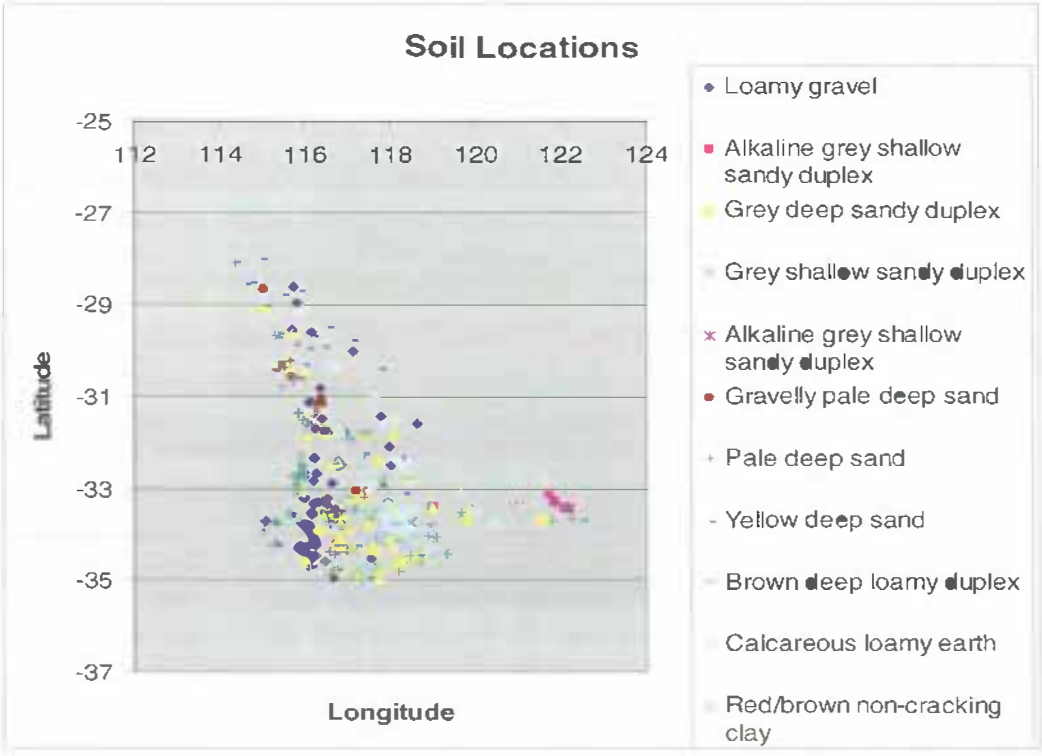
Field	Description
SOIL CLASSIFICATION	WA Soil Group code
MAP UNIT	Soil-landscape map unit (first three are zone)
AGENCY_CODE	Site's Agency code
PROJ_CODE	Sites Project code
S_ID	Site ID
O_ID	Observation ID (usually = 1 and largely redundant)
SAMP_ID	Sample ID unique identifier for sample taken within a site if null no sample taken
H_NO	Horizon (or layer) ID, from field morphology observations, sequence numbers may be missing
SAMP_H_MATCH	code indicating the degree of matching between the layer depth and sample depth A exact, B sample a subset of layer, C sample crosses layer but predominantly of layer, D other
SAMP_UPPER_DEPTH	sample upper depth
SAMP_LOWER_DEPTH	sample lower depth avDepth sample average depth
CACO3	CaCO3 %
CACO3_imp	HCl fizz test where from field observations N nil, S slight, M moderate, H high, V very high
OC	Organic Carbon %
PH	pH in CaCO3
Clay	clay %
EC	EC, ms/m
ExCA	Exchangeable Ca cmol(+)/kg
ExMG	Exchangeable Mg cmol(+)/kg
ExK	Exchangeable K cmol(+)/kg
ExNA	ExCEC CEC cmol(+)/kg
ExSUM	Sum Exchangeable Ca, Mg, K, Na cmol(+)/kg
ExESP	Exchangeable Na % of Sum
ExH	Exchangeable H cmol(+)/kg
ExMN	Exchangeable Mn cmol(+)/kg
ExAL	Exchangeable Al cmol(+)/kg
ExSAT_PC	Saturation % (100*Sum/CEC)
ExBASE	Base Status (100*Sum/clay)
ExCaP	Exchangeable Ca % of Sum
ExMgP	Exchangeable Mg % of Sum
ExKP	Exchangeable K % of Sum

(Griffin, 2005)

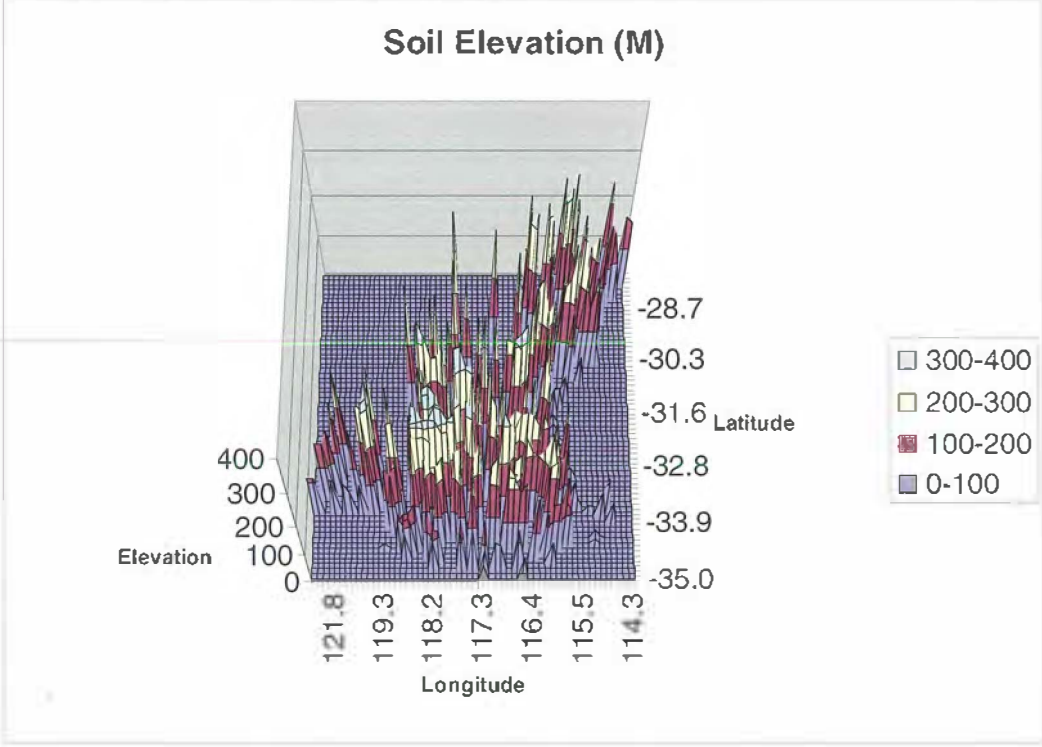
8.3 Statistical Analysis

8.3.1 Stage 1: Normal data – All soil types

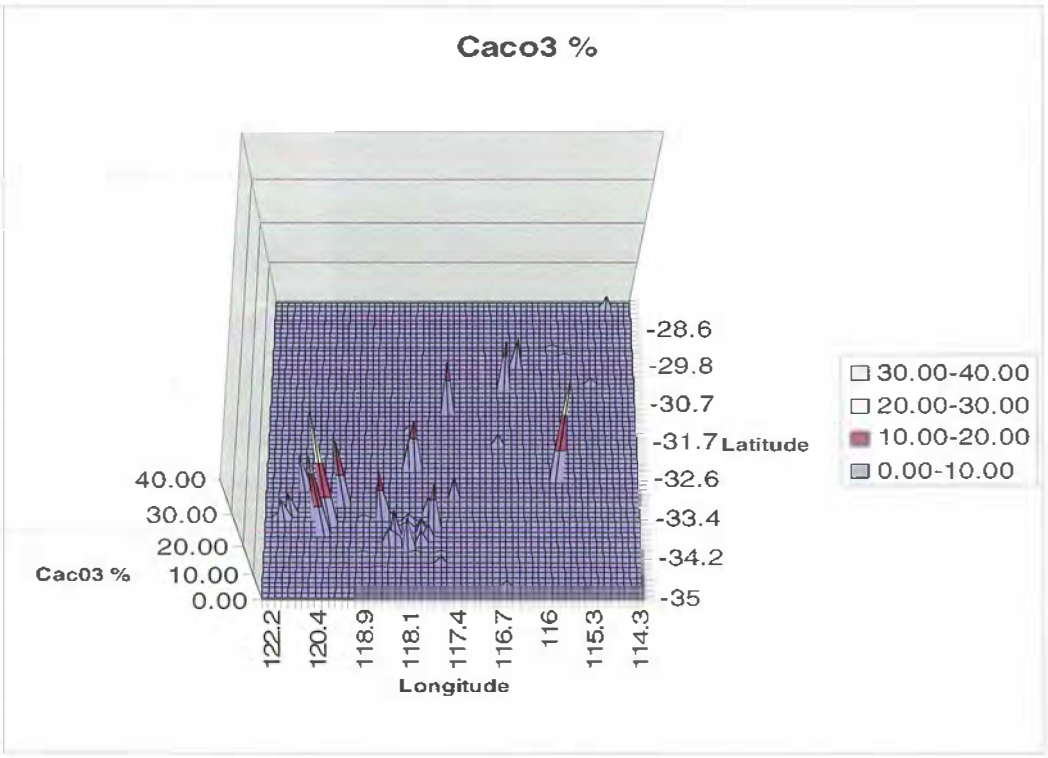
The graphs contained within this section are all 3D surface maps using the agdata full dataset. The graphic display the normal data as supplied by the DAFWA and each trait is graphed against it longitude and latitude. The values are averages for each location and show the traits distribution across the South Western agricultural region.



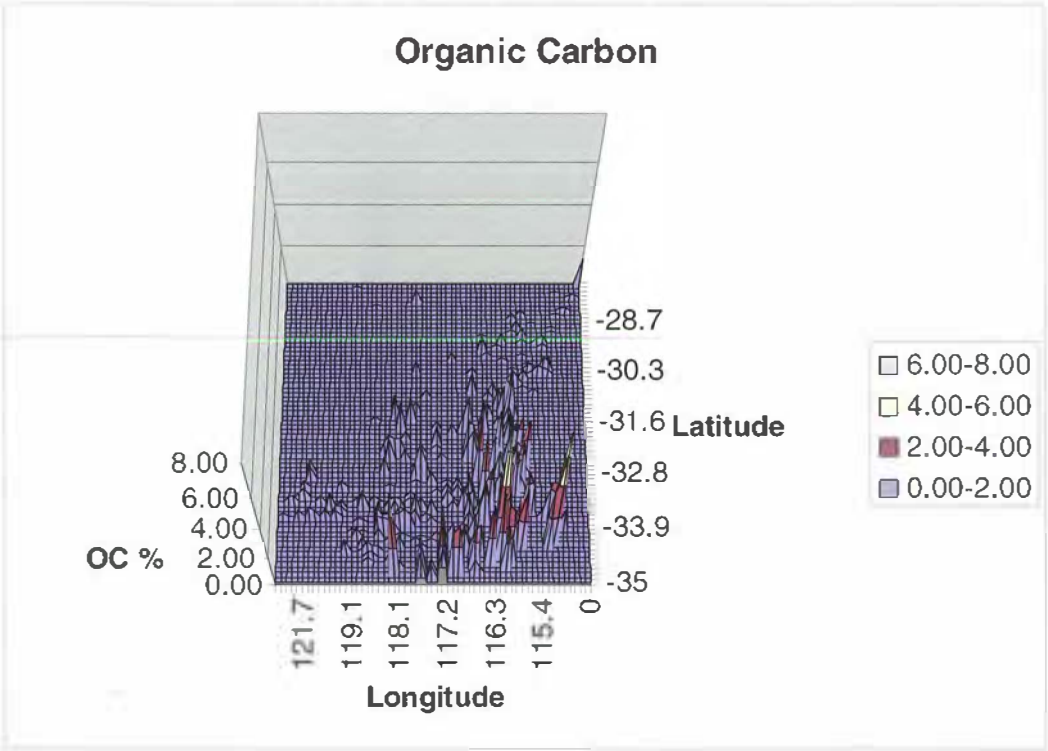
Initial analysis – Location of soils in the south west agricultural region.



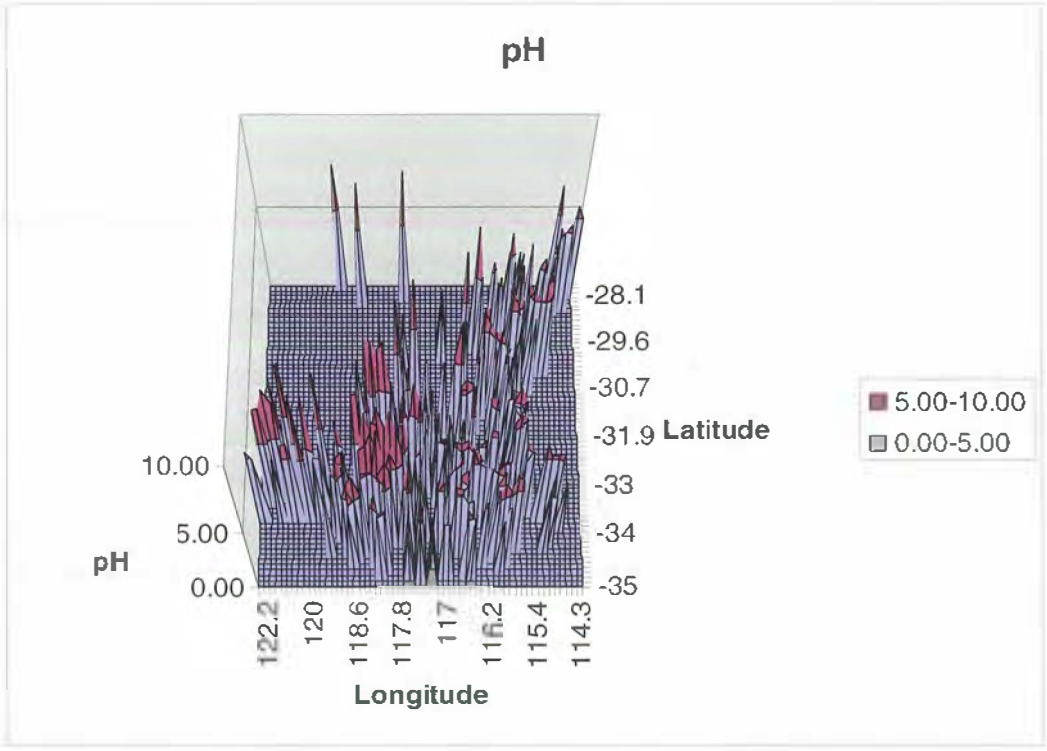
Elevation of each sample location.



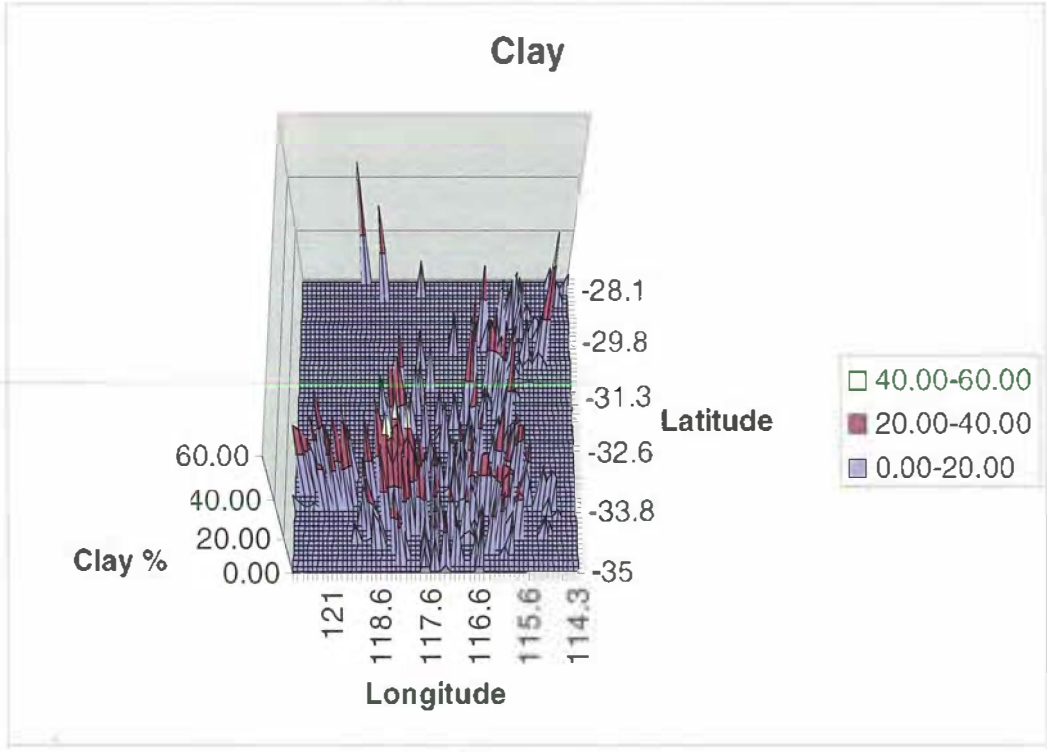
All soils (Normal data) – CAC03 %



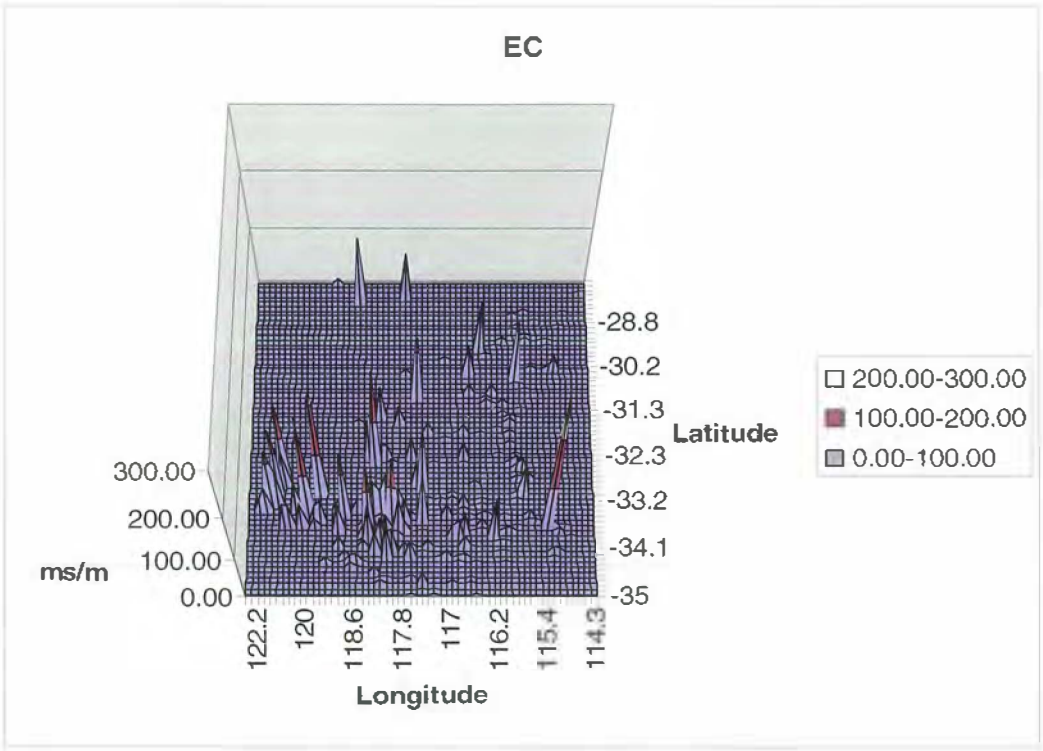
All soils (Normal data) – OC %



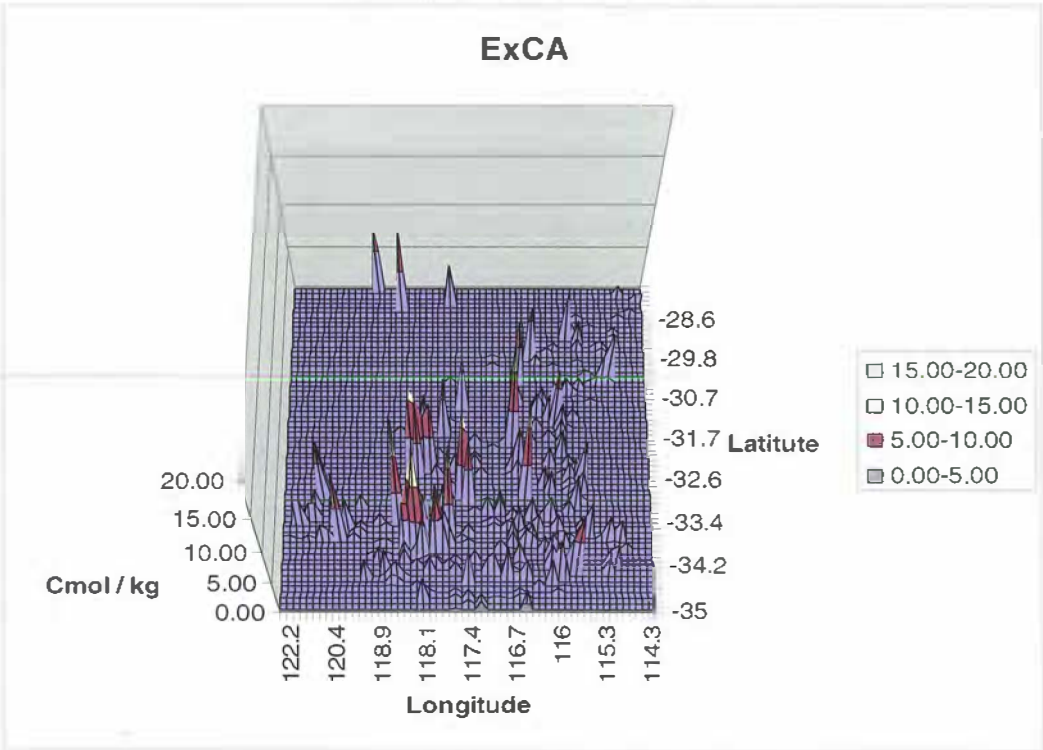
All soils (Normal data) – pH



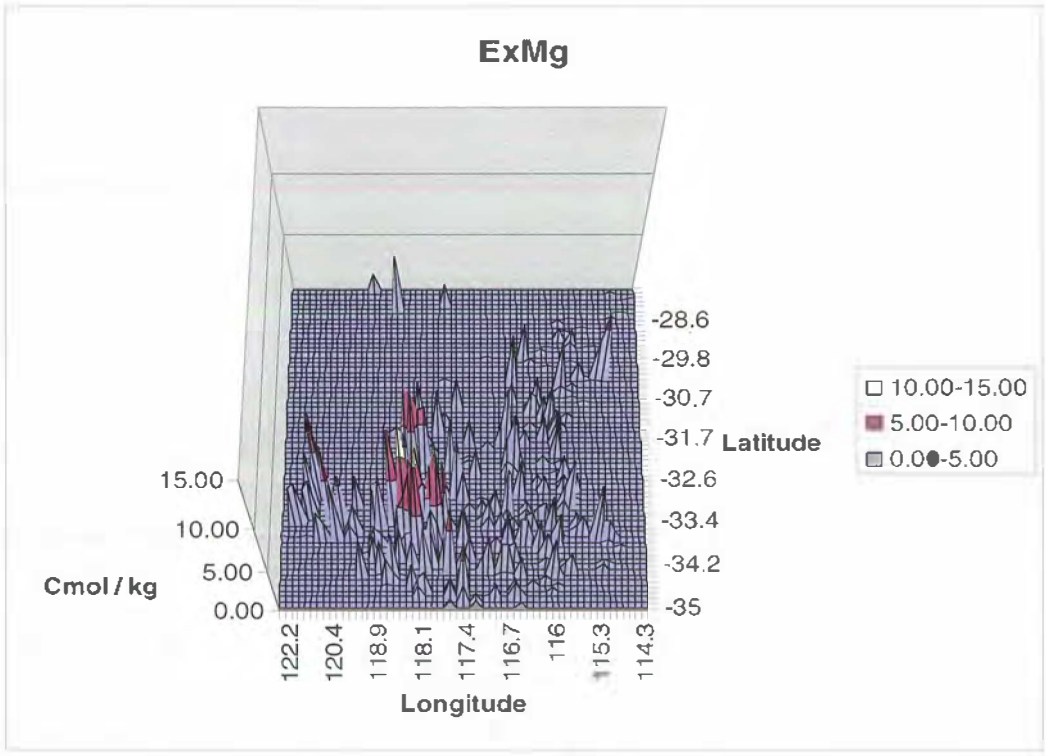
All soils (Normal data) – Clay



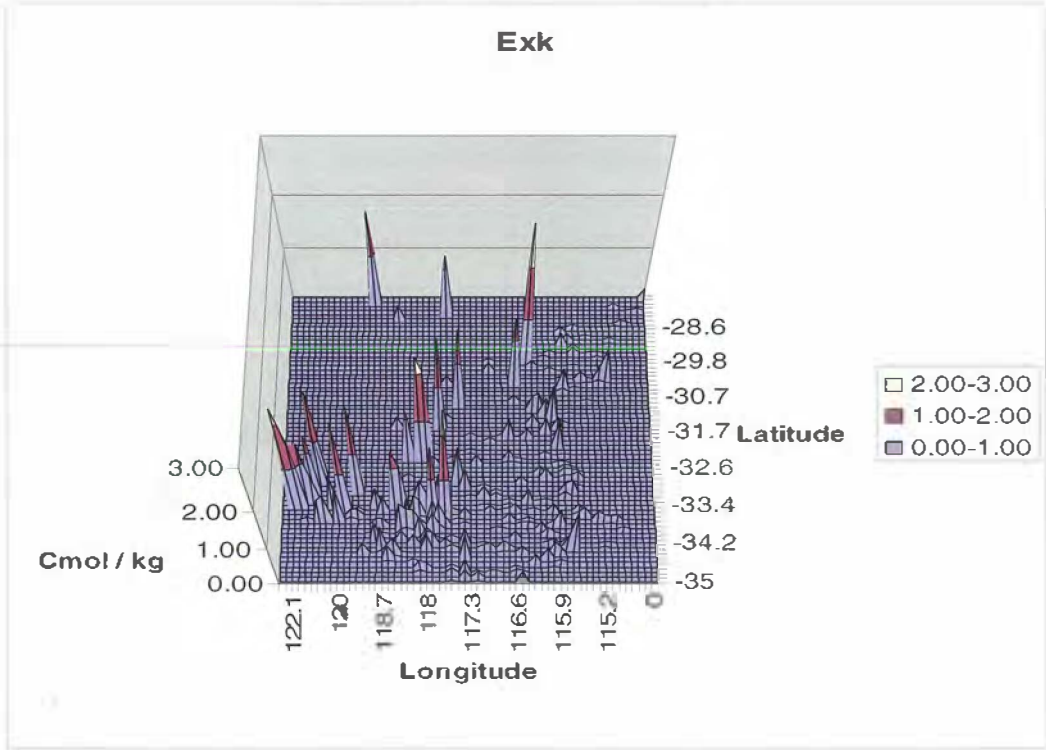
All soils (Normal data) – EC



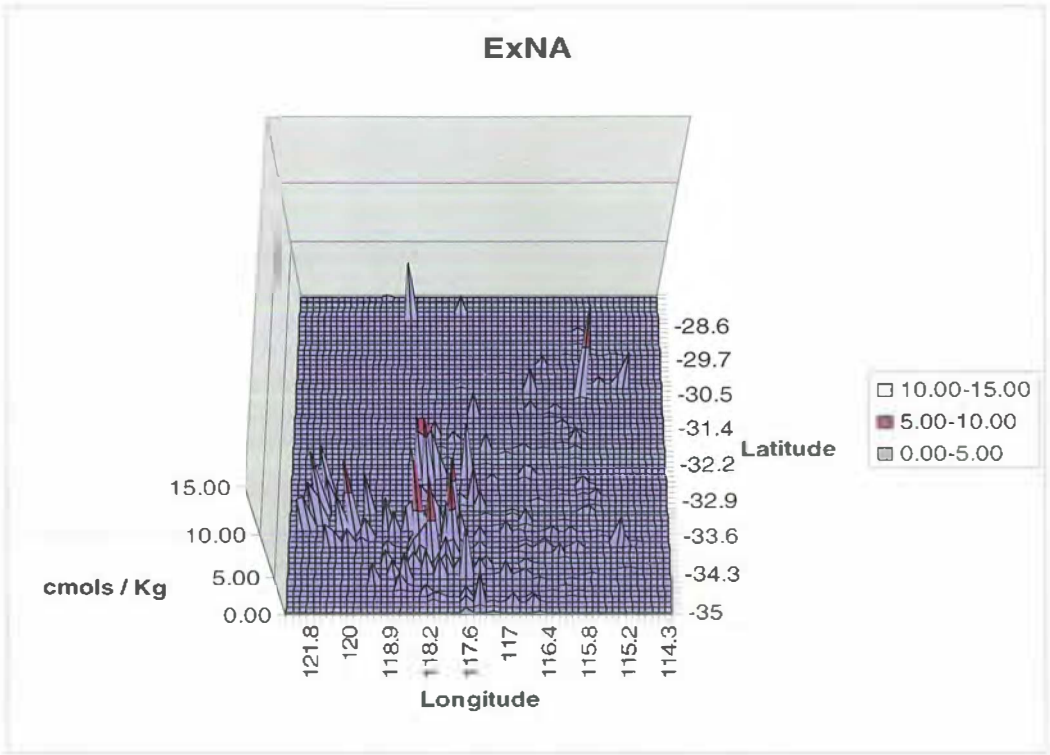
All soils (Normal data) – ExCA



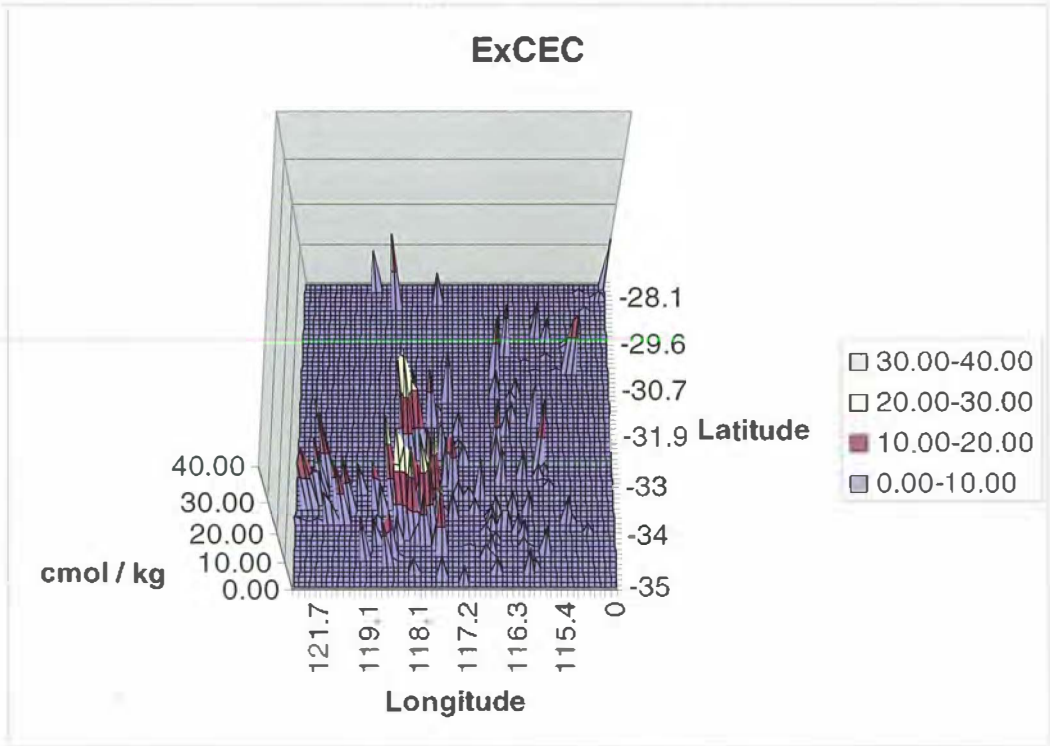
All soils (Normal data) – ExMG



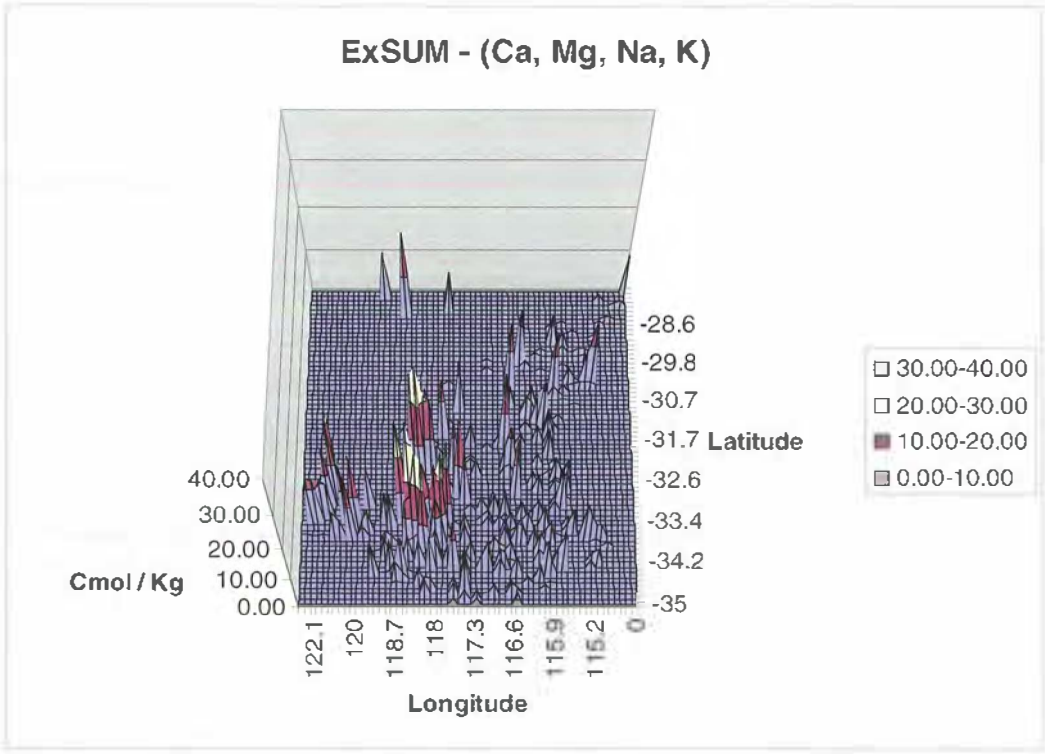
All soils (Normal data) – ExK



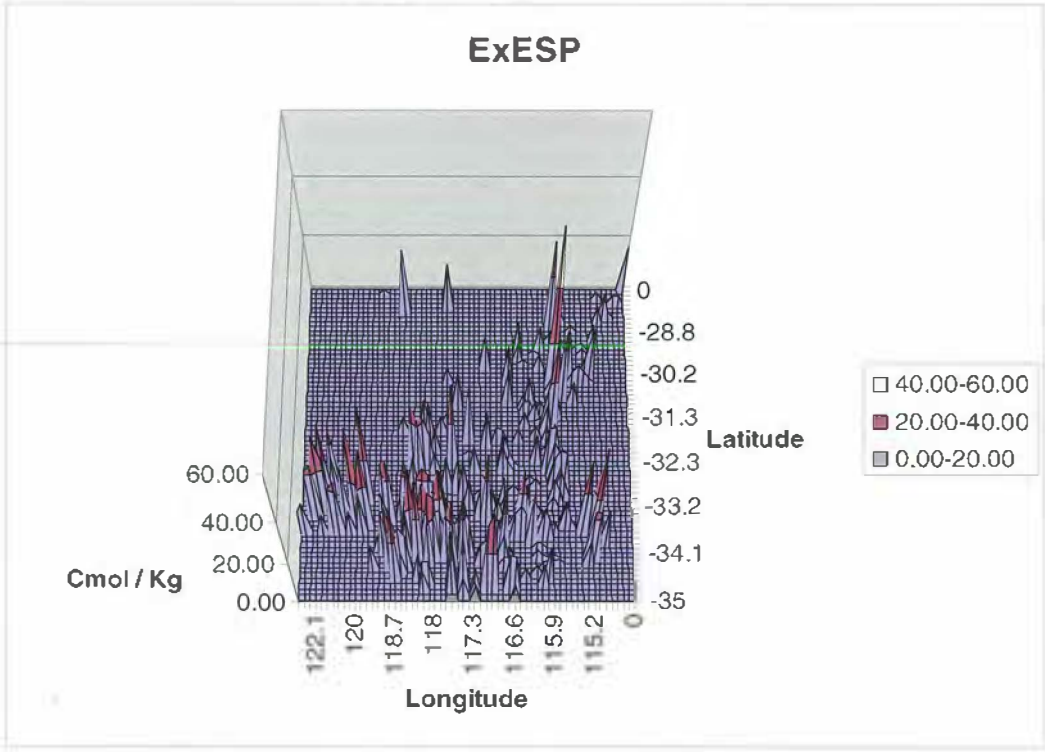
All soils (Normal data) – ExNA



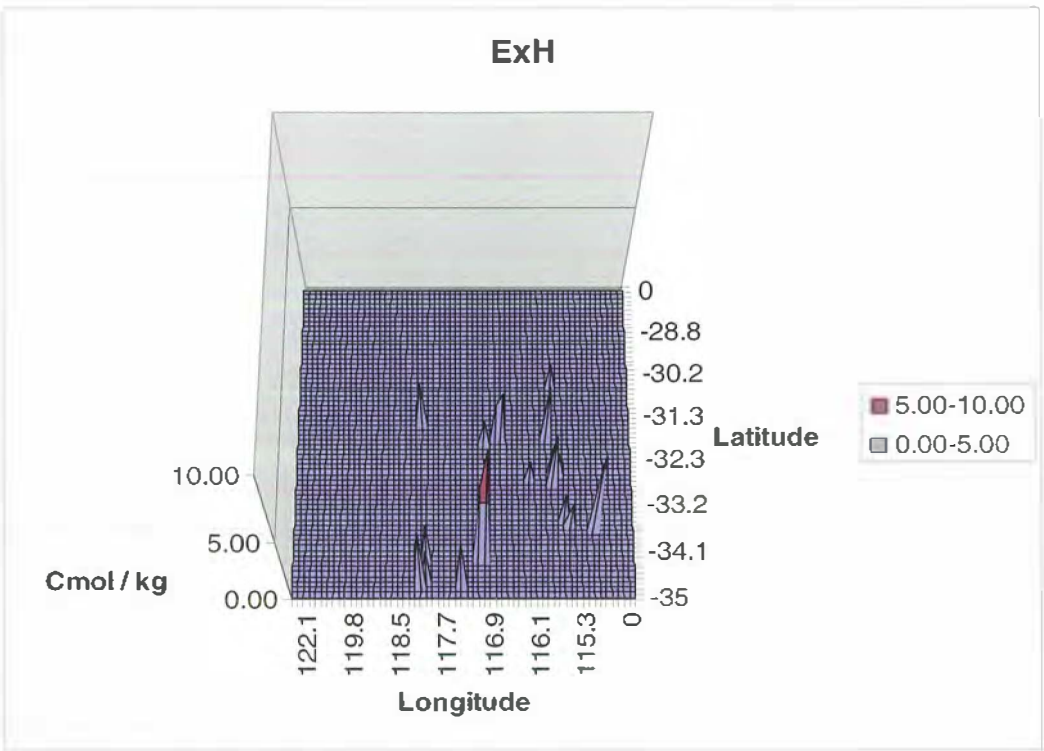
All soils (Normal data) – ExCEC



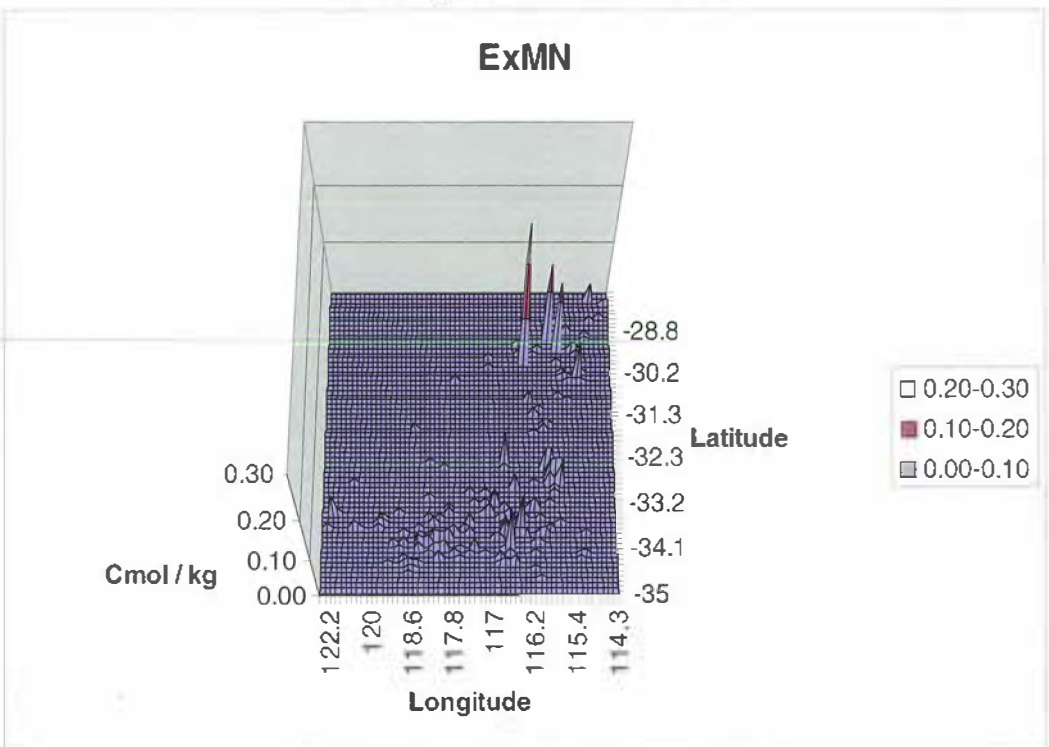
All soils (Normal data) – ExSUM



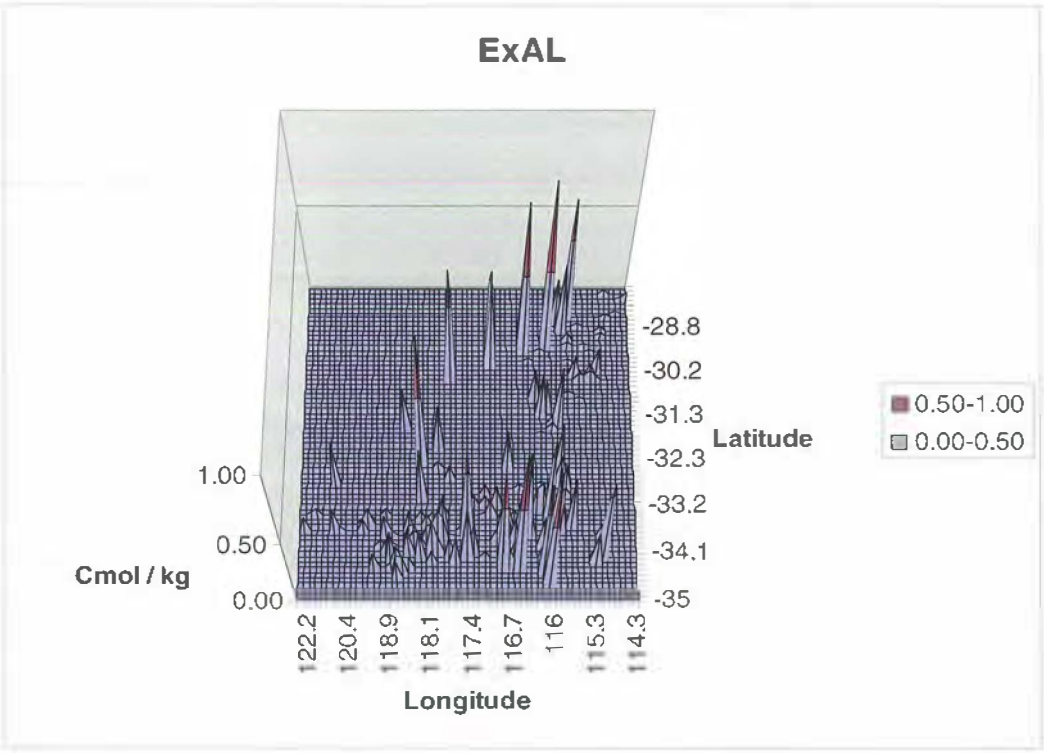
All soils (Normal data) – ExESP



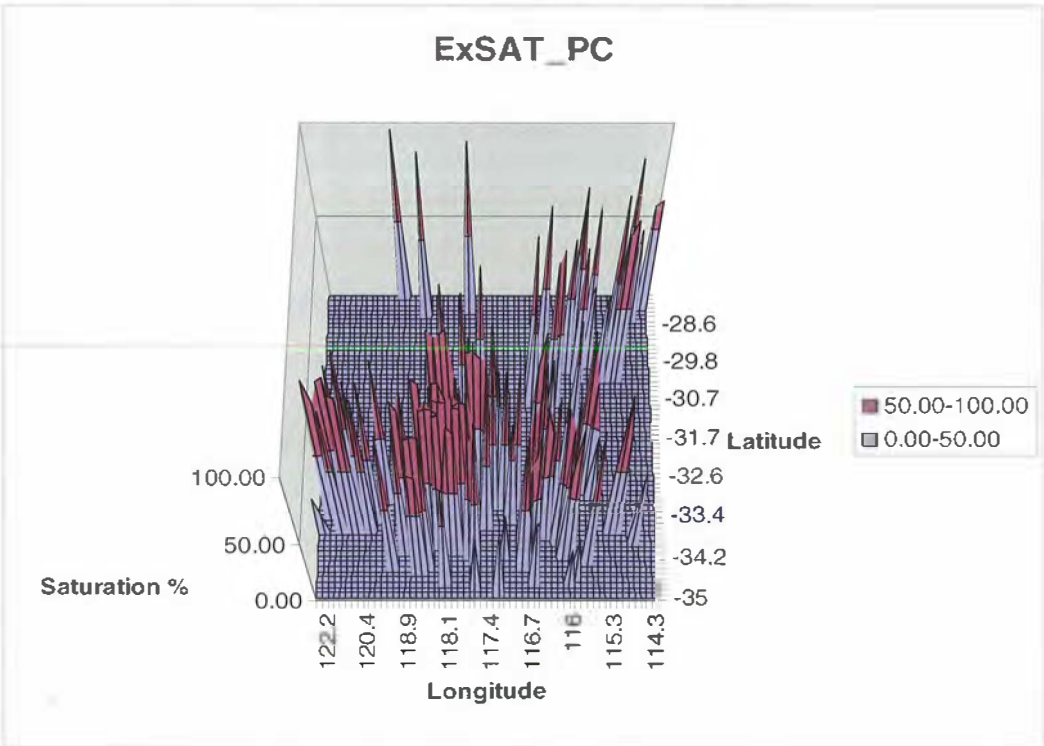
All soils (Normal data) – ExH



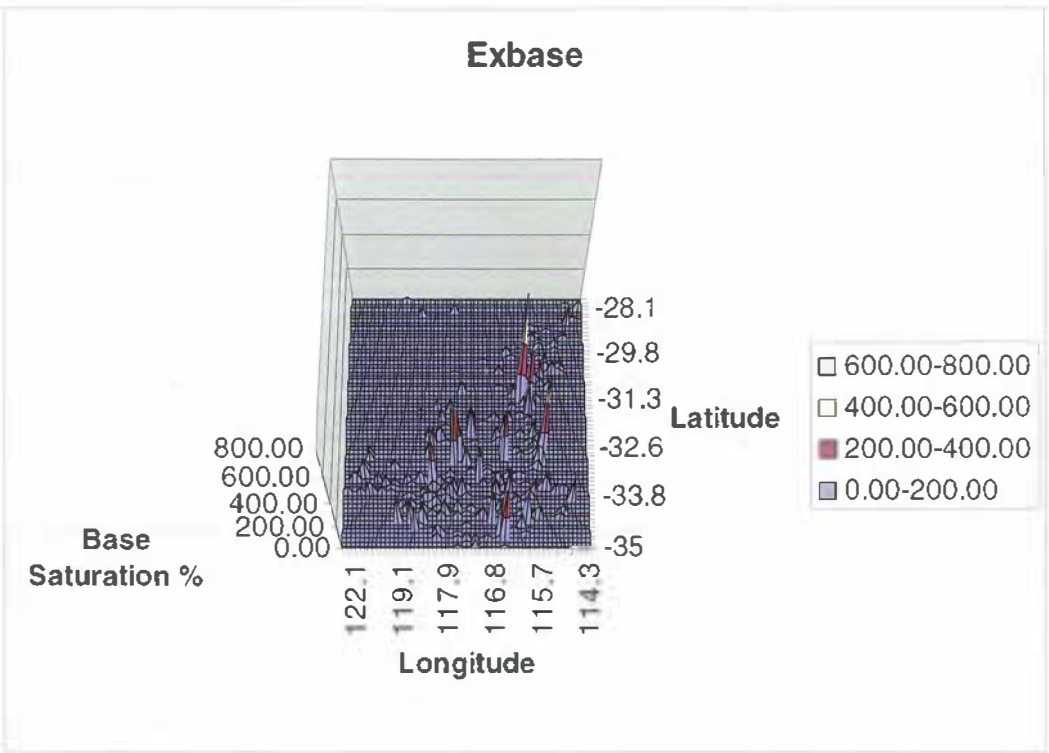
All soils (Normal data) – ExMN



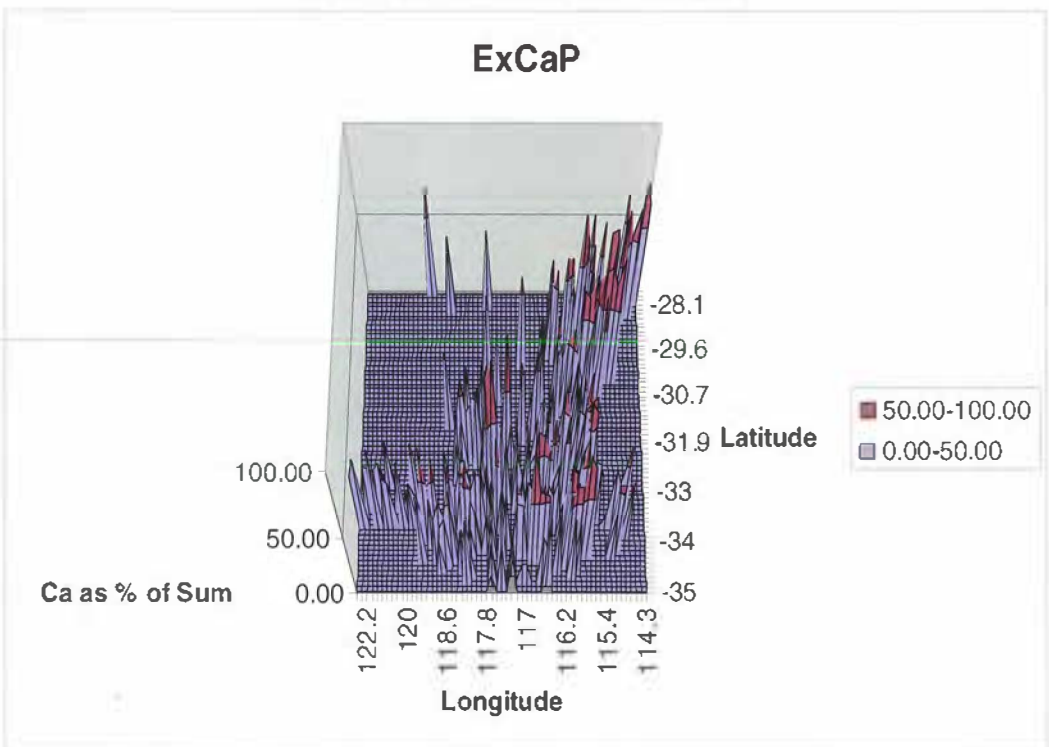
All soils (Normal data) – ExAL



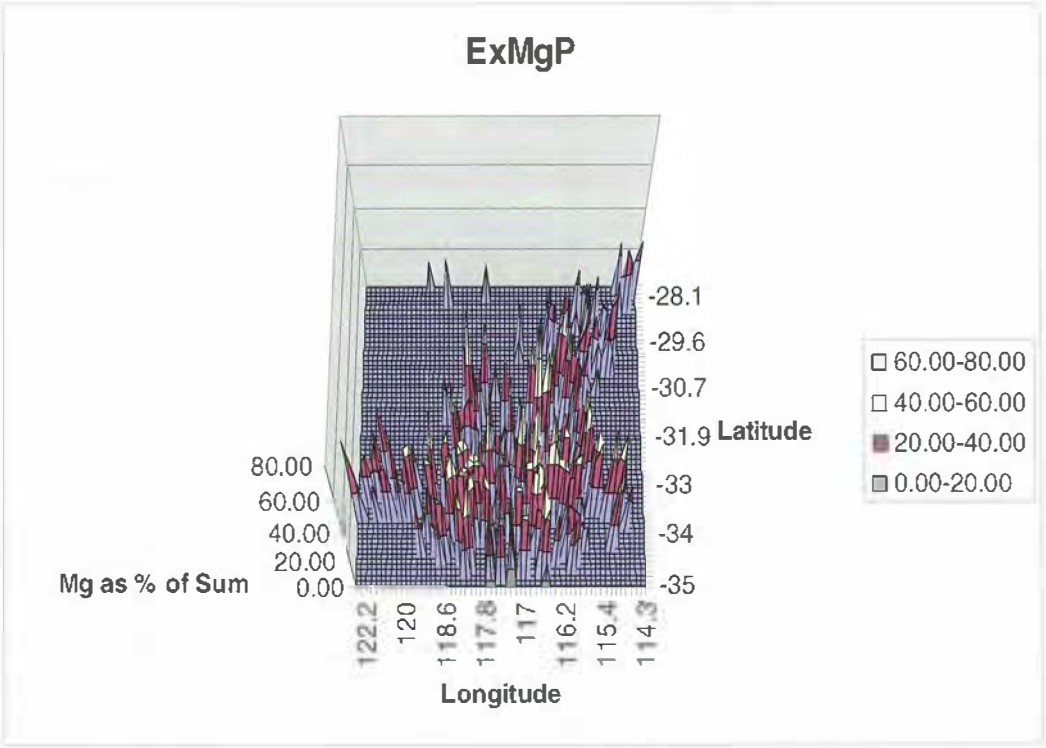
All soils (Normal data) – ExSAT_PC



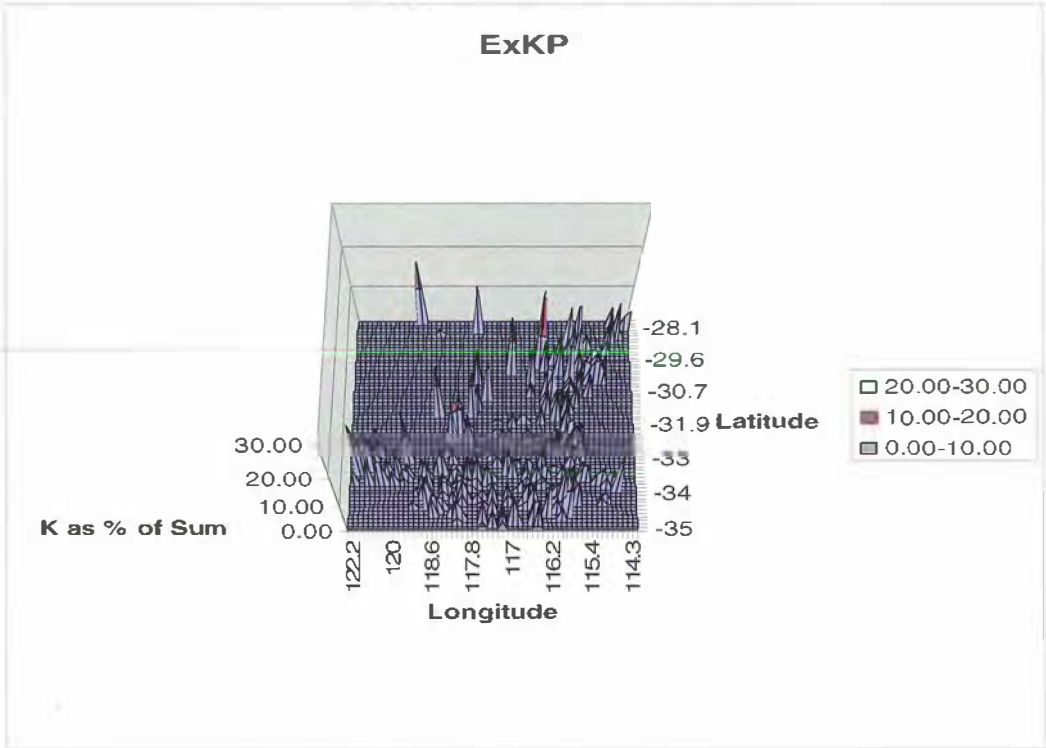
All soils (Normal data) – ExBASE



All soils (Normal data) – ExCaP



All soils (Normal data) – ExMgP



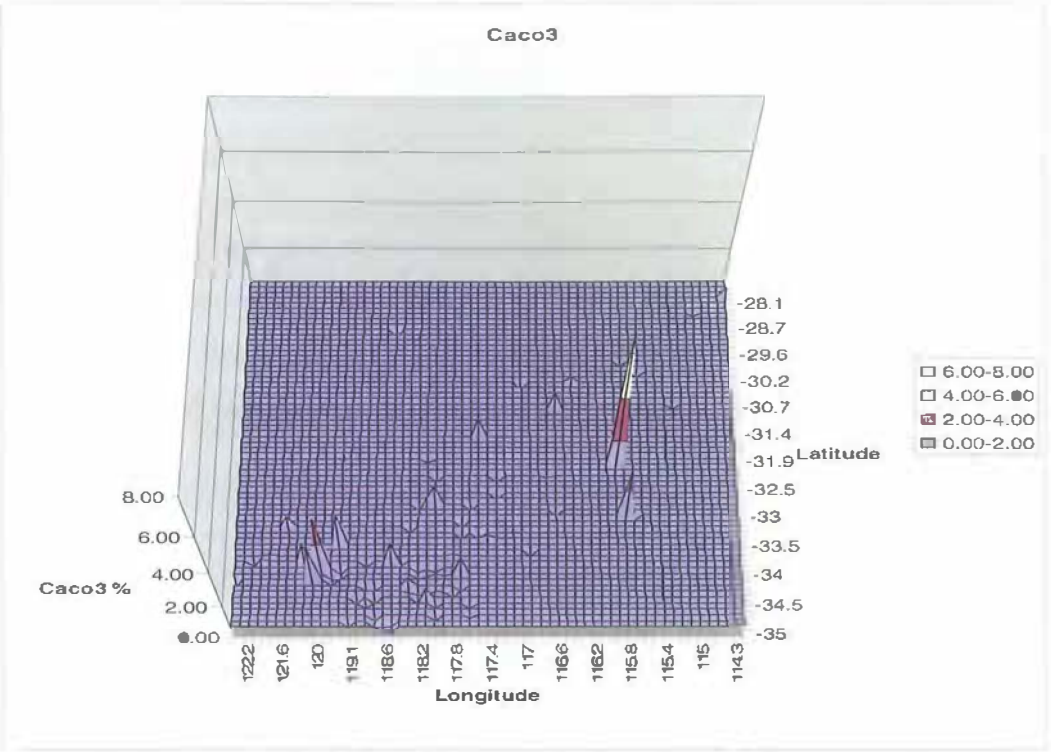
All soils (Normal data) – ExKP

Data	Mean	Min	Max
CACO3	4.65	0.00	69.00
OC	1.04	0.00	16.50
PH	5.59	3.20	9.30
CLAY	19.30	0.00	75.00
EC	23.14	0.00	1000.00
ExCA	2.43	0.00	36.00
ExMG	2.26	0.00	28.90
ExK	0.31	0.00	5.10
ExNA	1.08	0.00	26.60
ExCEC	10.91	0.40	43.00
ExSUM	6.08	0.05	51.80
ExESP	12.74	0.00	82.00
ExH	4.38	0.20	26.40
ExMN	0.02	0.00	2.40
ExAL	0.18	0.00	3.10
ExSAT_PC	84.16	12.00	100.00
ExBASE	66.89	0.00	2937.00
ExCaP	46.79	0.00	98.00
ExMgP	35.26	0.00	99.00
ExKP	5.22	0.00	37.00

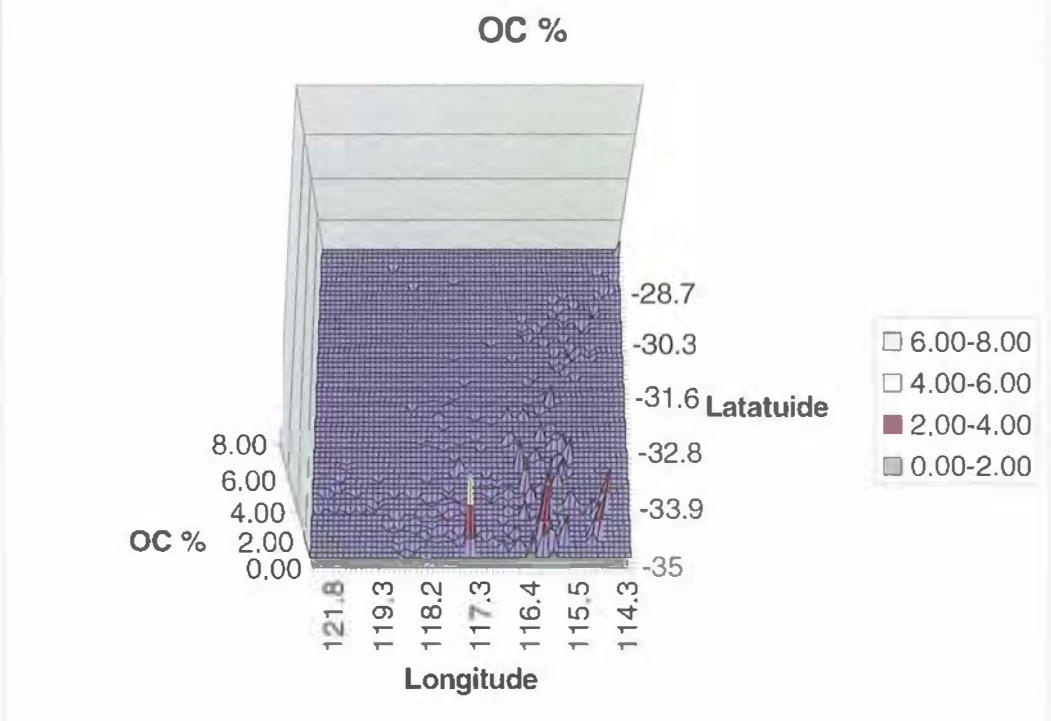
All soils (Normal data) – Data summary

8.3.2 Stage 2: Standardized data - All soil types

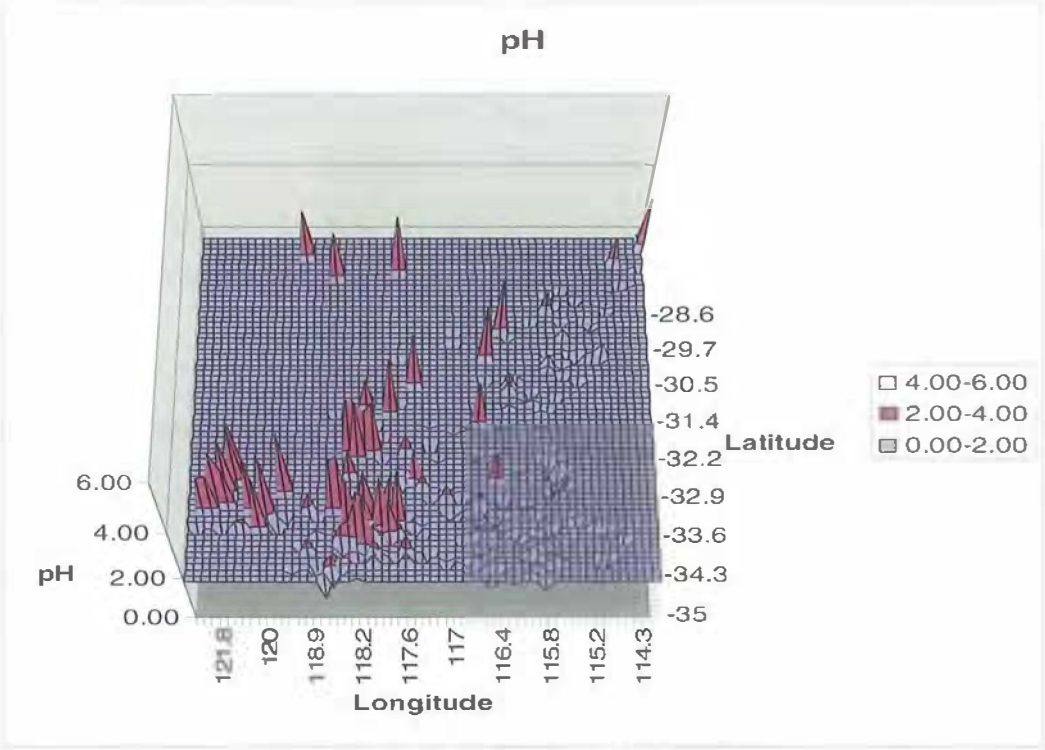
The graphs contained within this section are all 3D surface maps using the agdata full dataset. The graphic display the standardized data changed from that as supplied by the DAFWA and each trait is graphed against it longitude and latitude. The values are averages for each location and show the traits distribution across the South Western agricultural region.



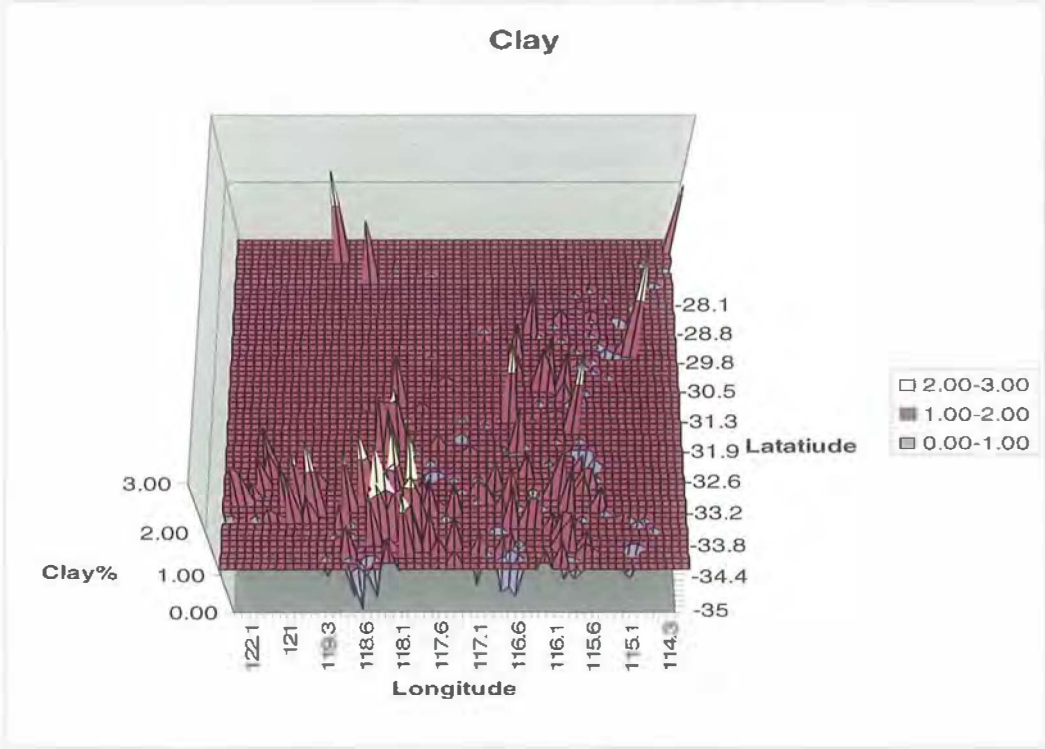
All soils (Standardized data) – CAC03 %



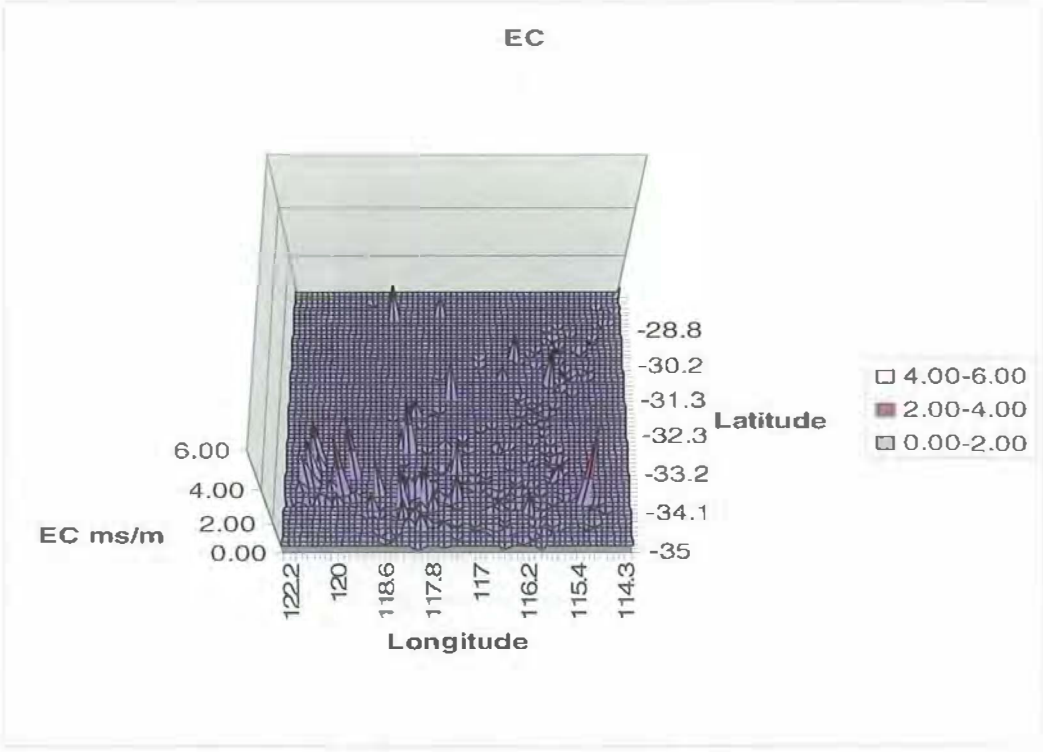
All soils (Standardized data) – ●C%



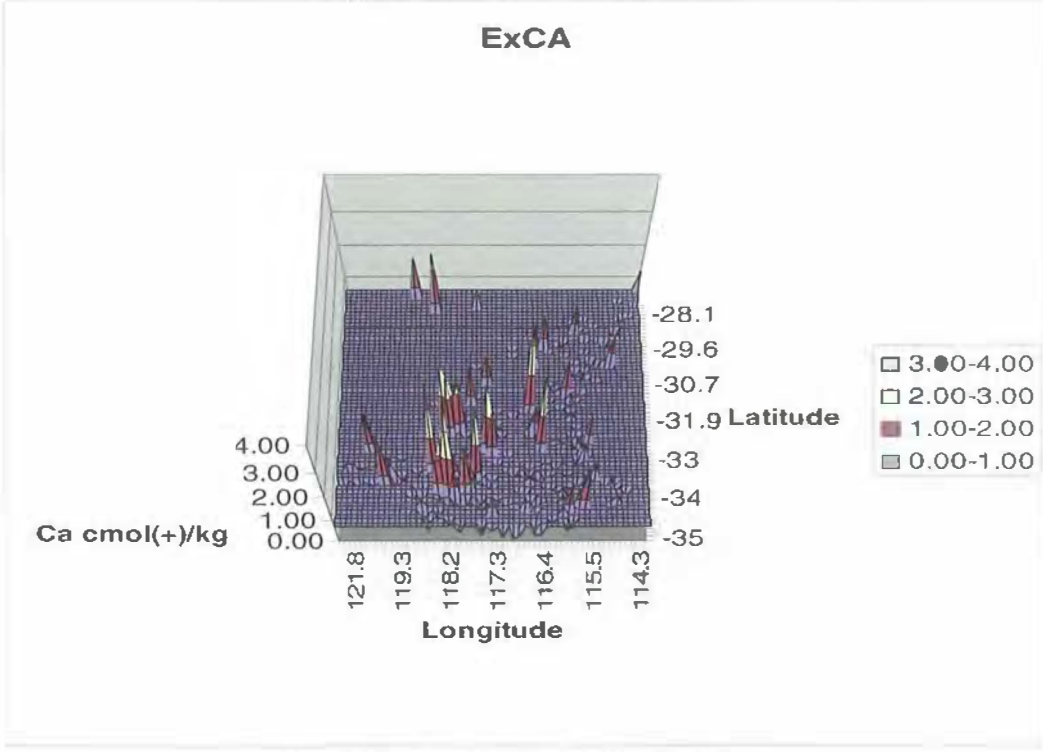
All soils (Standardized data) – pH



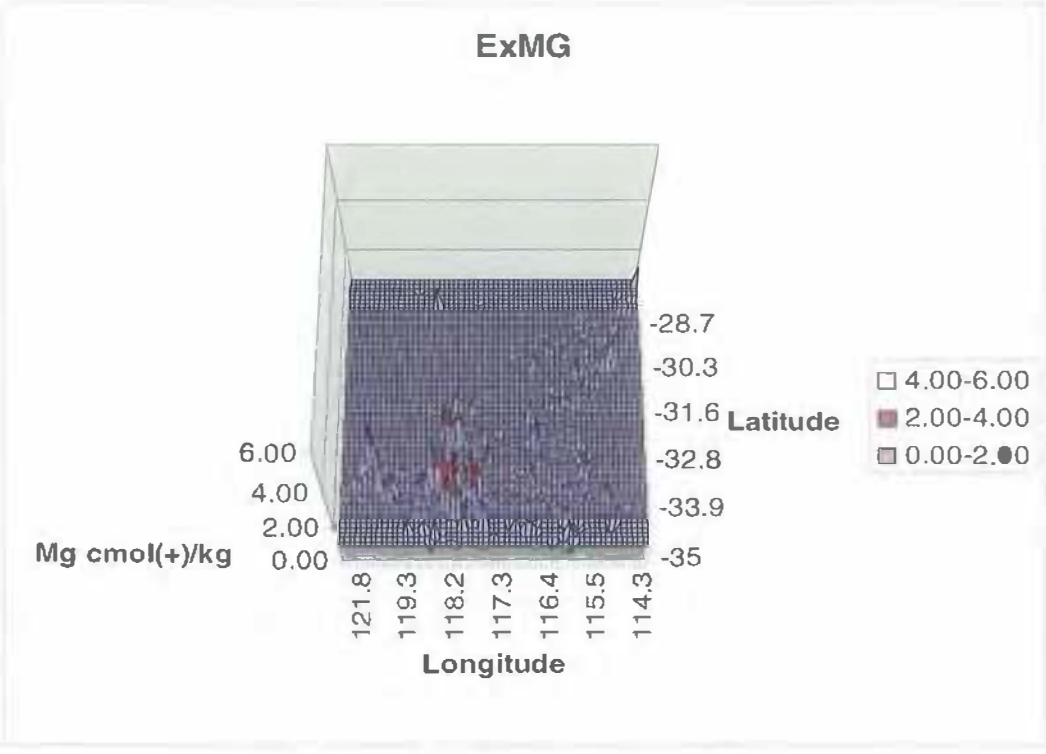
All soils (Standardized data) – Clay



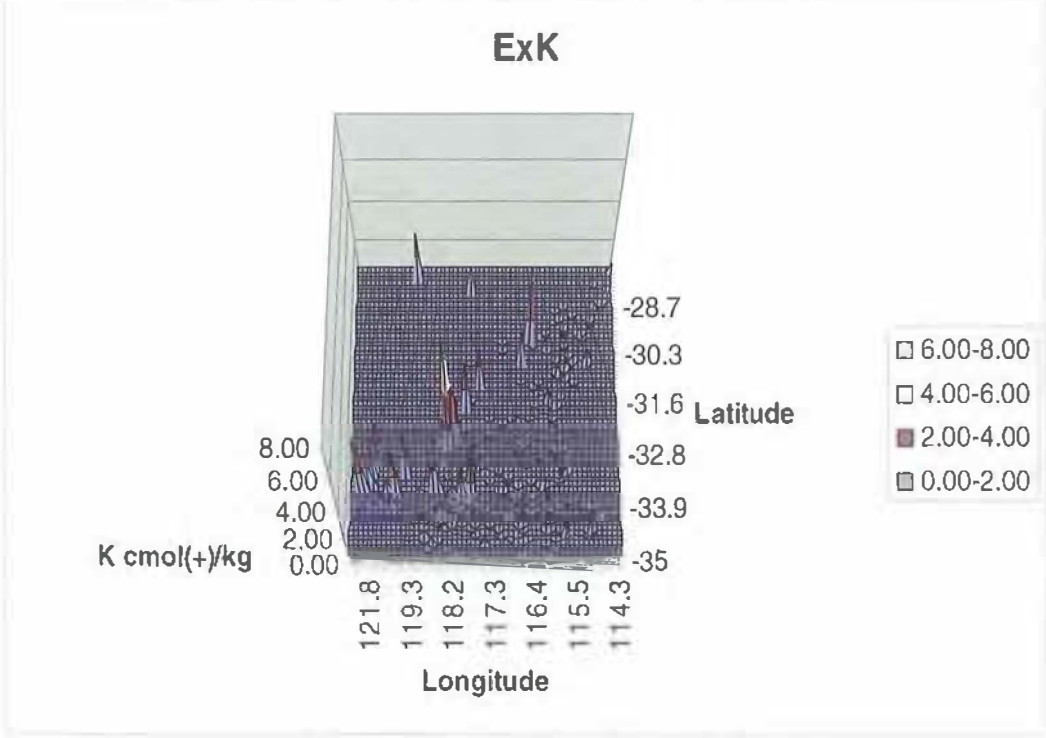
All soils (Standardized data) – EC



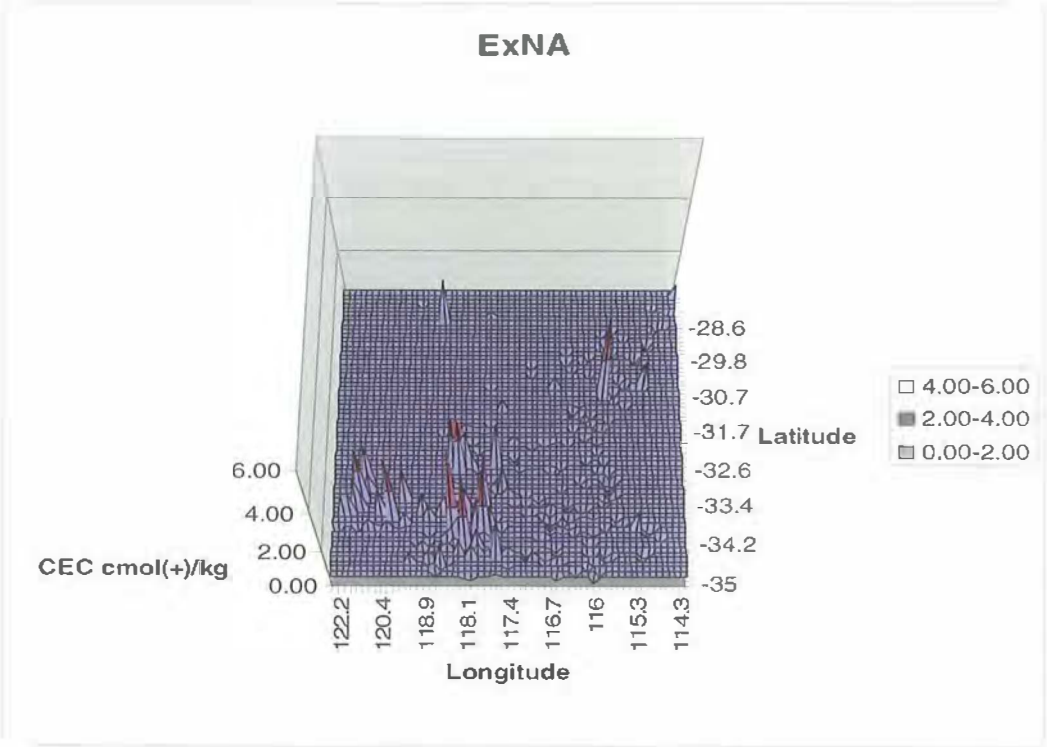
All soils (Standardized data) – ExCA



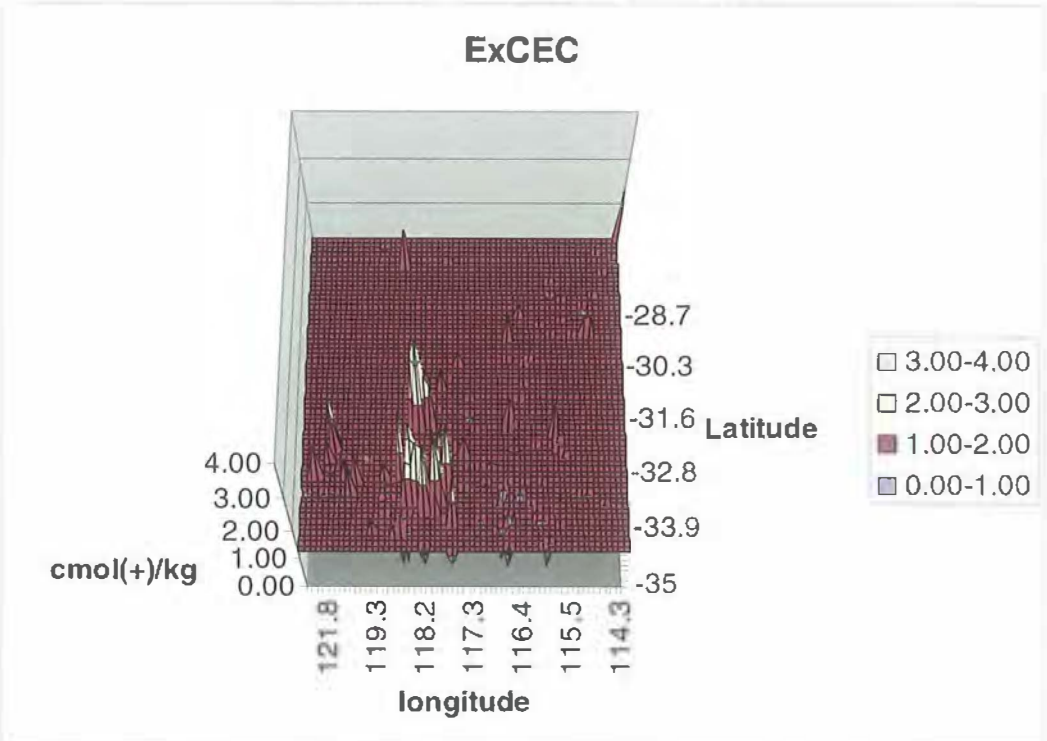
All soils (Standardized data) – ExMG



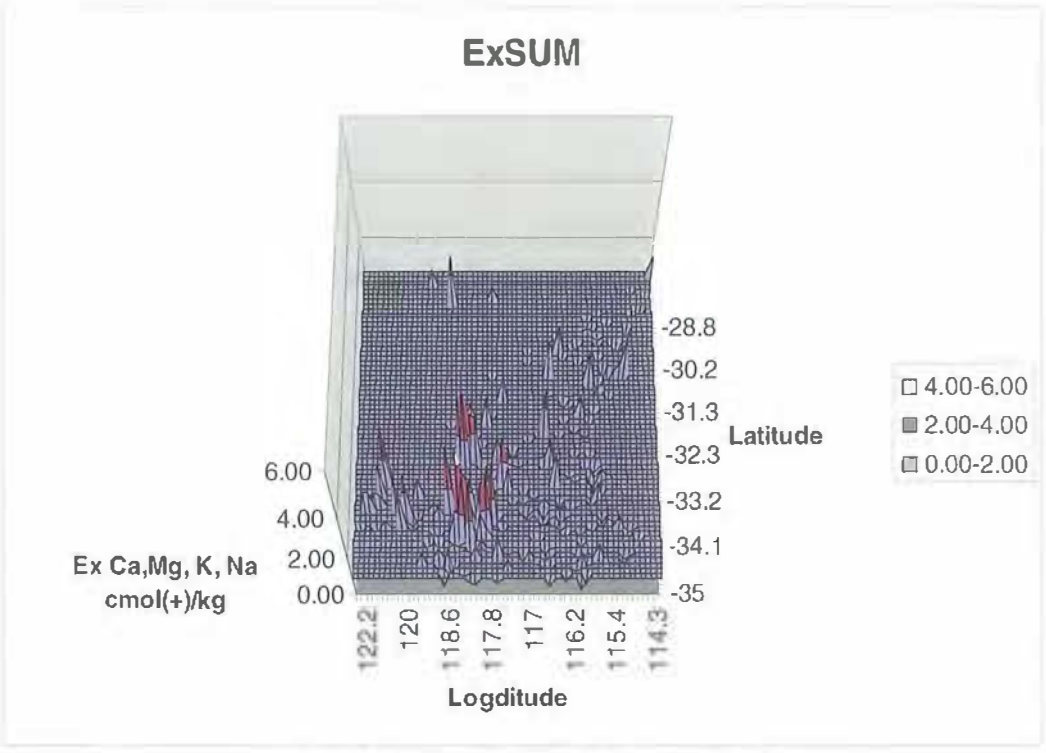
All soils (Standardized data) – ExK



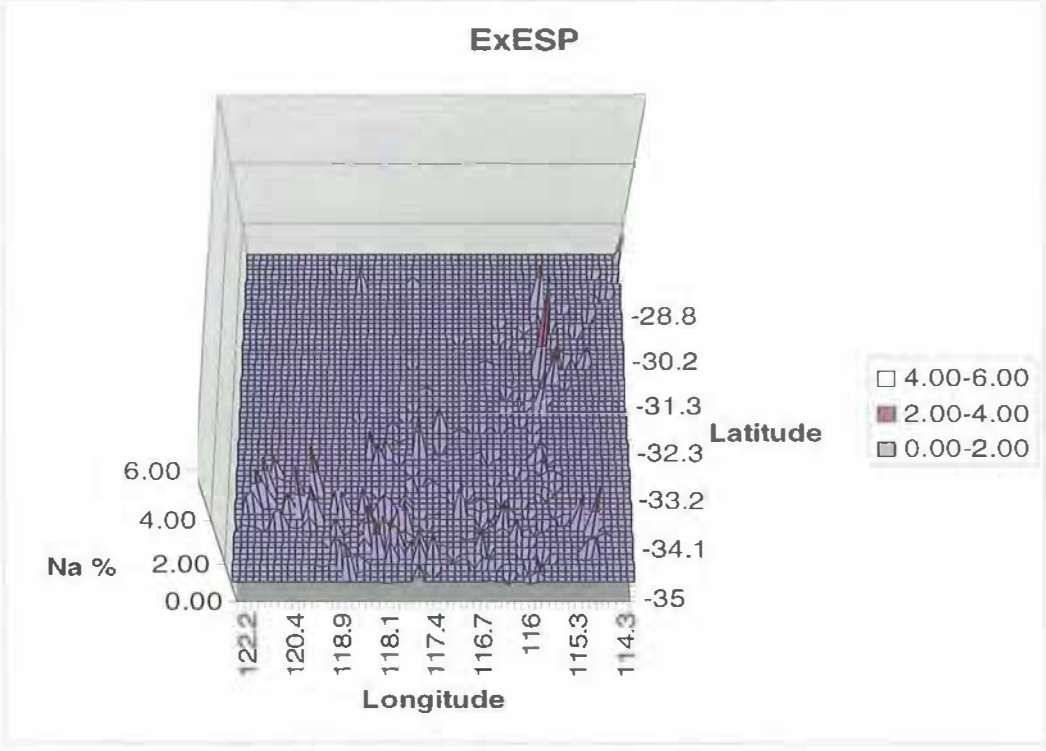
All soils (Standardized data) – ExNA



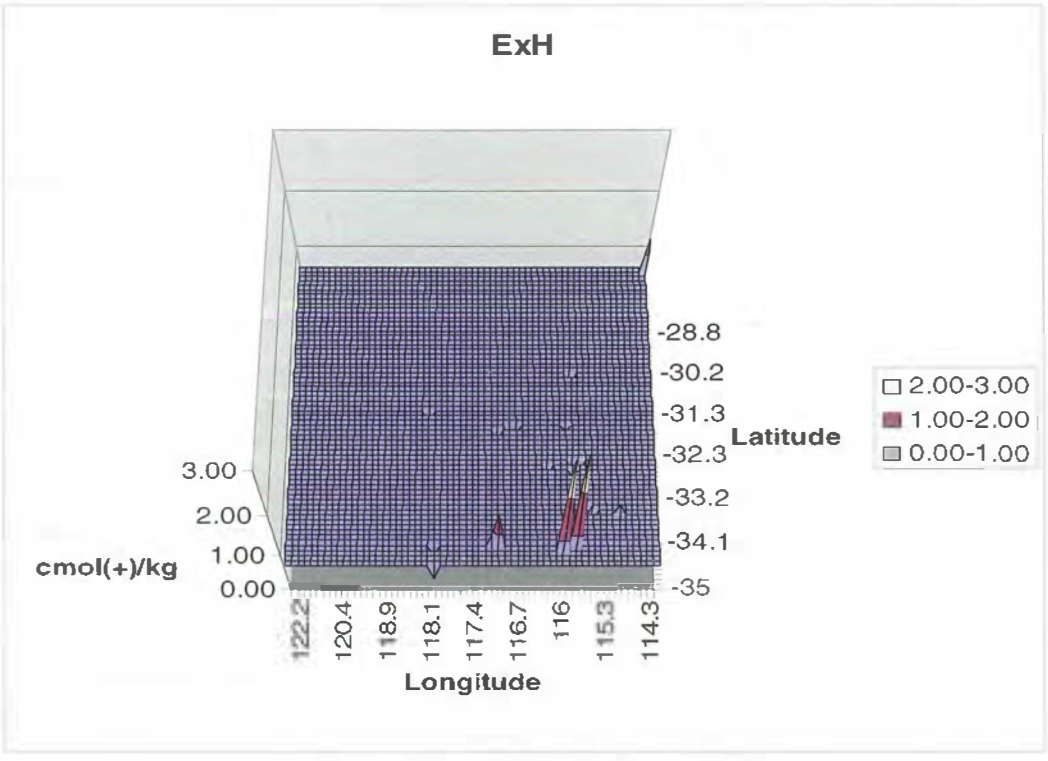
All soils (Standardized data) – ExCEC



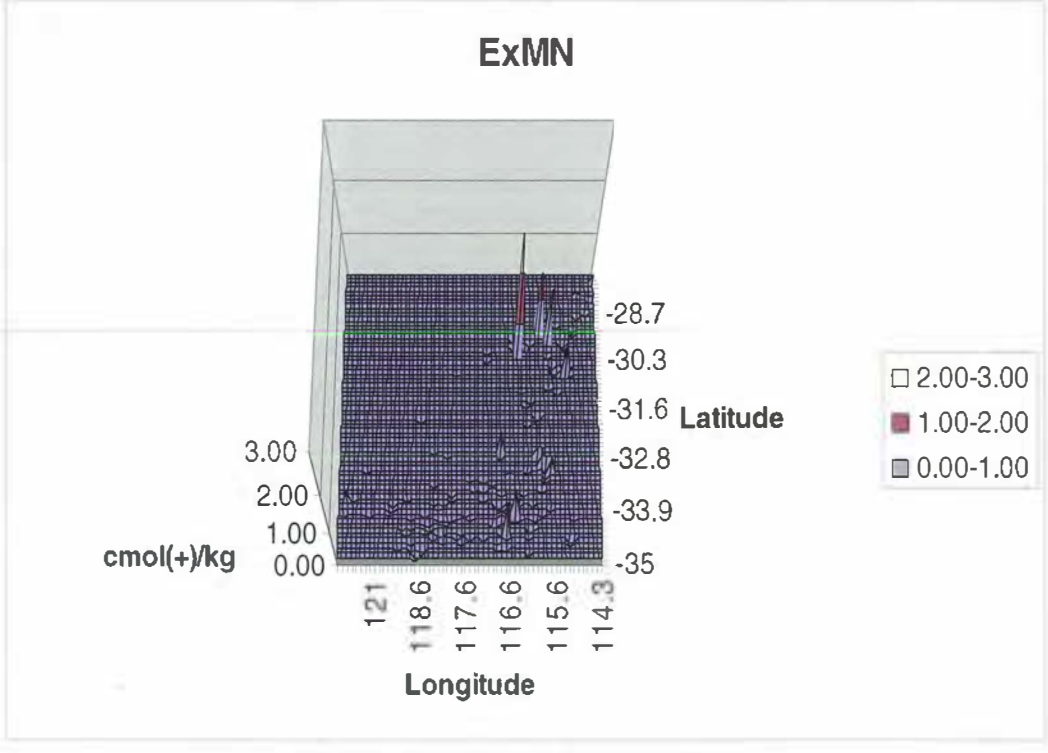
All soils (Standardized data) – ExSUM



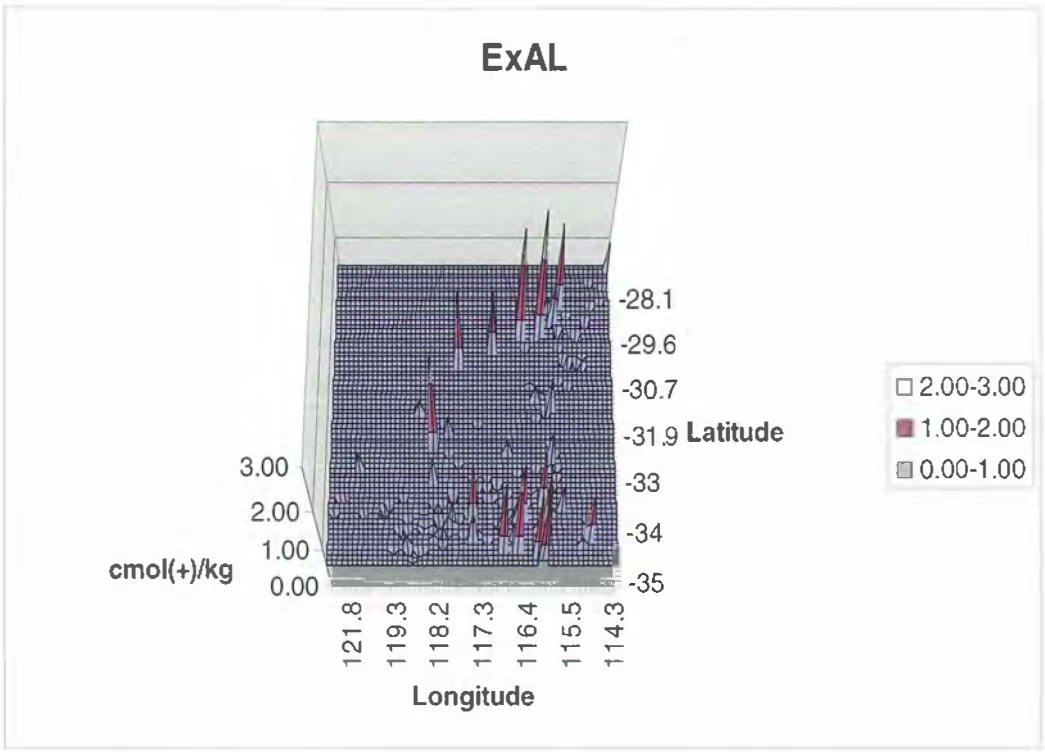
All soils (Standardized data) – ExESP



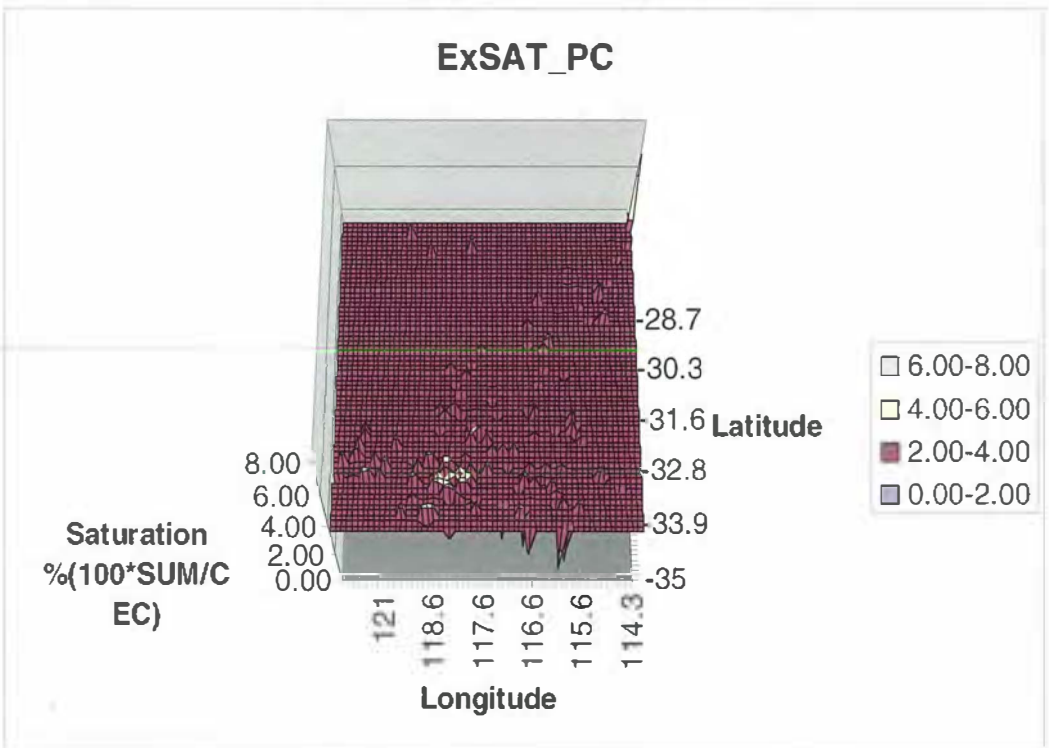
All soils (Standardized data) – ExH



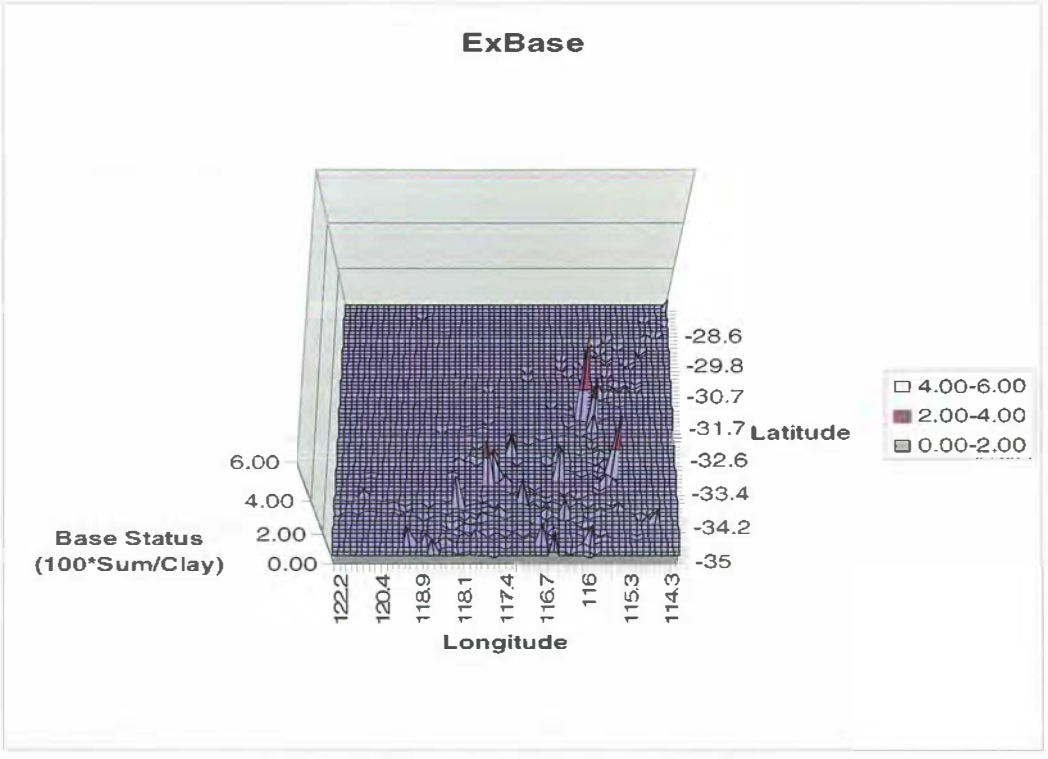
All soils (Standardized data) – ExMN



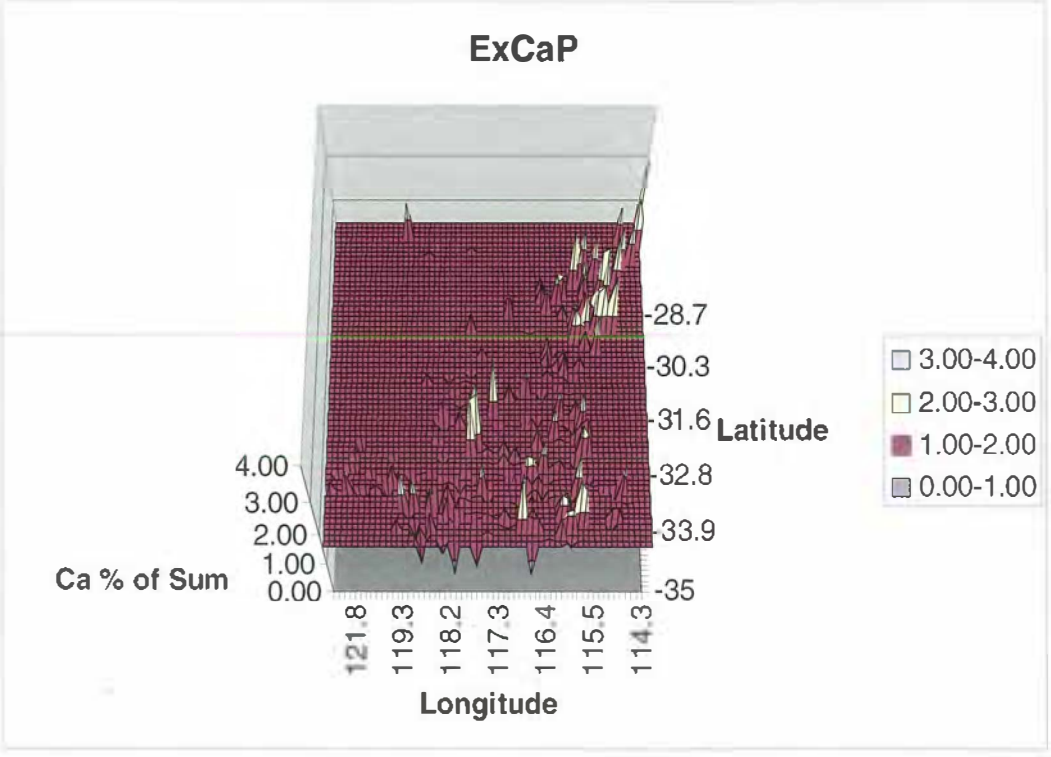
All soils (Standardized data) – ExAL



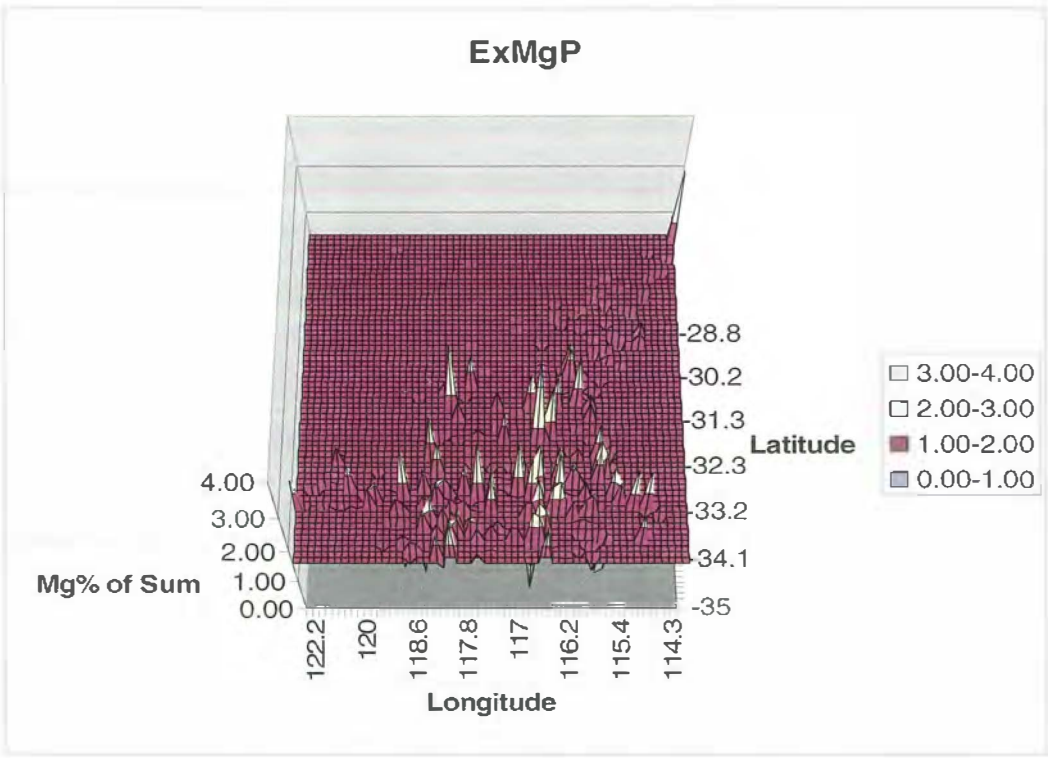
All soils (Standardized data) – ExSAT_PC



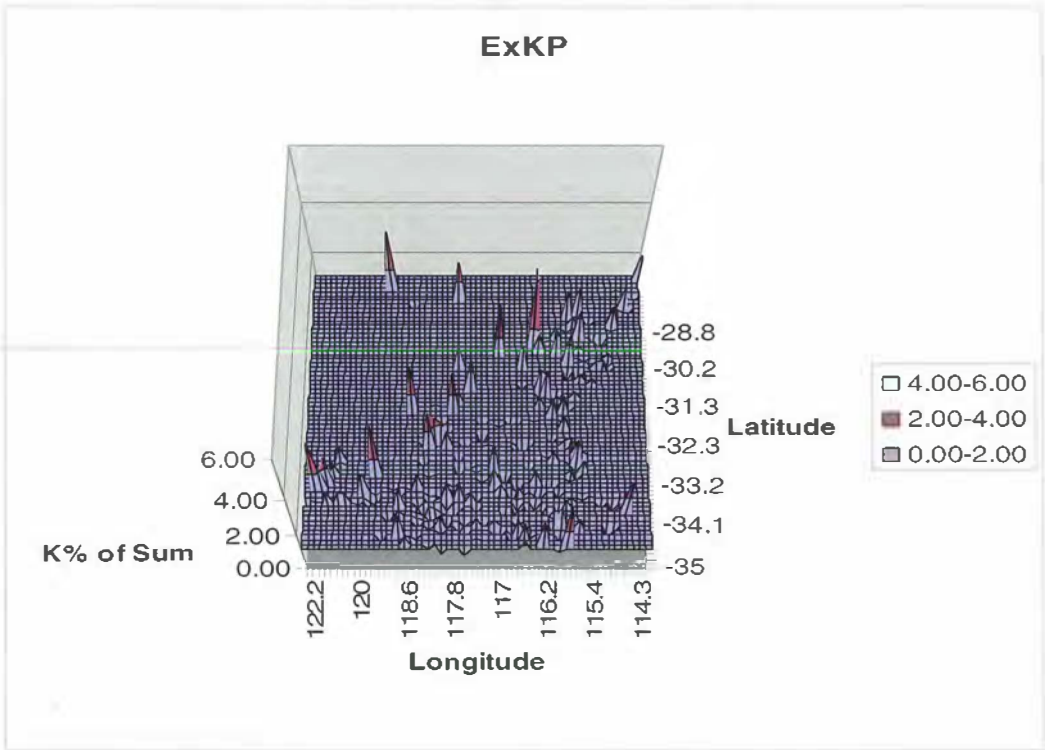
All soils (Standardized data) – ExBASE



All soils (Standardized data) – ExCaP



All soils (Standardized data) – ExMgP



All soils (Standardized data) – ExKP

8.3.3 Stage 3: Correlation Table

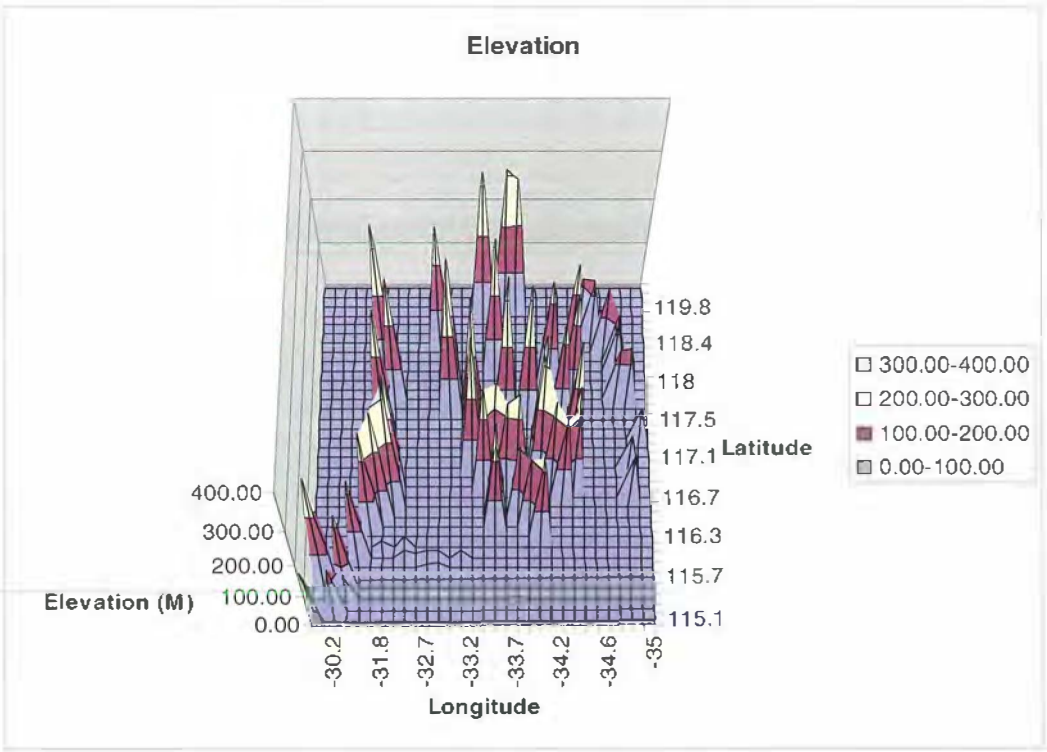
	CACO3	*OC*	*PH*	*clay*	*EC*	*ExCA*	*ExMG*	*ExK*	*ExNA*	*ExCEC*
CACO3	**1.00**									
OC	-0.05	**1.00**								
PH	0.44	-0.15	**1.00**							
clay	0.23	-0.19	0.42	**1.00**						
EC	0.32	-0.09	0.43	0.33	**1.00**					
ExCA	0.12	0.45	0.39	0.13	0.15	**1.00**				
ExMG	0.26	0.00	**0.60**	**0.61**	0.43	0.41	**1.00**			
ExK	0.30	0.03	**0.64**	0.33	0.40	0.45	0.46	**1.00**		
ExNA	0.27	-0.14	**0.58**	0.45	**0.57**	0.13	**0.75**	0.45	**1.00**	
ExCEC	0.19	0.09	**0.69**	**0.57**	0.28	**0.62**	**0.84**	**0.60**	**0.65**	**1.00**
ExSUM	0.30	0.19	**0.67**	**0.48**	0.46	**0.74**	**0.88**	**0.63**	**0.71**	**0.93**
ExESP	0.22	-0.27	0.28	0.32	0.45	-0.18	0.39	0.22	**0.67**	0.32
ExH	**-0.77**	**0.82**	-0.12	-0.20	-0.08	0.42	0.14	0.11	-0.11	#DIV/0!
ExMN	-0.04	0.11	0.11	0.02	0.33	0.12	0.05	0.12	0.02	-0.28
ExAL	-0.14	0.14	**-0.49**	0.08	0.09	-0.01	0.28	0.04	0.19	-0.14
ExSAT_PC	0.25	-0.15	**0.49**	0.42	0.30	0.30	**0.47**	0.33	0.35	0.37
ExBASE	-0.04	0.35	-0.02	-0.24	0.02	0.30	0.06	0.05	0.01	0.03
ExCaP	-0.22	0.34	-0.24	**-0.55**	-0.32	0.28	**-0.48**	-0.16	**-0.47**	-0.20
ExMgP	0.04	-0.24	0.08	**0.52**	0.08	-0.23	0.38	-0.06	0.15	-0.03
ExKP	0.20	-0.12	0.13	-0.05	0.08	-0.05	-0.12	**0.46**	0.00	0.04

	ExSUM	*ExESP*	*ExH*	*ExMN*	*ExAL*	*ExSAT_PC*	*ExBASE*	*ExCaP*	*ExMgP*	*ExKP*
CACO3										
OC										
PH										
clay										
EC										
ExCA										
ExMG										
ExK										
ExNA										
ExCEC										
ExSUM	**1.00**									
ExESP	0.29	**1.00**								
ExH	0.30	-0.23	**1.00**							
ExMN	0.10	0.00	#DIV/0!	**1.00**						
ExAL	0.19	0.12	#DIV/0!	-0.01	**1.00**					
ExSAT_PC	**0.49**	0.25	#DIV/0!	0.45	**-0.47**	**1.00**				
ExBASE	0.18	-0.10	**0.76**	0.03	-0.03	0.13	**1.00**			
ExCaP	-0.21	**-0.69**	0.27	-0.01	-0.11	-0.17	0.20	**1.00**		
ExMgP	0.08	0.19	-0.17	0.01	0.06	0.01	-0.16	**-0.82**	**1.00**	
ExKP	-0.03	0.07	-0.20	0.02	0.02	-0.02	-0.11	-0.05	-0.23	**1.00**

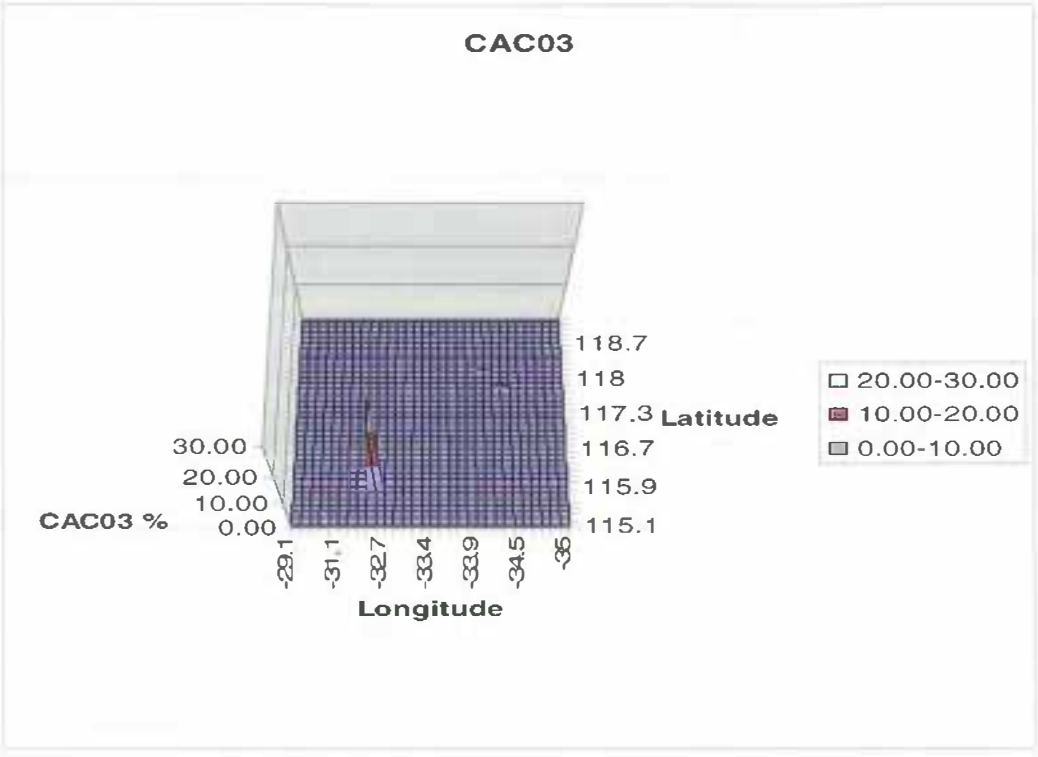
8.3.4 Stage 4: Normal data – 3 Main soil types

The graphs contained within this section are all 3D surface maps using the subset of the agdata dataset. The graphic display the normal data as supplied by the DAFWA and each trait is graphed against it longitude and latitude. The values are averages for each location and show the traits distribution across the South Western agricultural region.

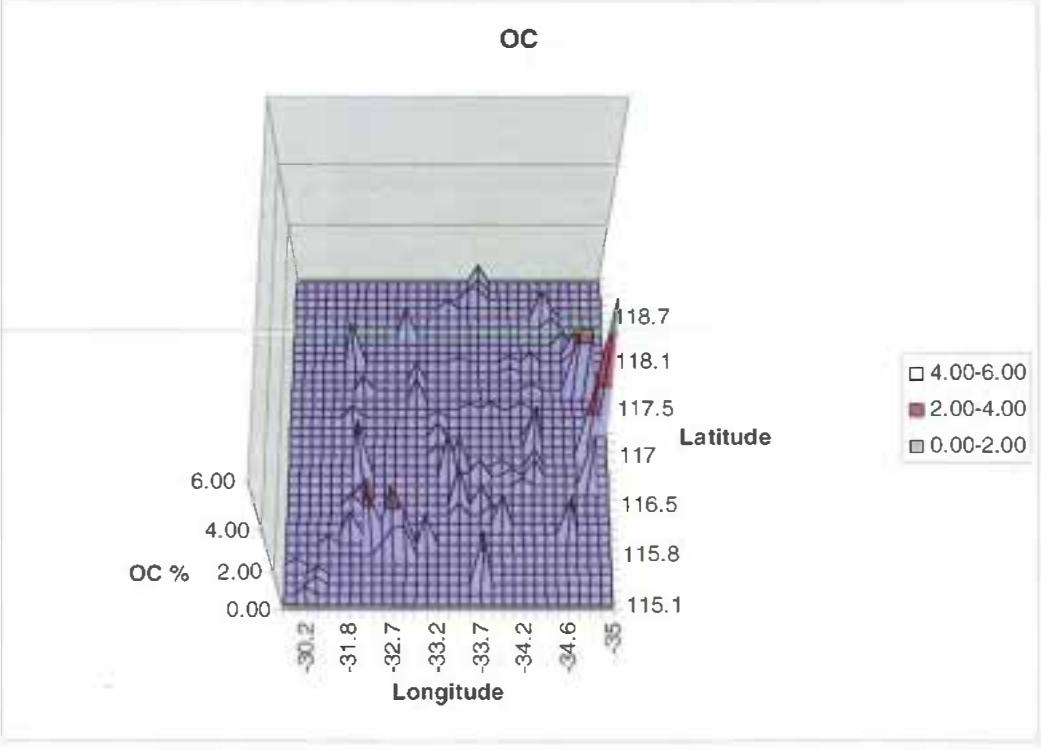
Soil 1 - Grey deep sandy duplex.



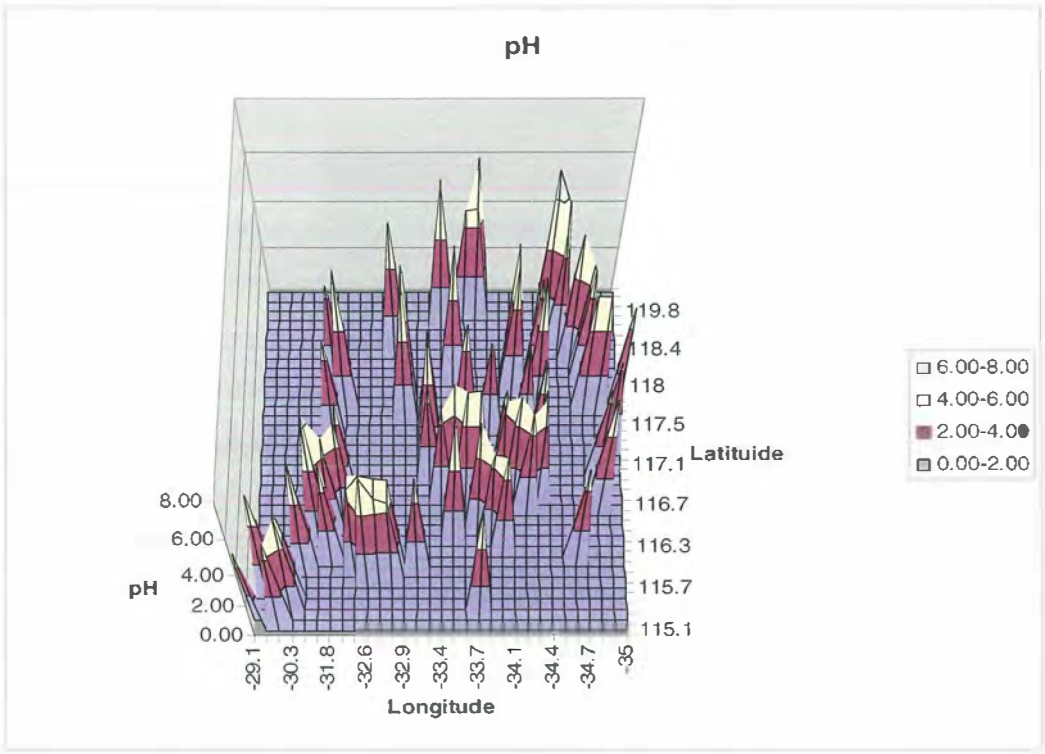
Grey deep sandy duplex (Normal data) – Elevation



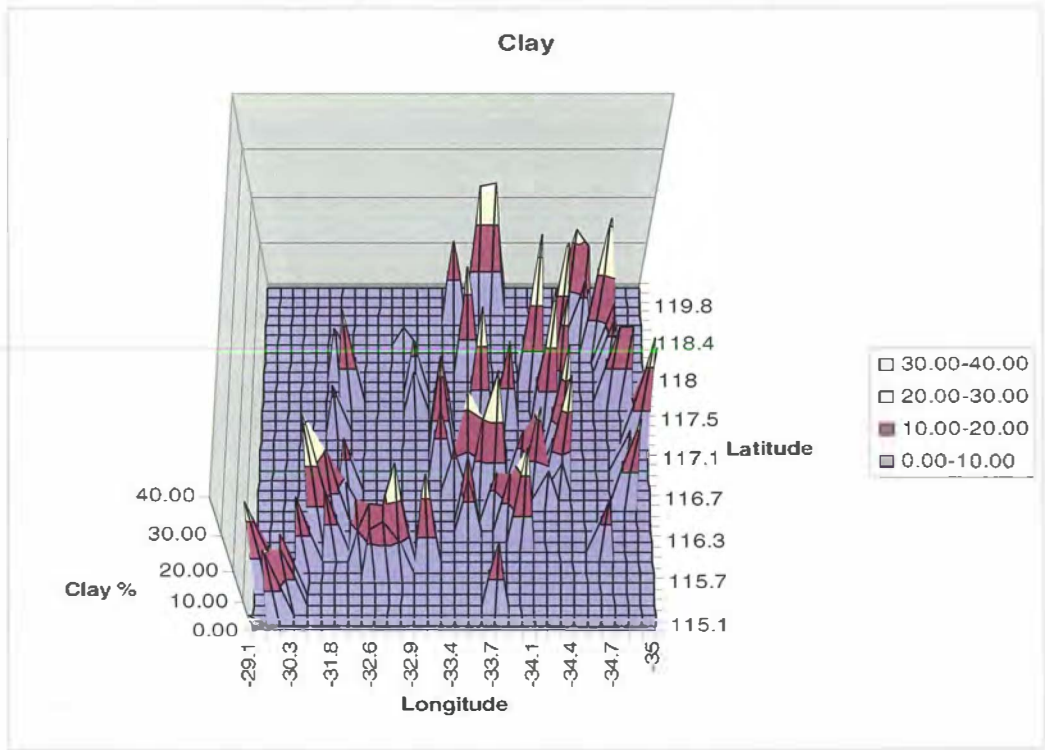
Grey deep sandy duplex (Normal data) – CAC03



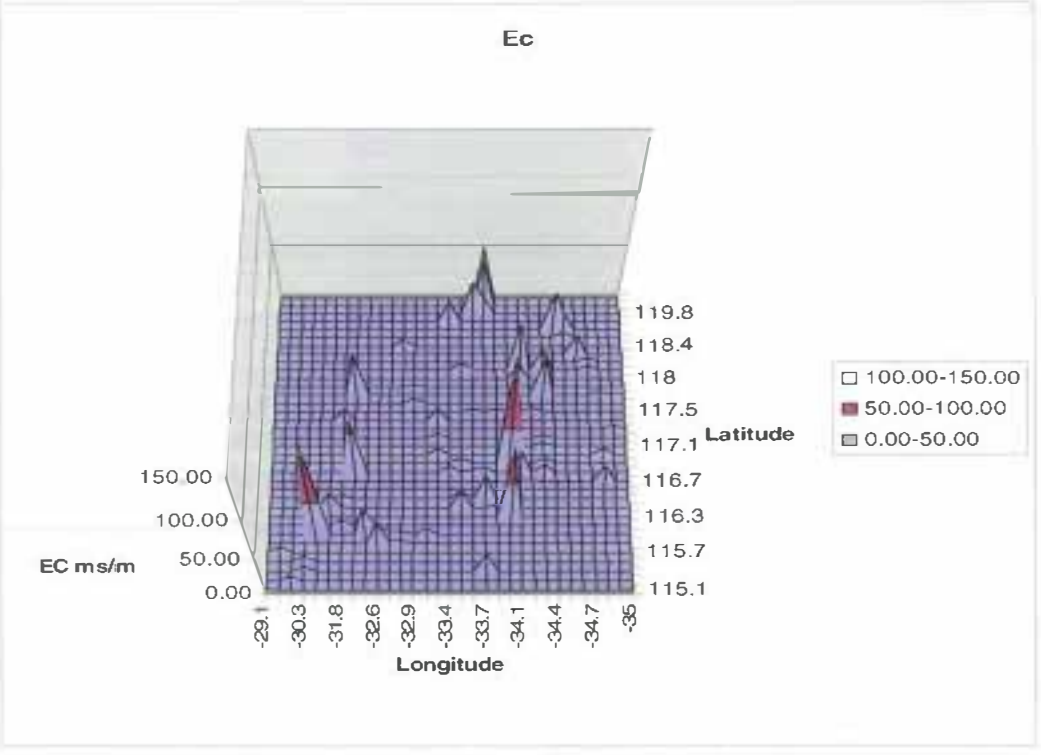
Grey deep sandy duplex (Normal data) – OC



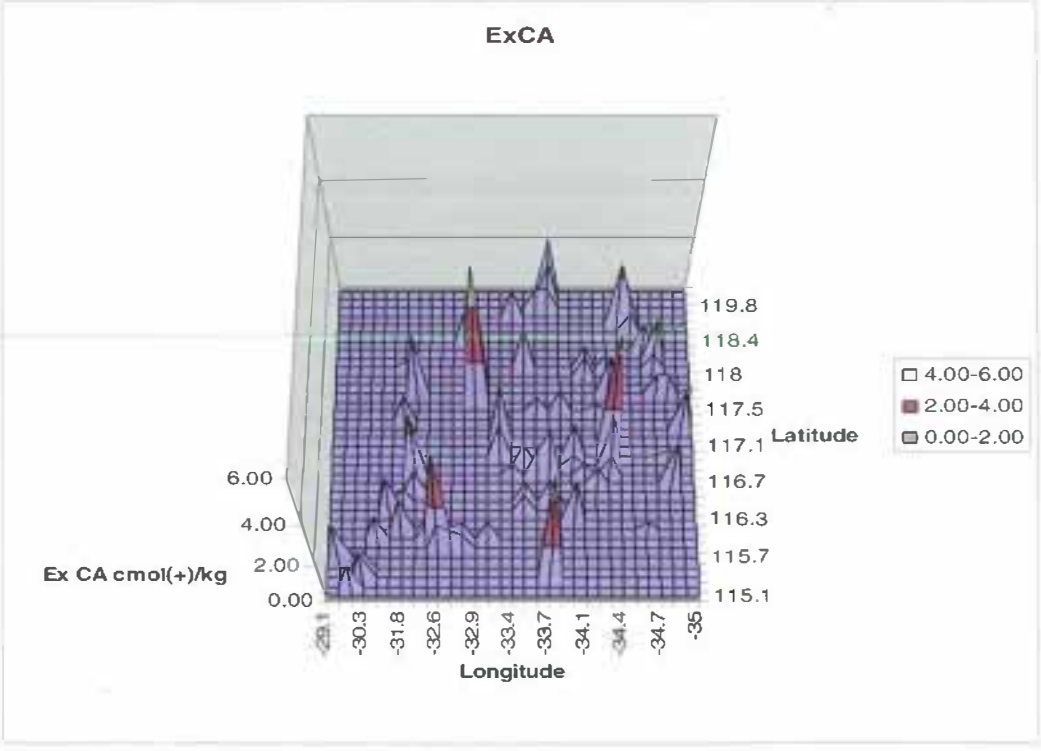
Grey deep sandy duplex (Normal data) – pH



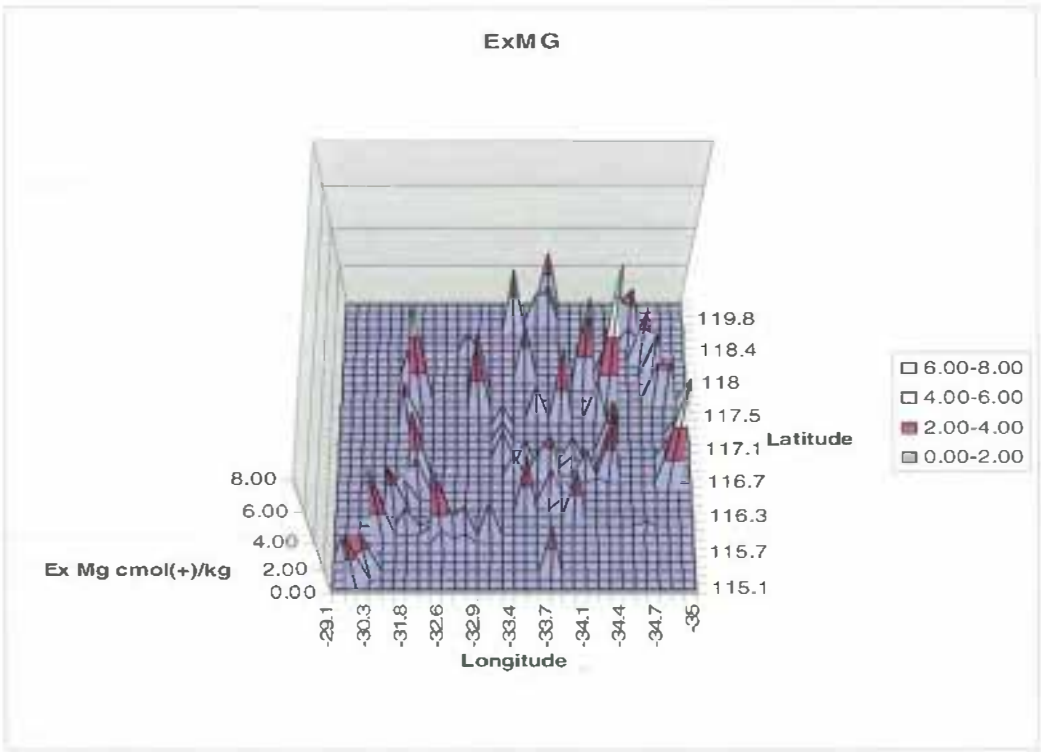
Grey deep sandy duplex (Normal data) – Clay



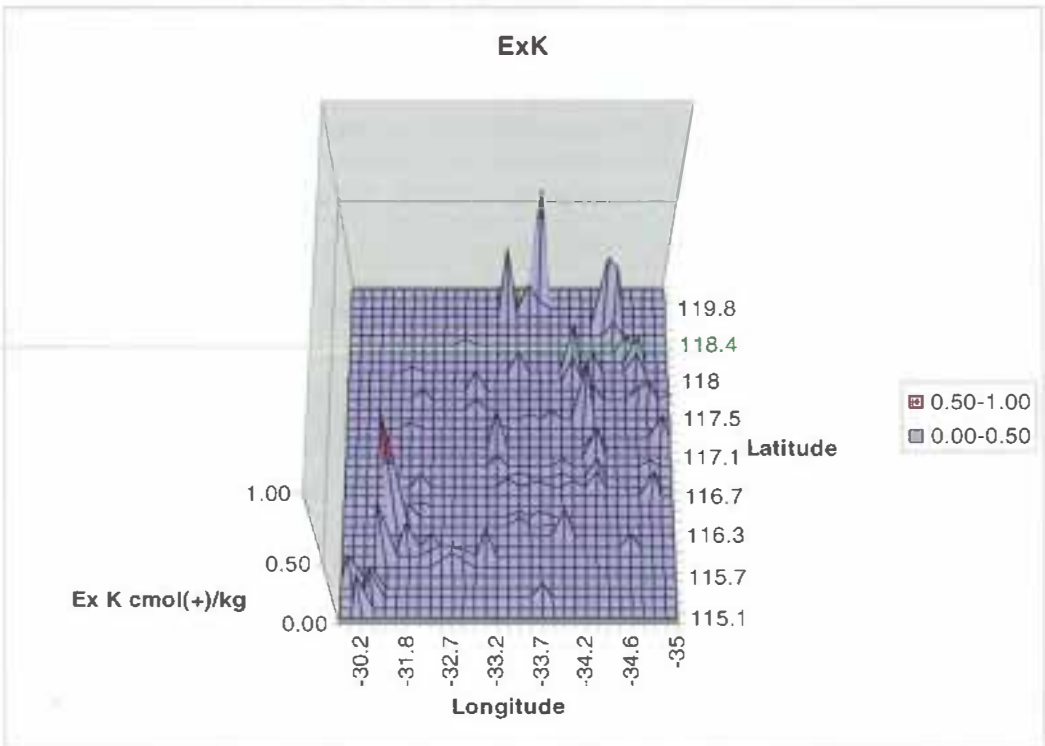
Grey deep sandy duplex (Normal data) – Ec



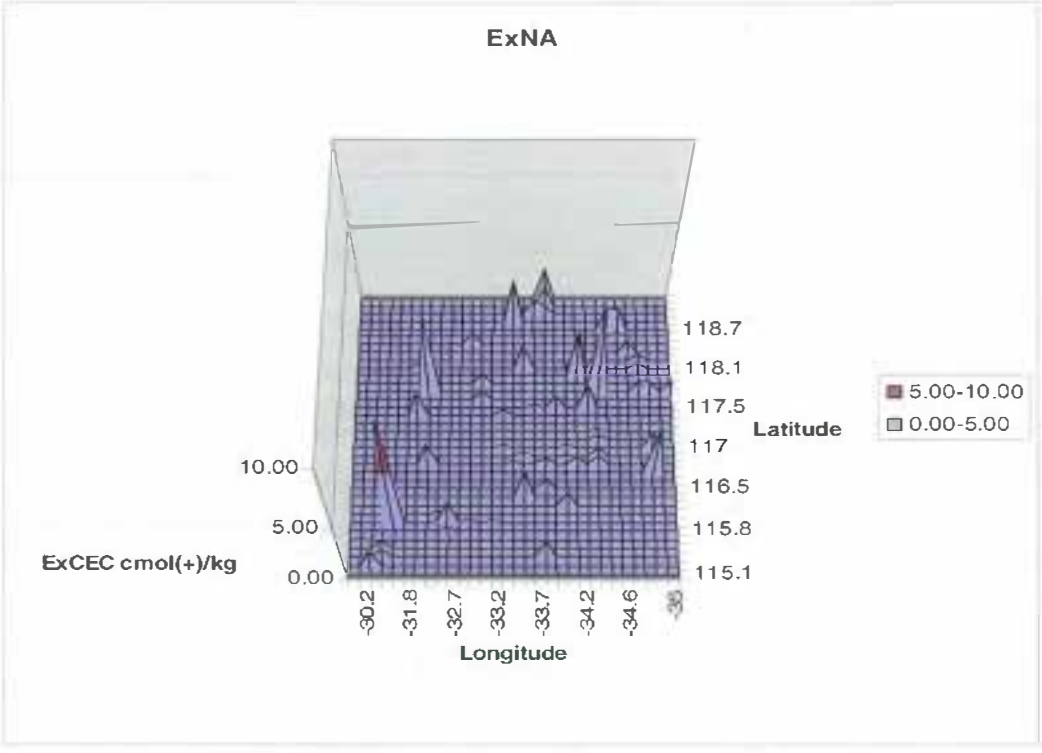
Grey deep sandy duplex (Normal data) – ExCA



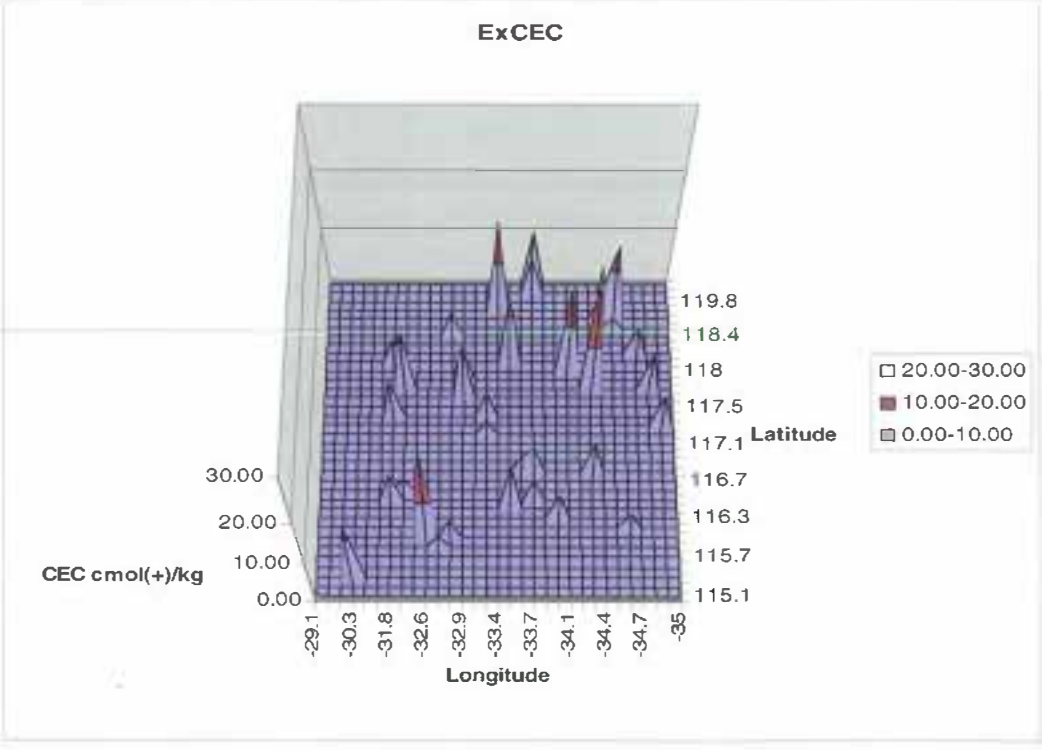
Grey deep sandy duplex (Normal data) – ExMG



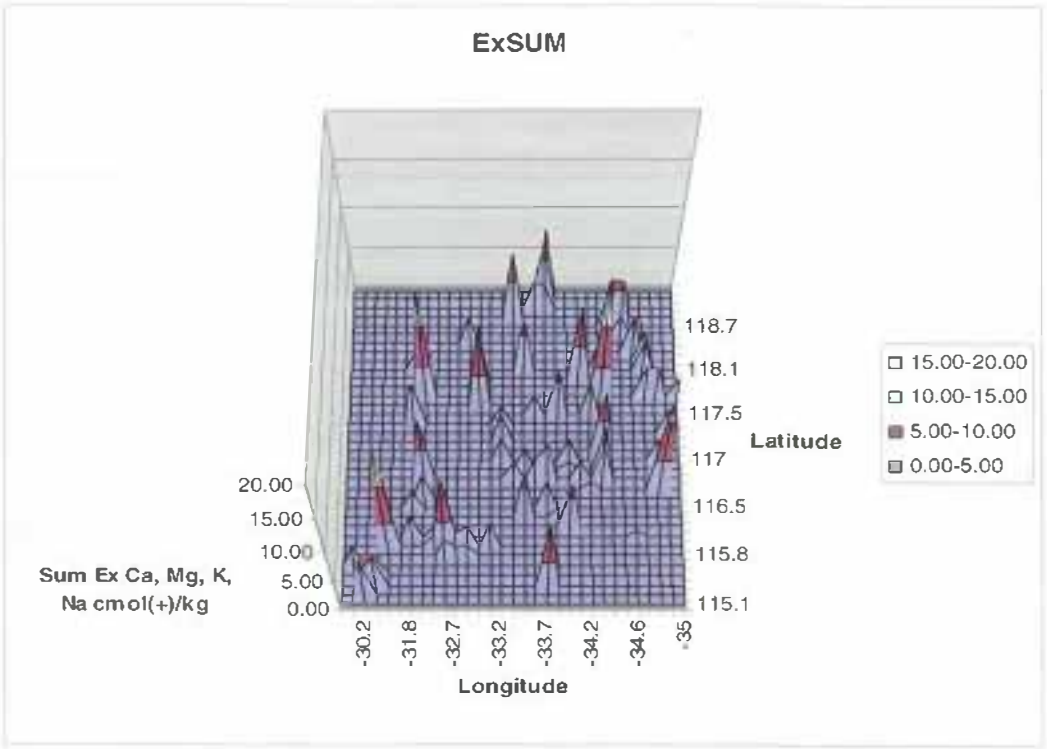
Grey deep sandy duplex (Normal data) – ExK



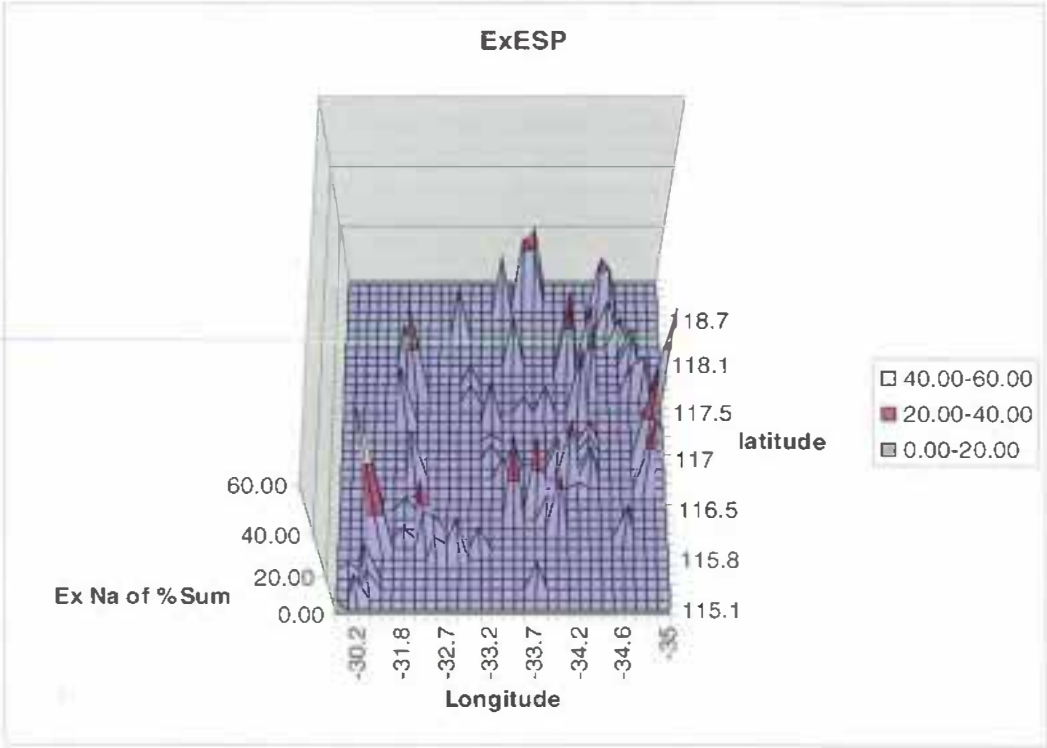
Grey deep sandy duplex (Normal data) – ExNA



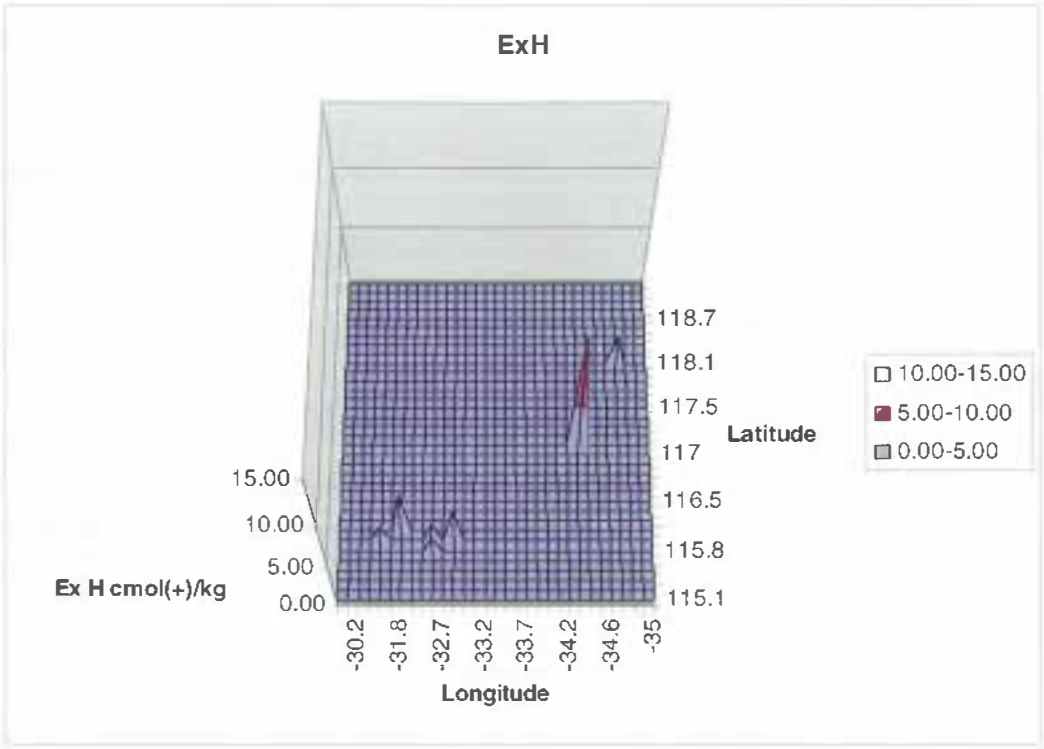
Grey deep sandy duplex (Normal data) – ExCEC



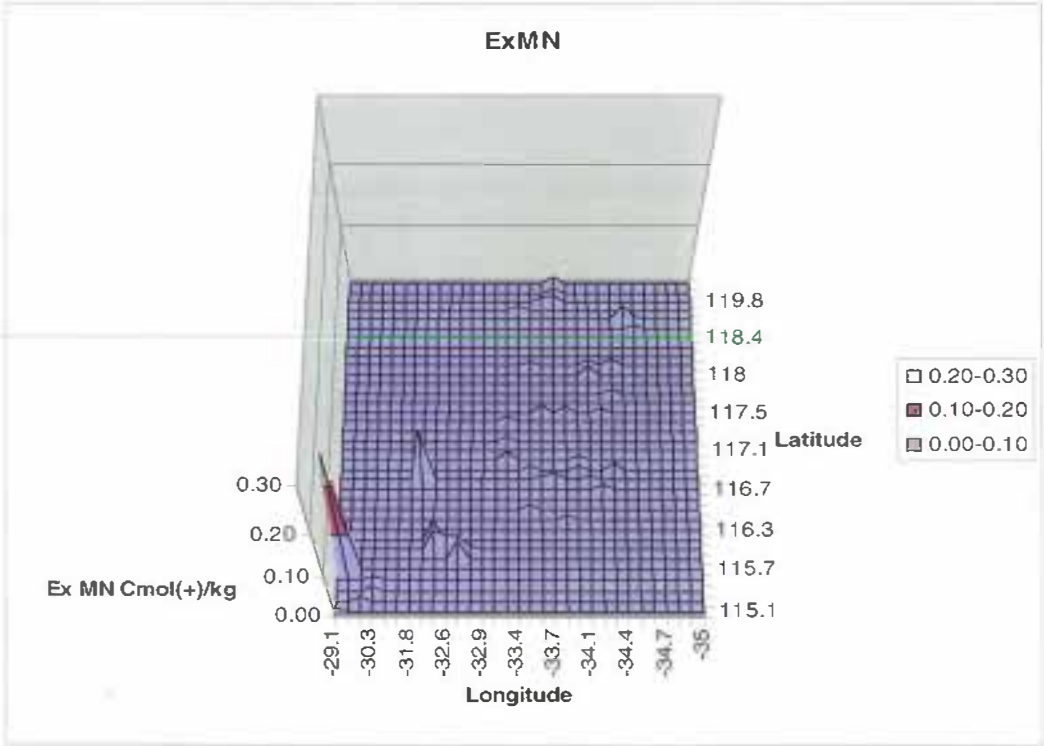
Grey deep sandy duplex (Normal data) – ExSUM



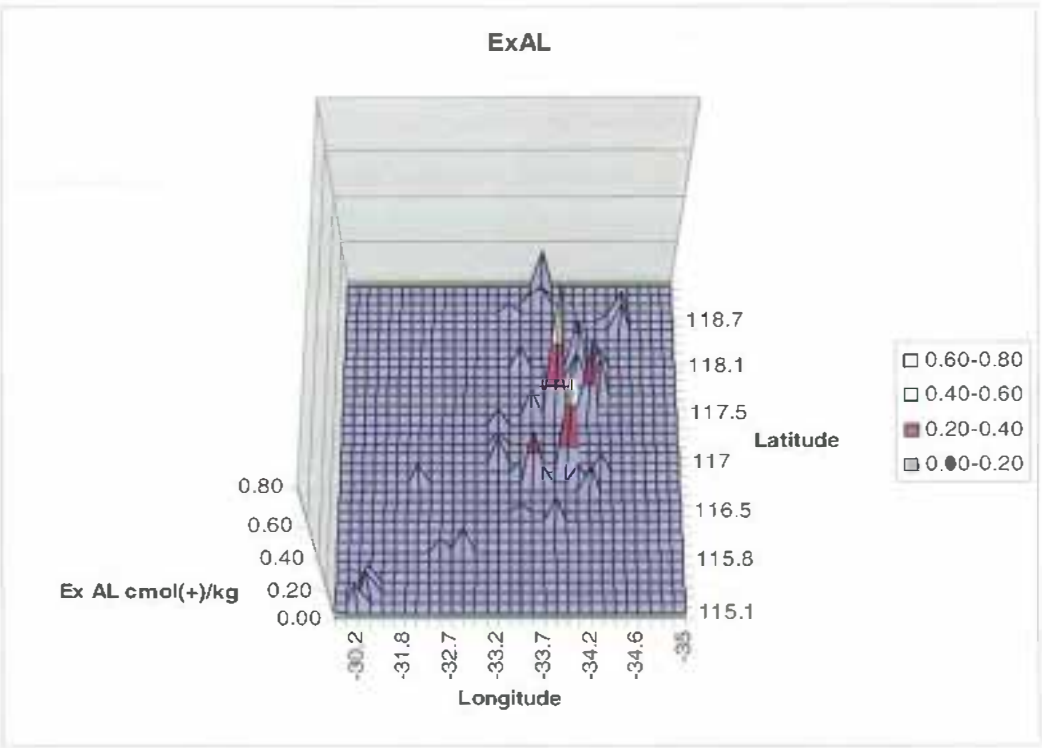
Grey deep sandy duplex (Normal data) – ExESP



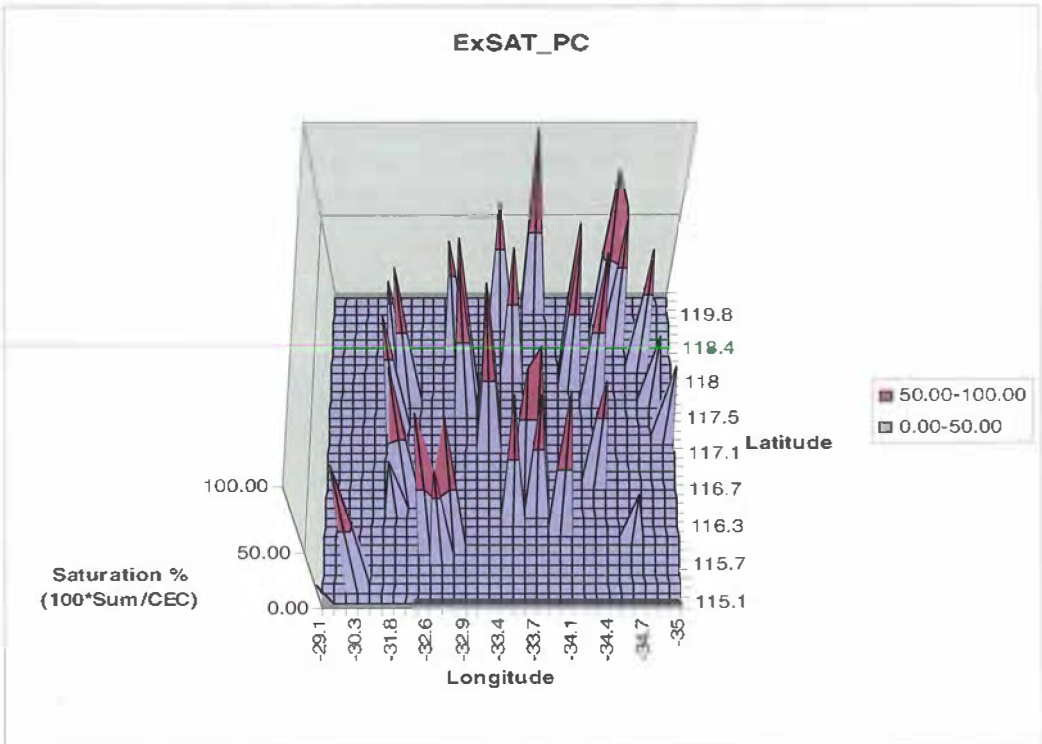
Grey deep sandy duplex (Normal data) – ExH



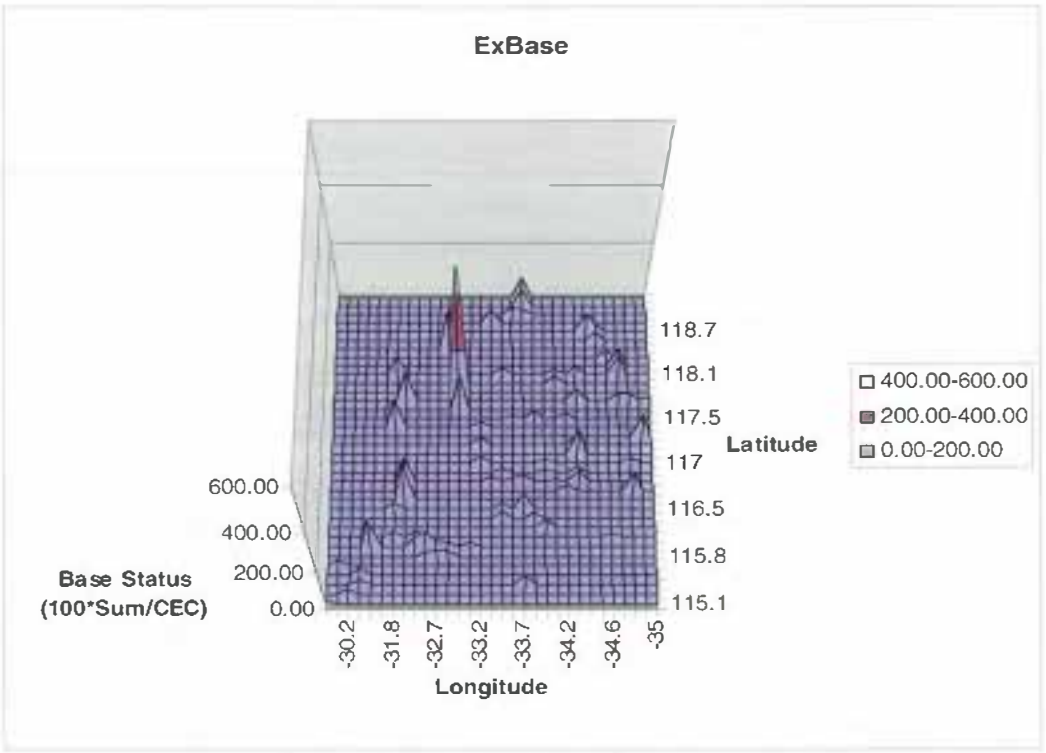
Grey deep sandy duplex (Normal data) – ExMN



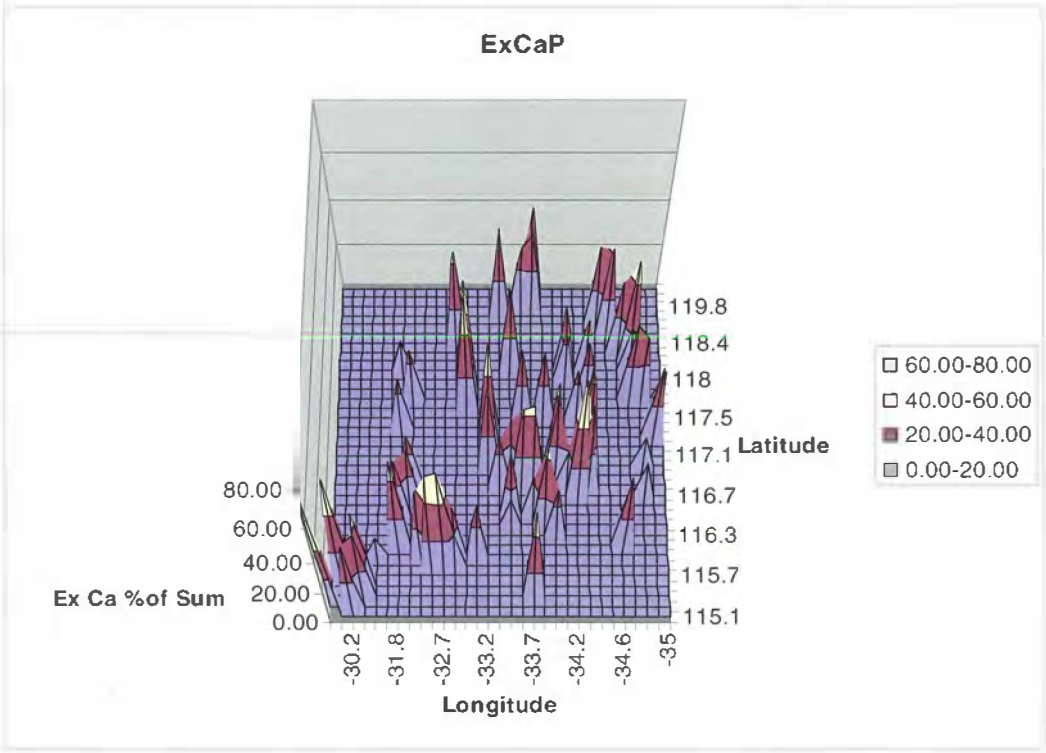
Grey deep sandy duplex (Normal data) – ExAL



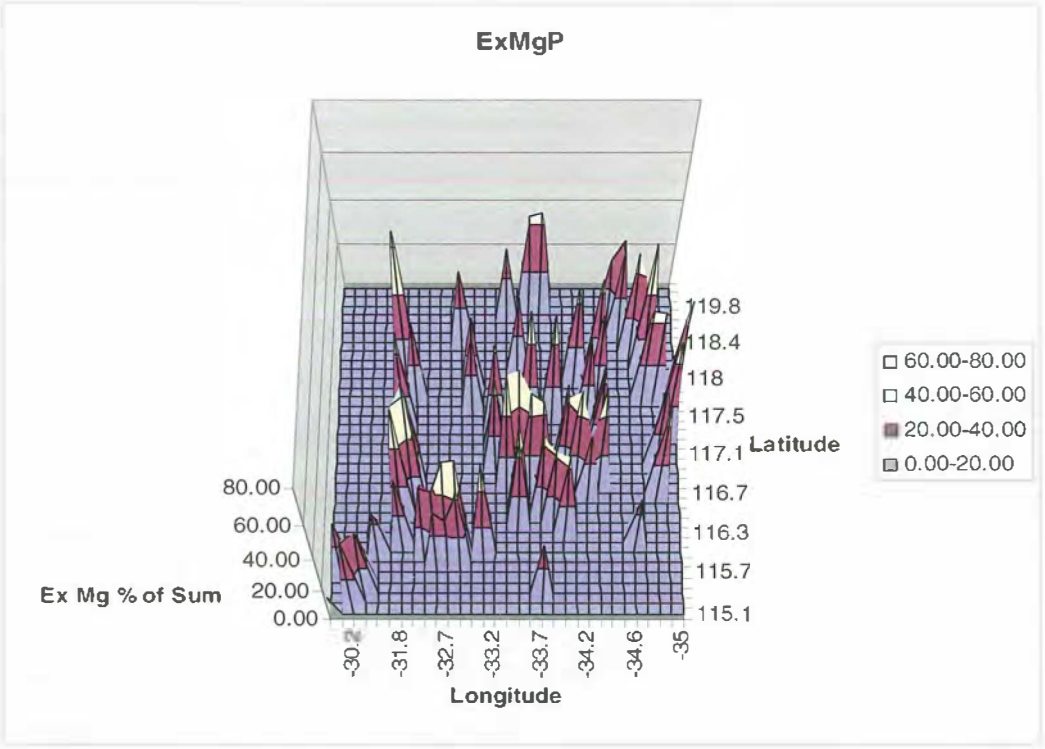
Grey deep sandy duplex (Normal data) – ExSAT_PC



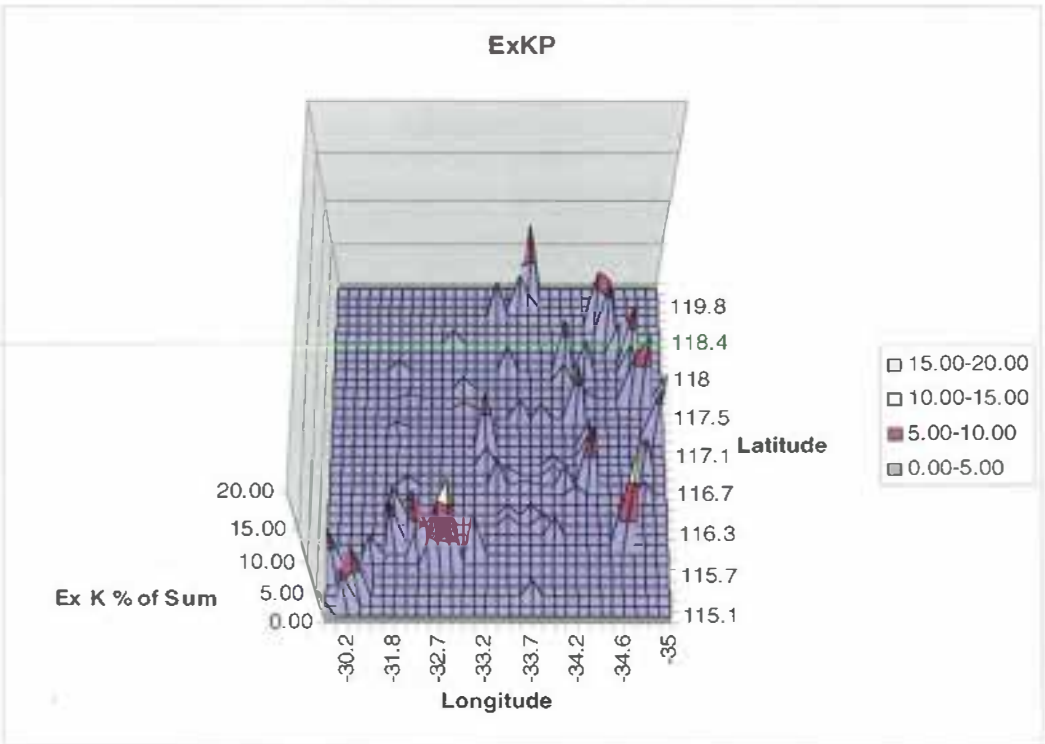
Grey deep sandy duplex (Normal data) – ExBASE



Grey deep sandy duplex (Normal data) – ExCaP

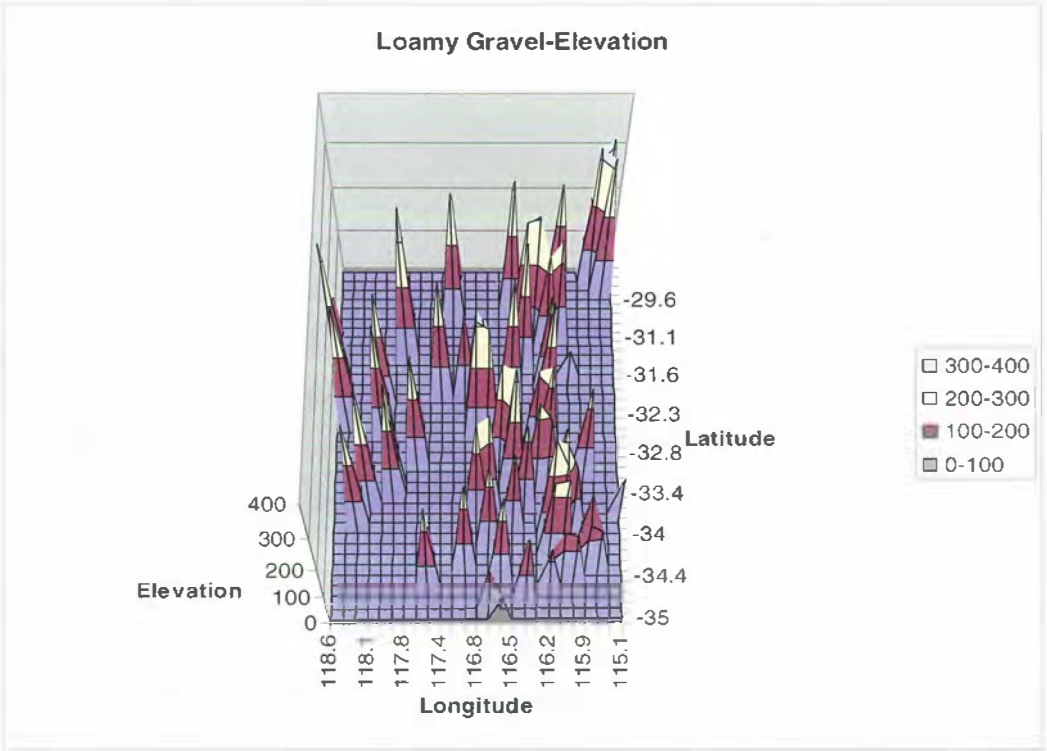


Grey deep sandy duplex (Normal data) – ExMgP

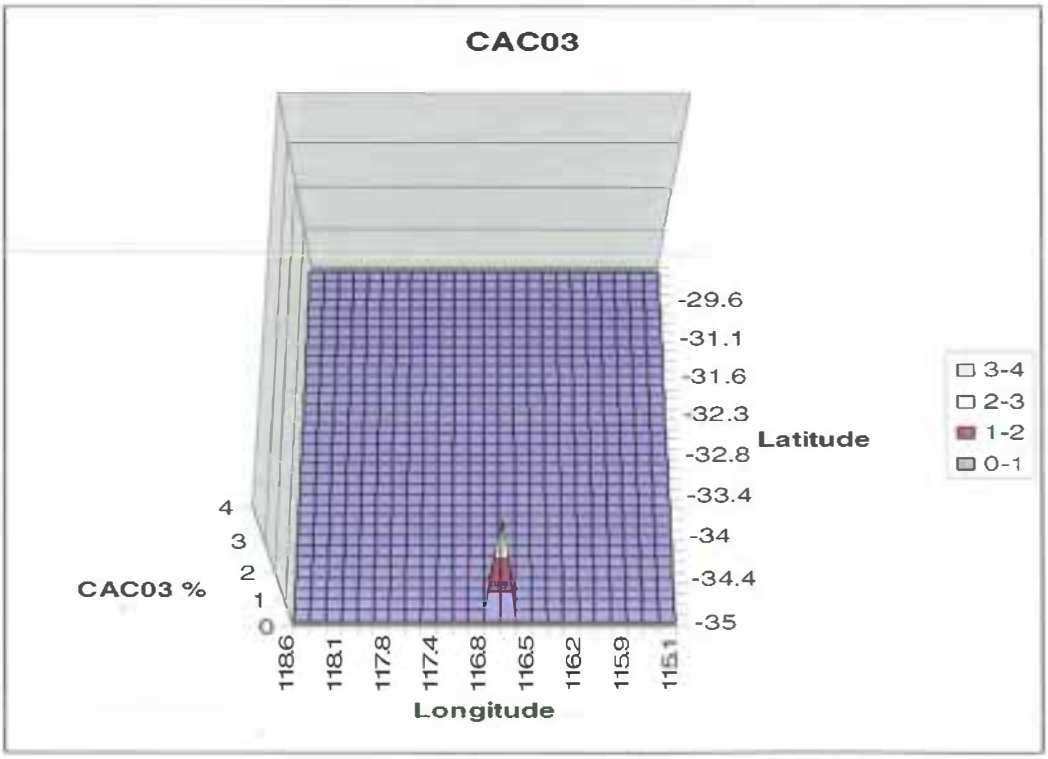


Grey deep sandy duplex (Normal data) – ExKP

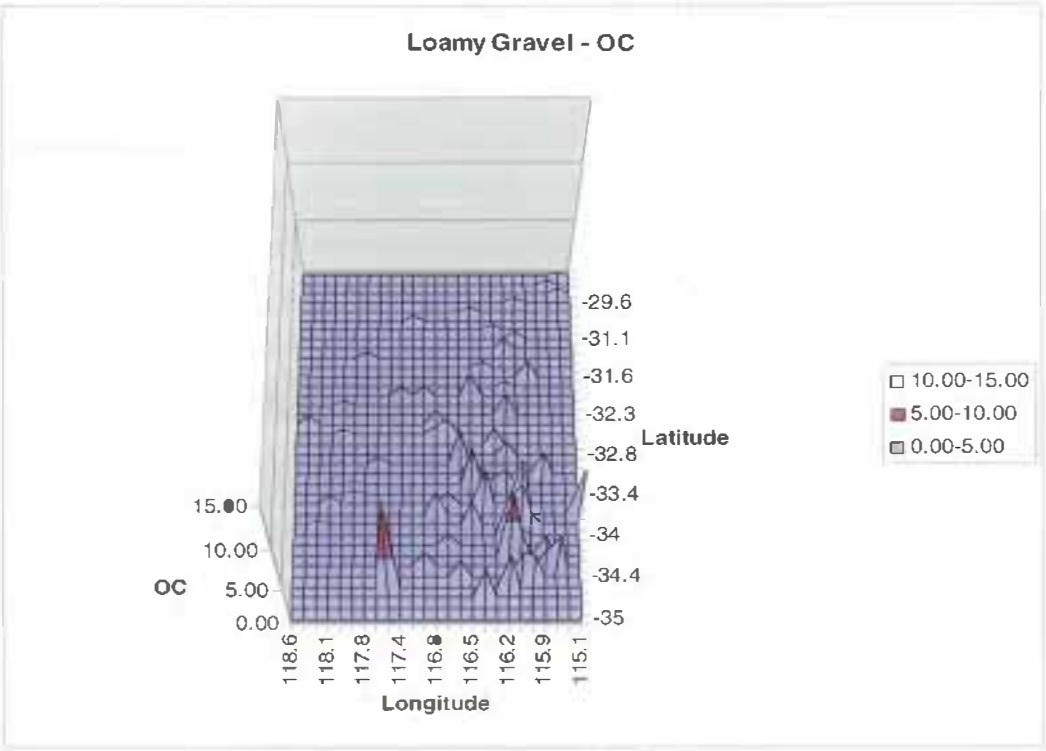
Soil 2 - Loamy gravel



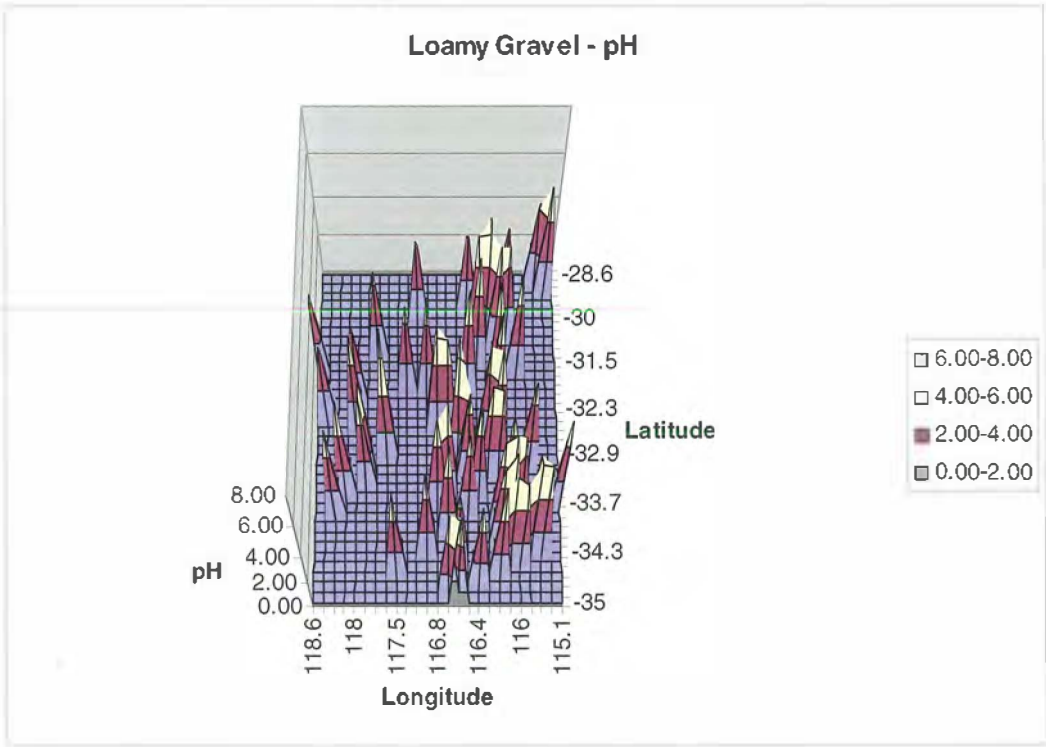
Loamy gravel (Normal data) – Elevation



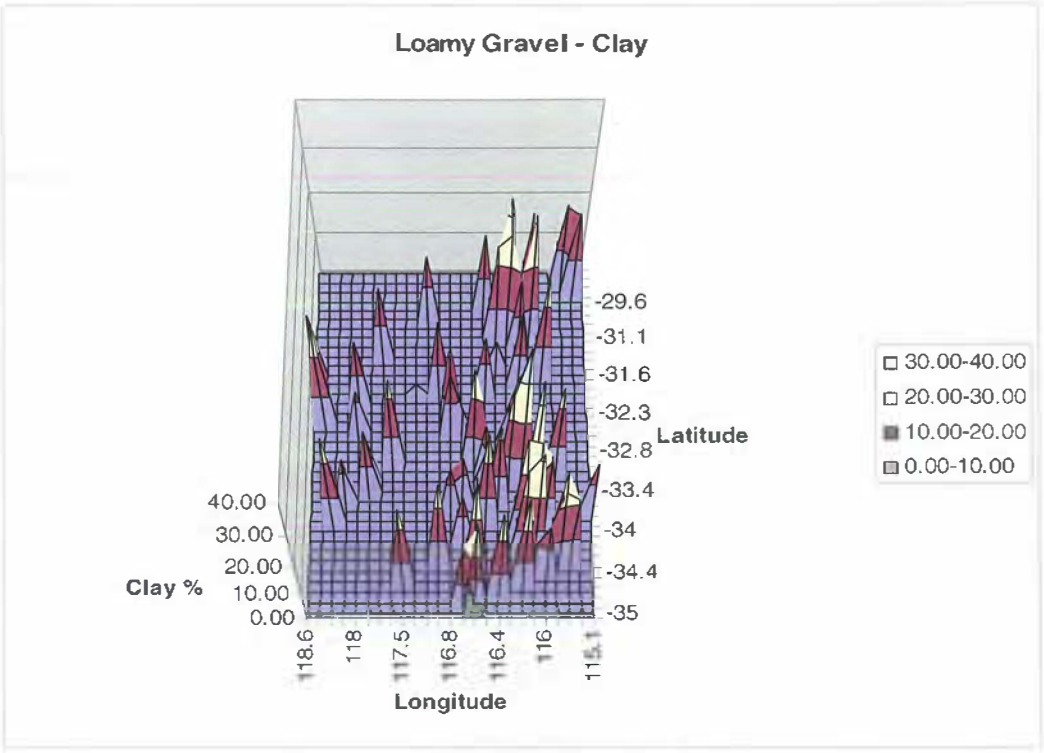
Loamy gravel (Normal data) – CAC03



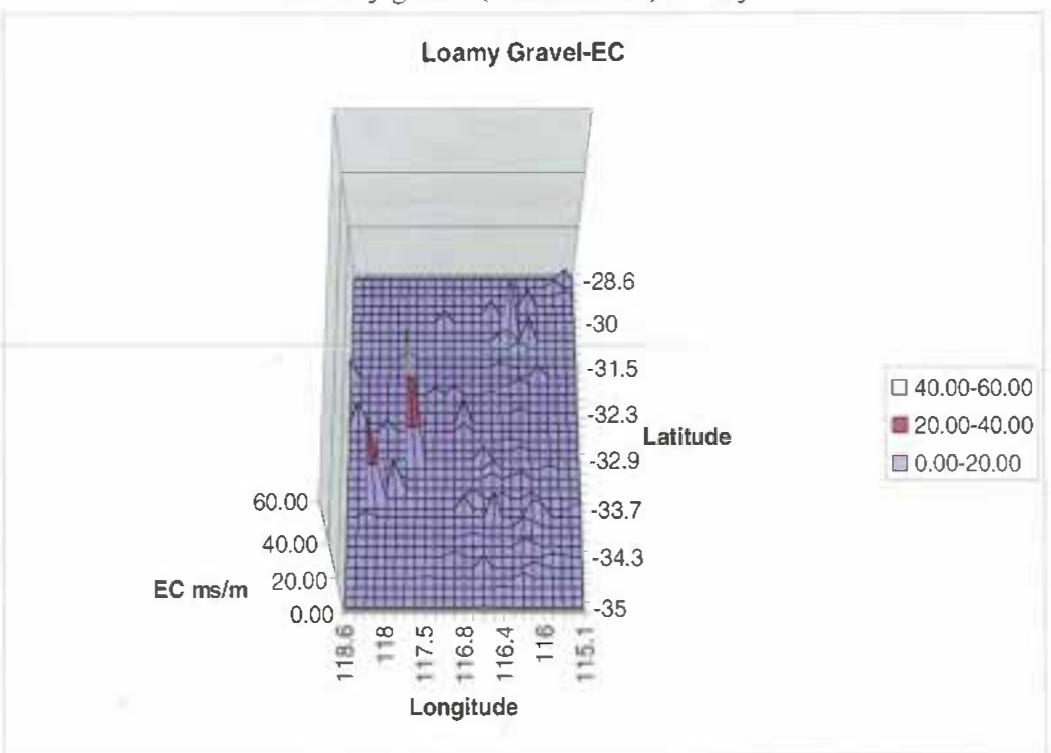
Loamy gravel (Normal data) – OC %



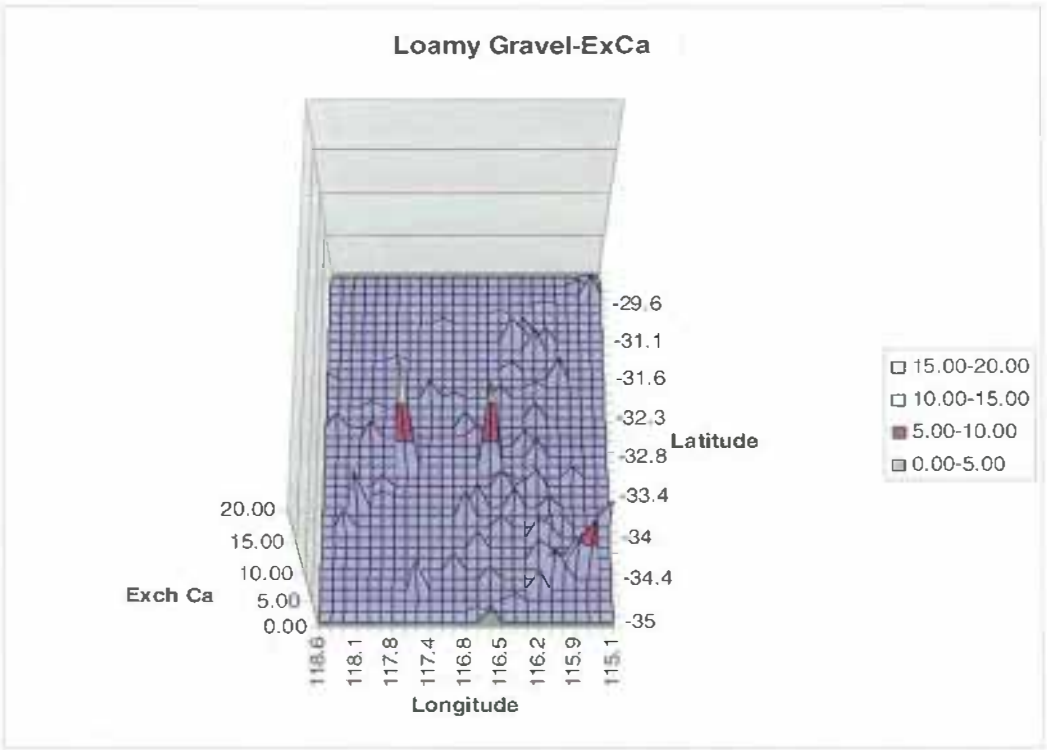
Loamy gravel (Normal data) – pH



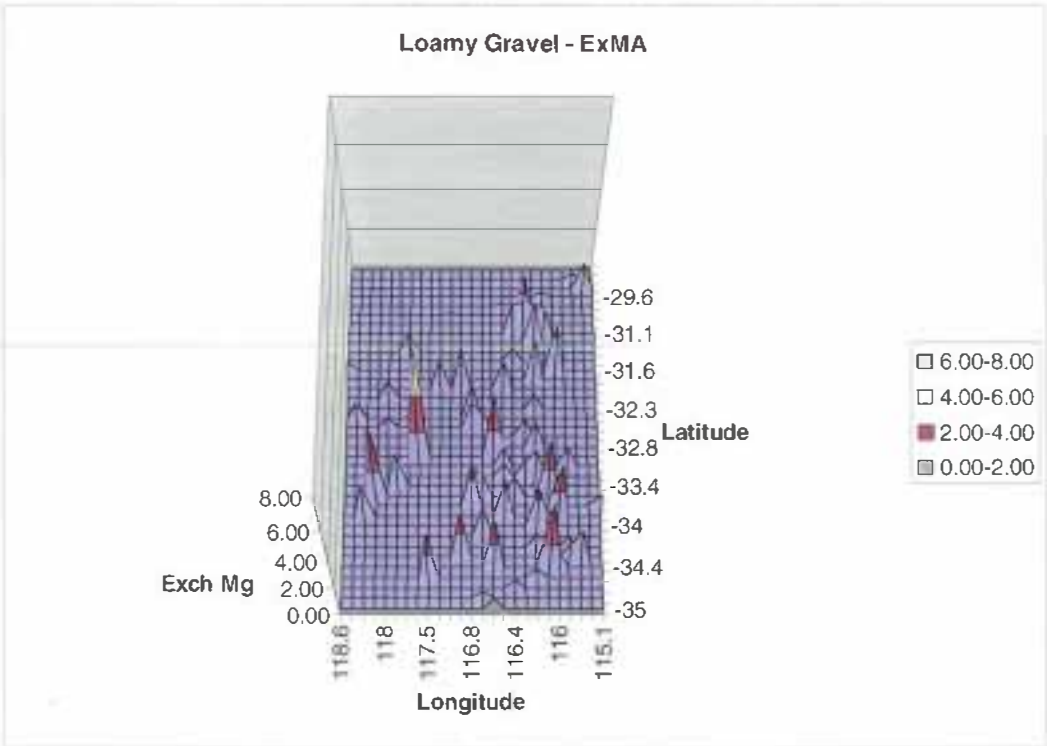
Loamy gravel (Normal data) – Clay



Loamy gravel (Normal data) – EC

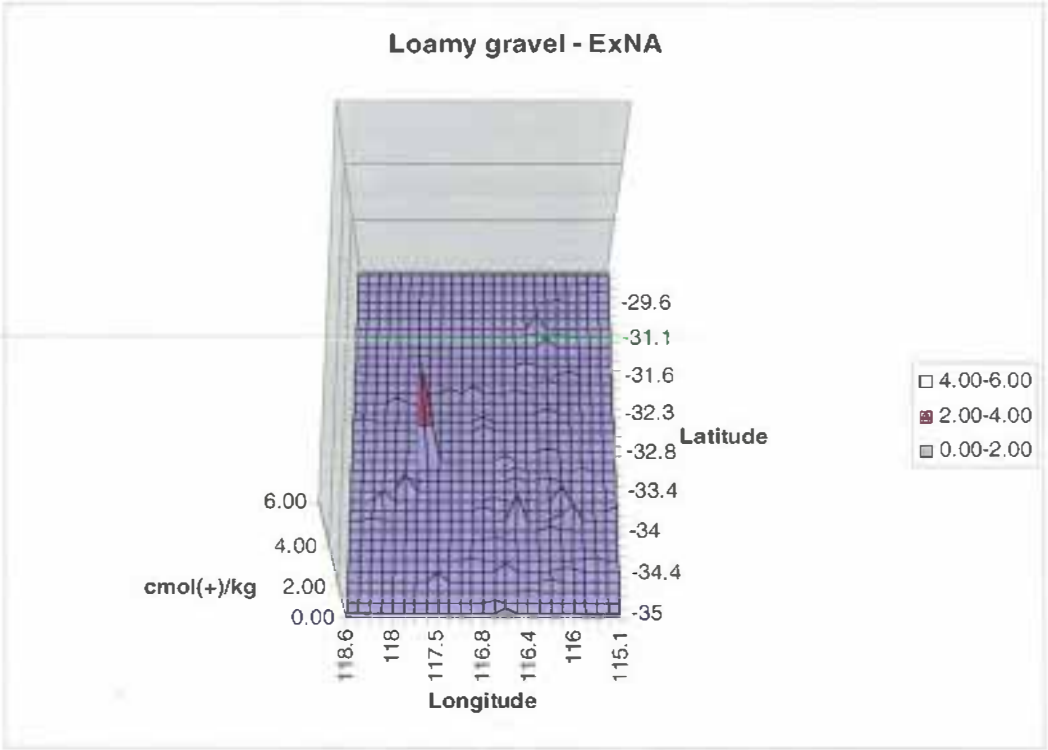


Loamy gravel (Normal data) – ExCA

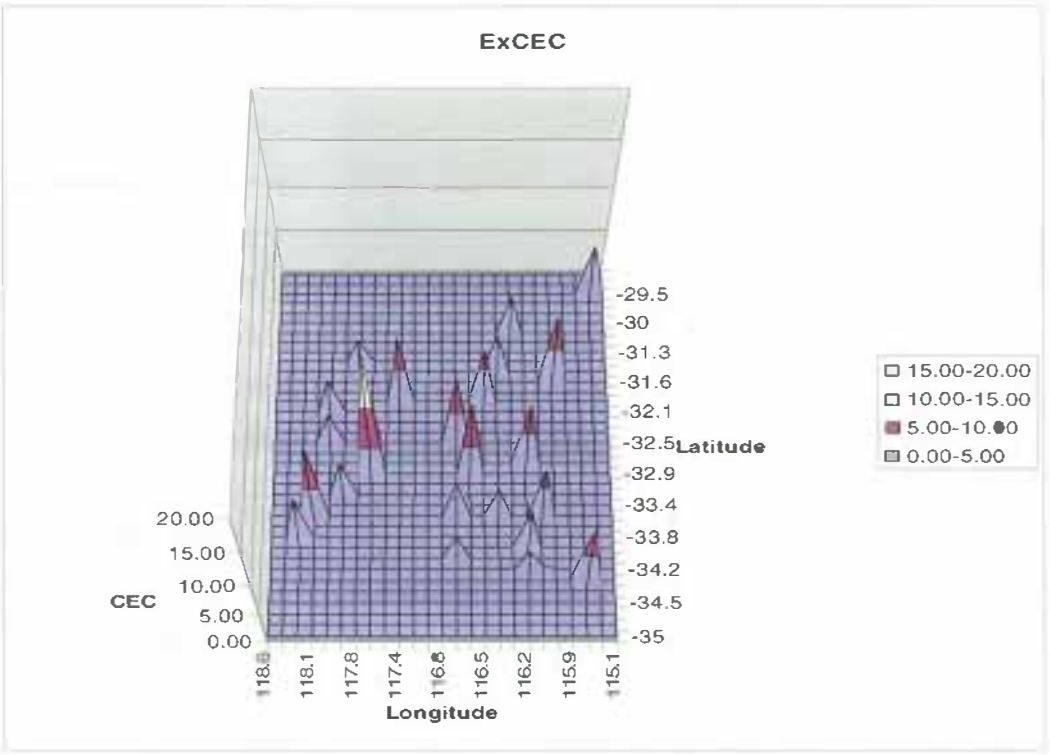


Loamy gravel (Normal data) – ExMA

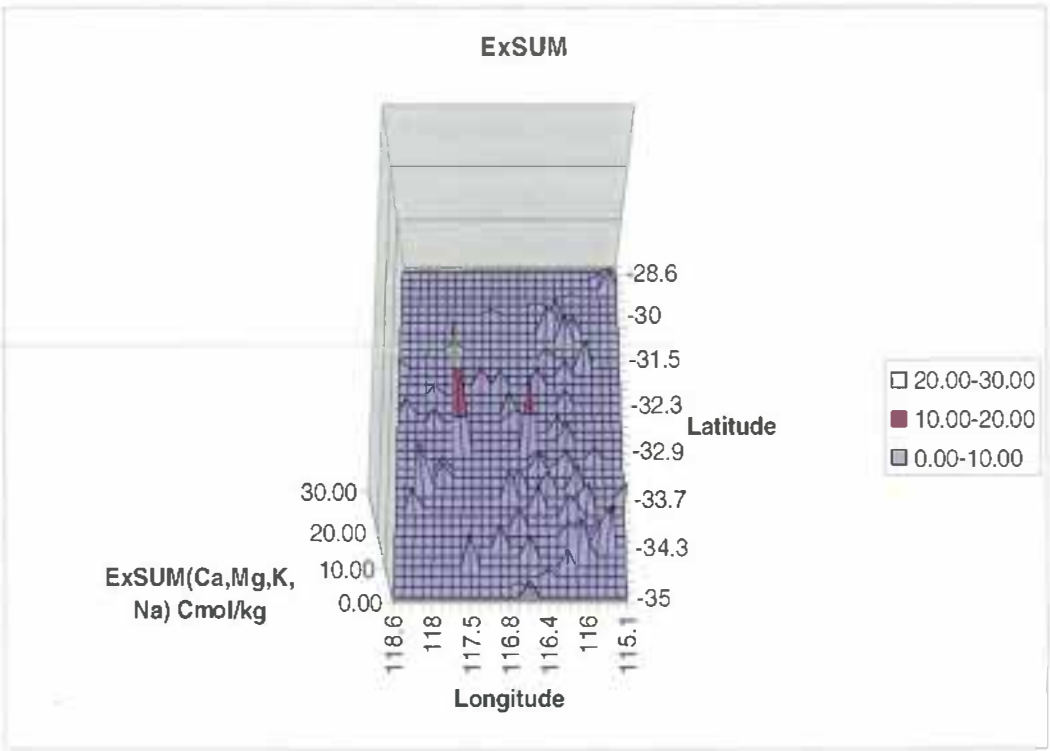
Loamy gravel (Normal data) – ExK



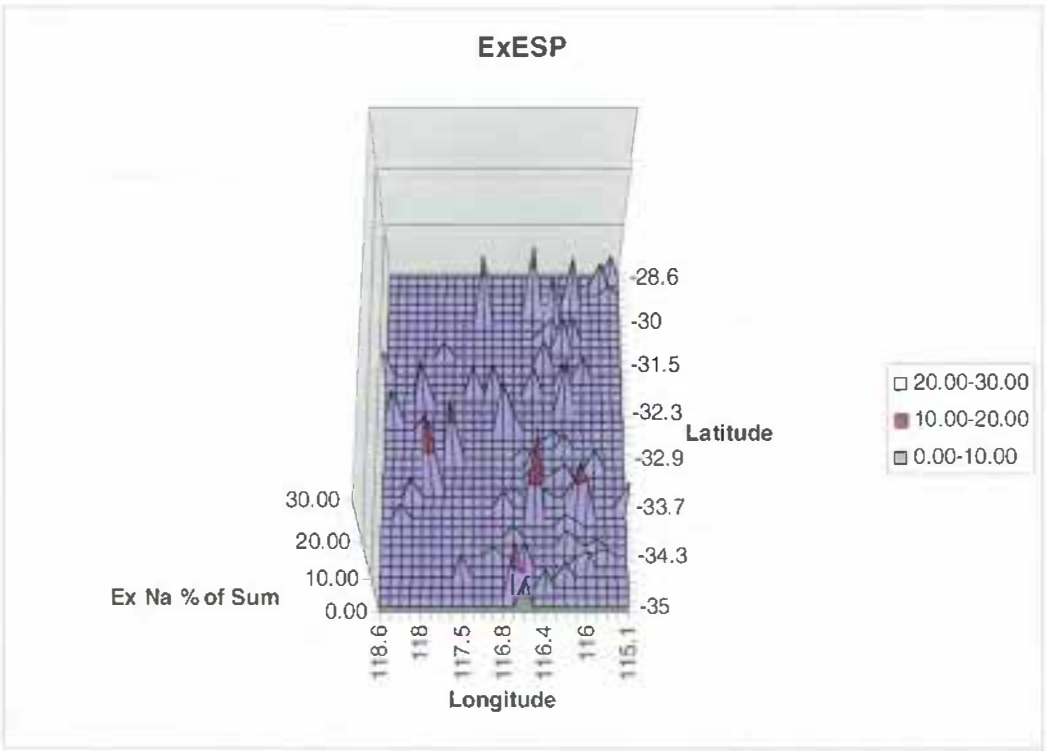
Loamy gravel (Normal data) – ExNA



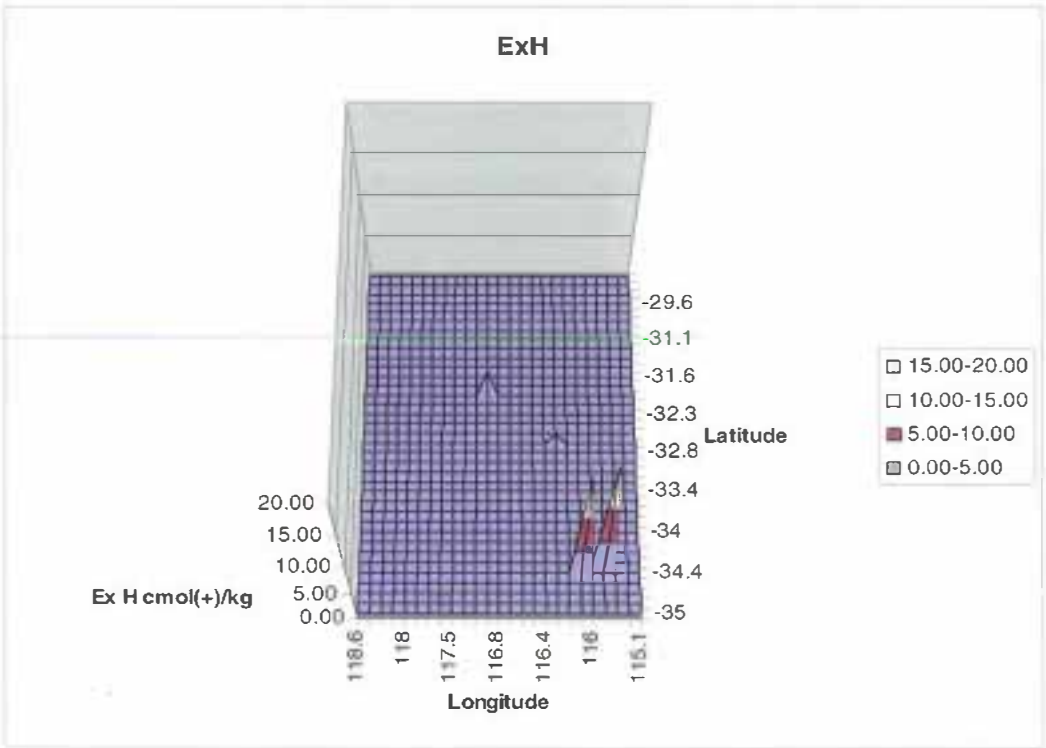
Loamy gravel (Normal data) – ExCEC



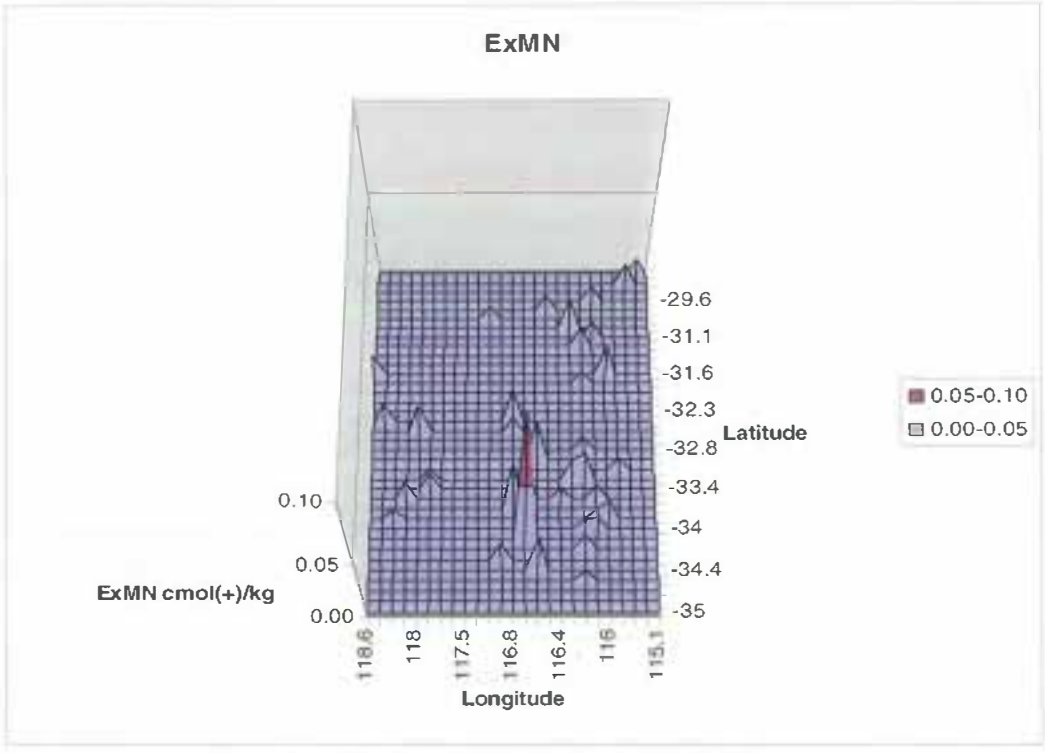
Loamy gravel (Normal data) – ExSUM



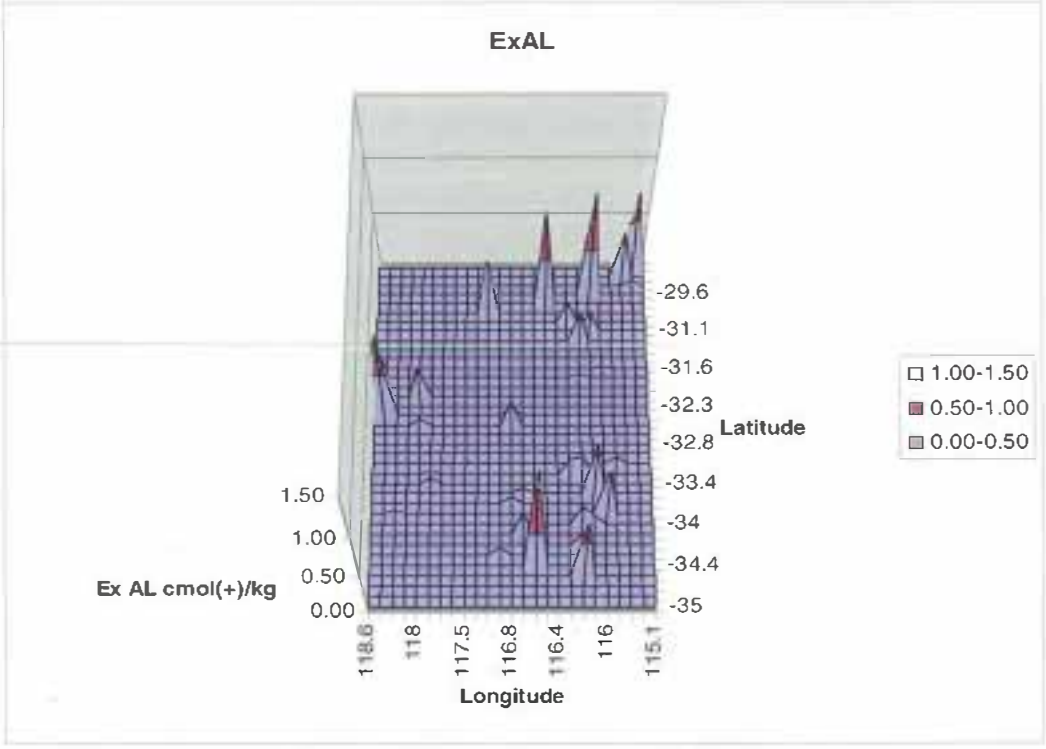
Loamy gravel (Normal data) – ExESP



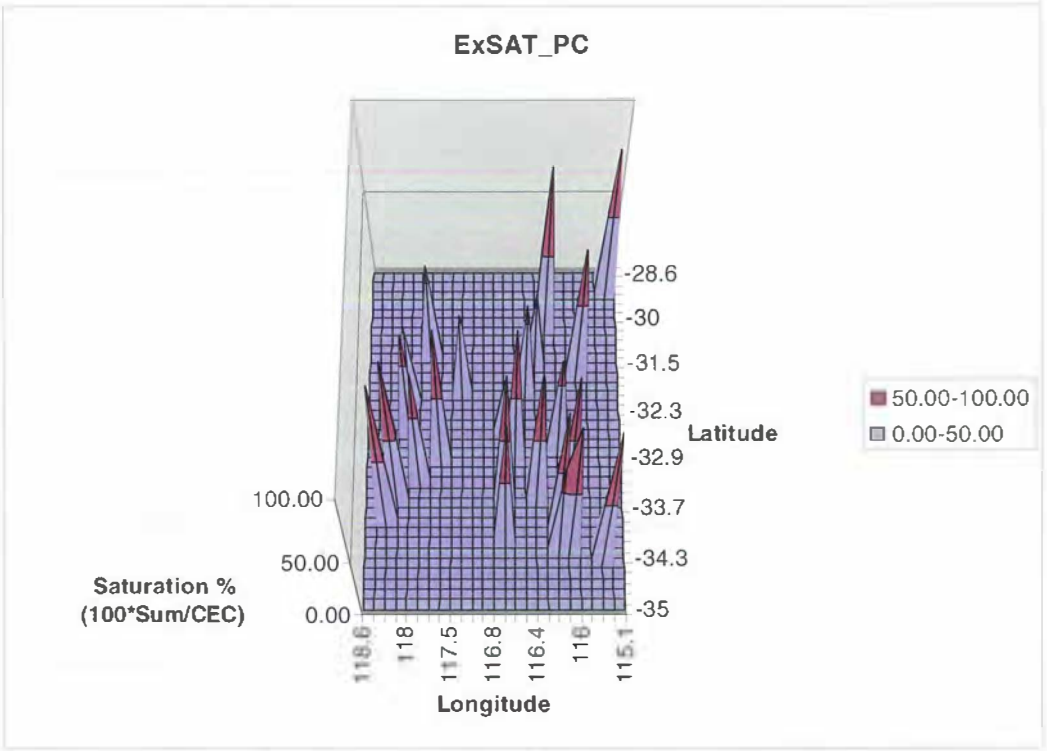
Loamy gravel (Normal data) – ExH



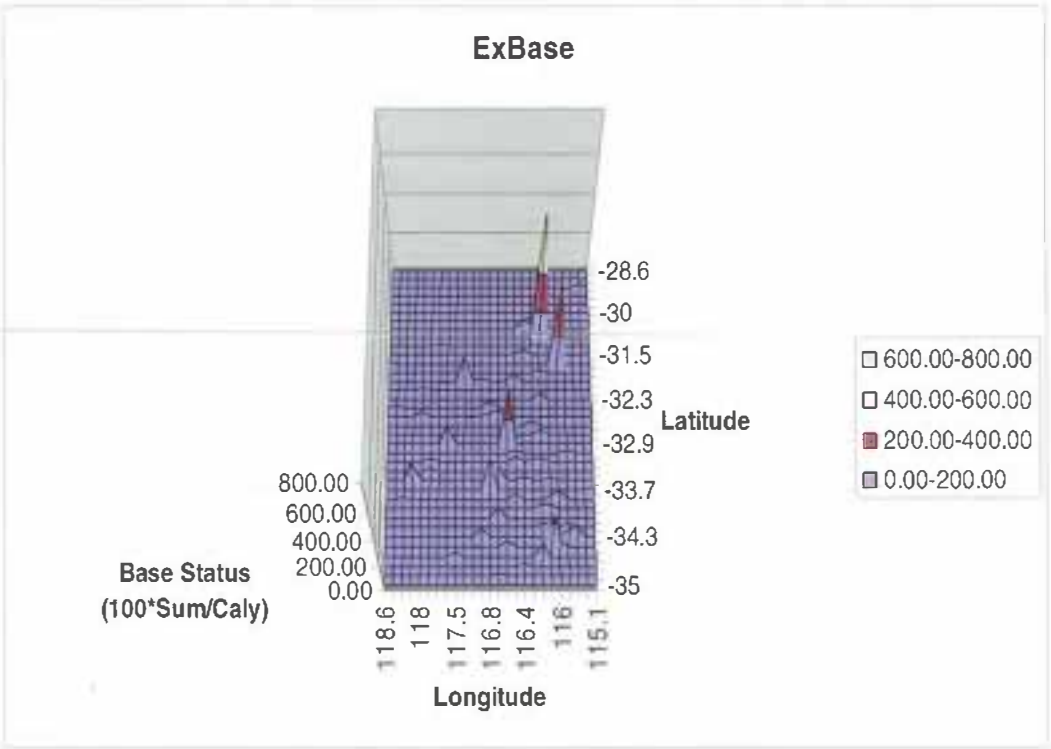
Loamy gravel (Normal data) – ExMN



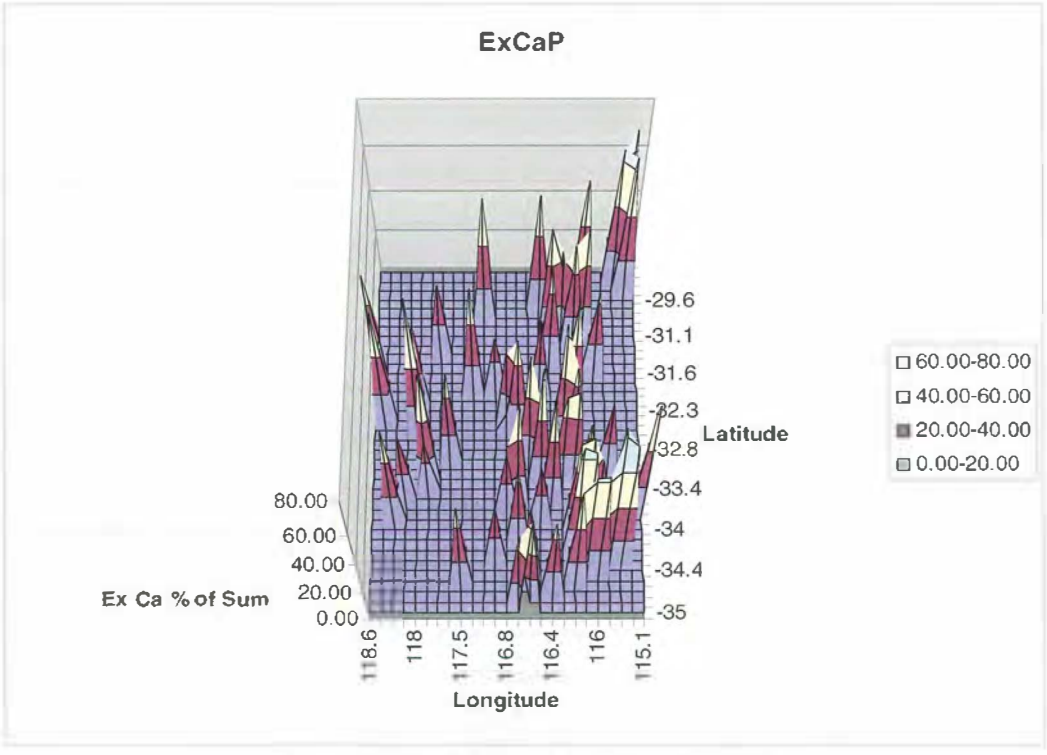
Loamy gravel (Normal data) – ExAL



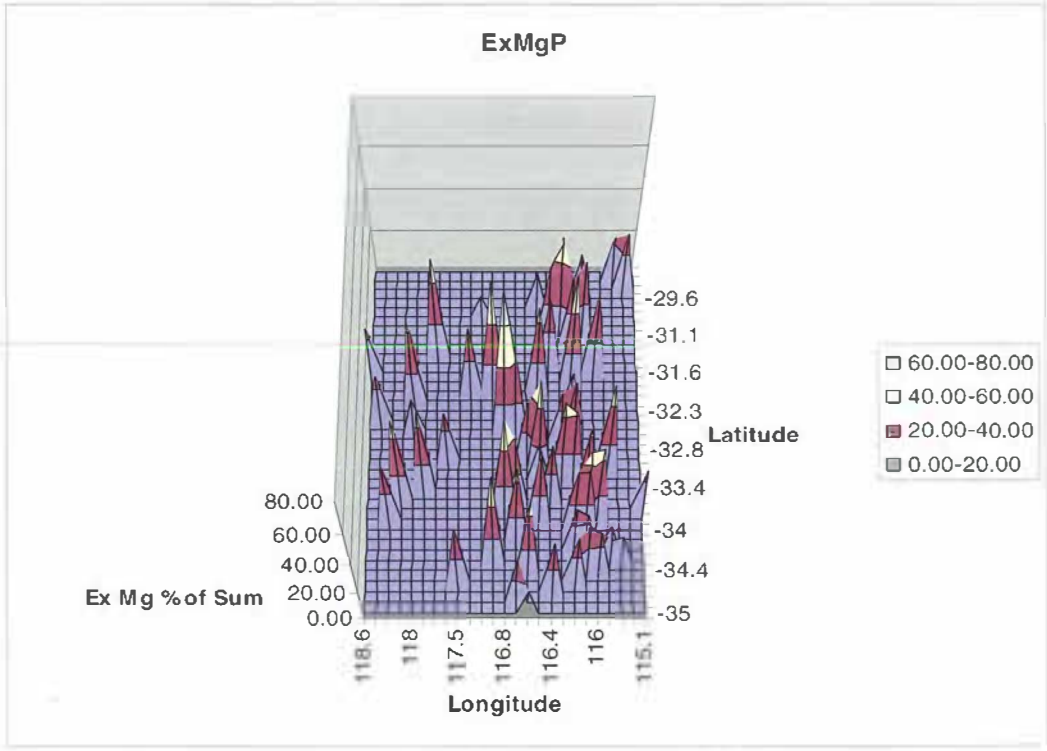
Loamy gravel (Normal data) – ExSAT_PC



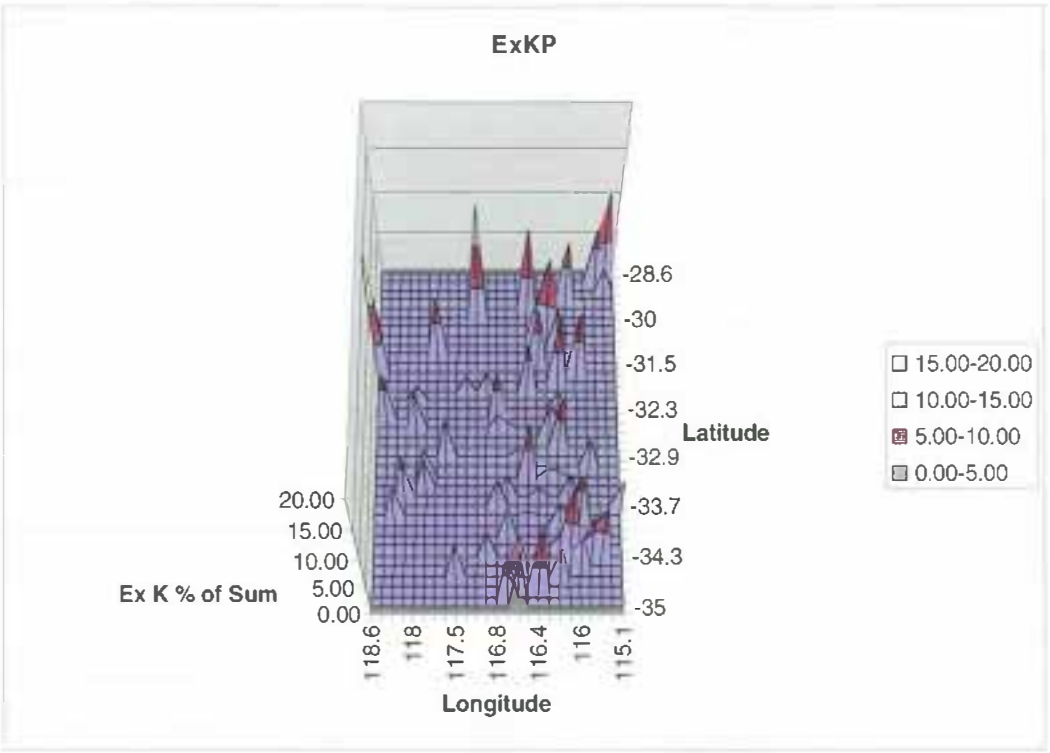
Loamy gravel (Normal data) – ExBASE



Loamy gravel (Normal data) – ExCaP

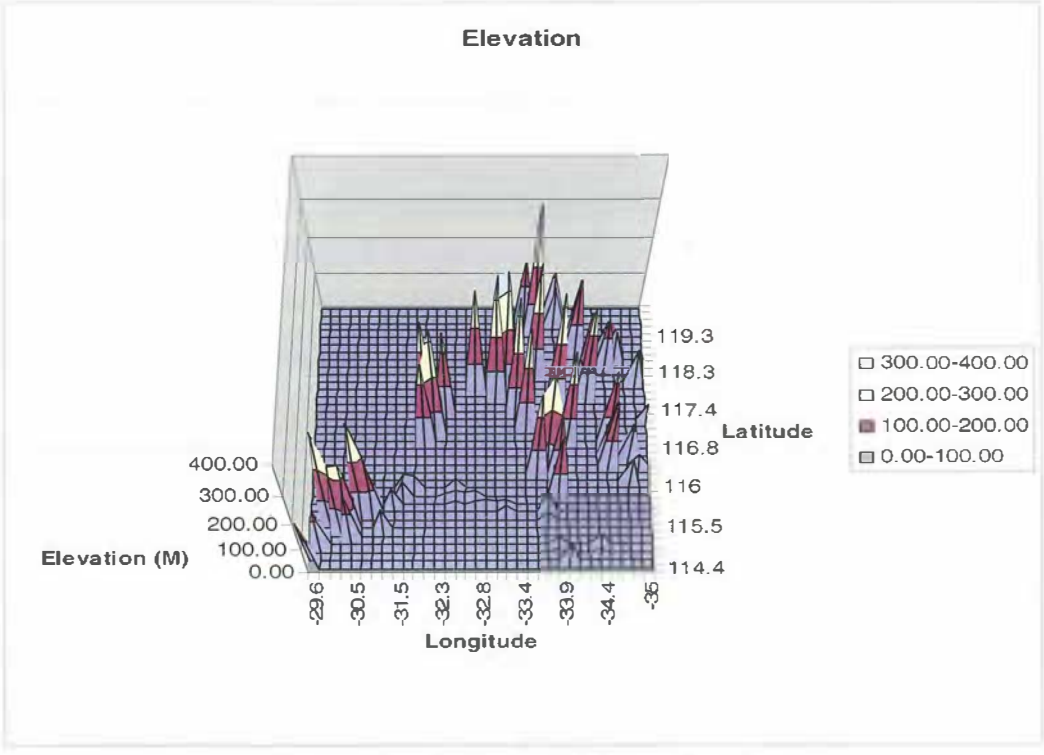


Loamy gravel (Normal data) – ExMgP

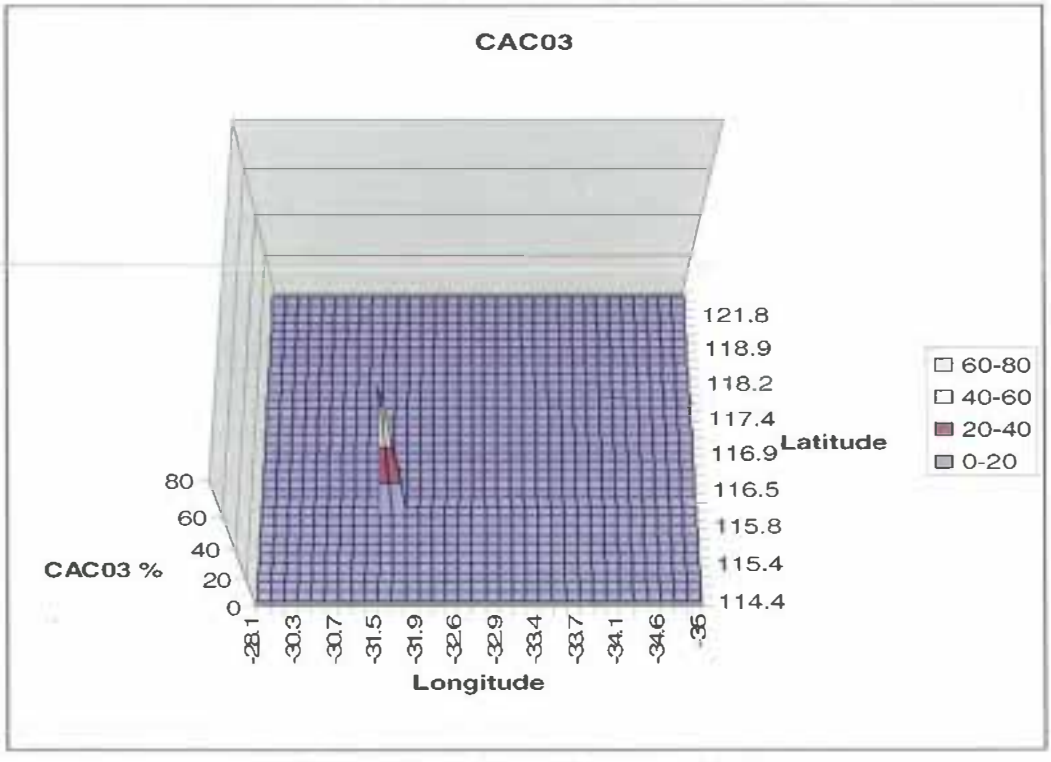


Loamy gravel (Normal data) – ExKP

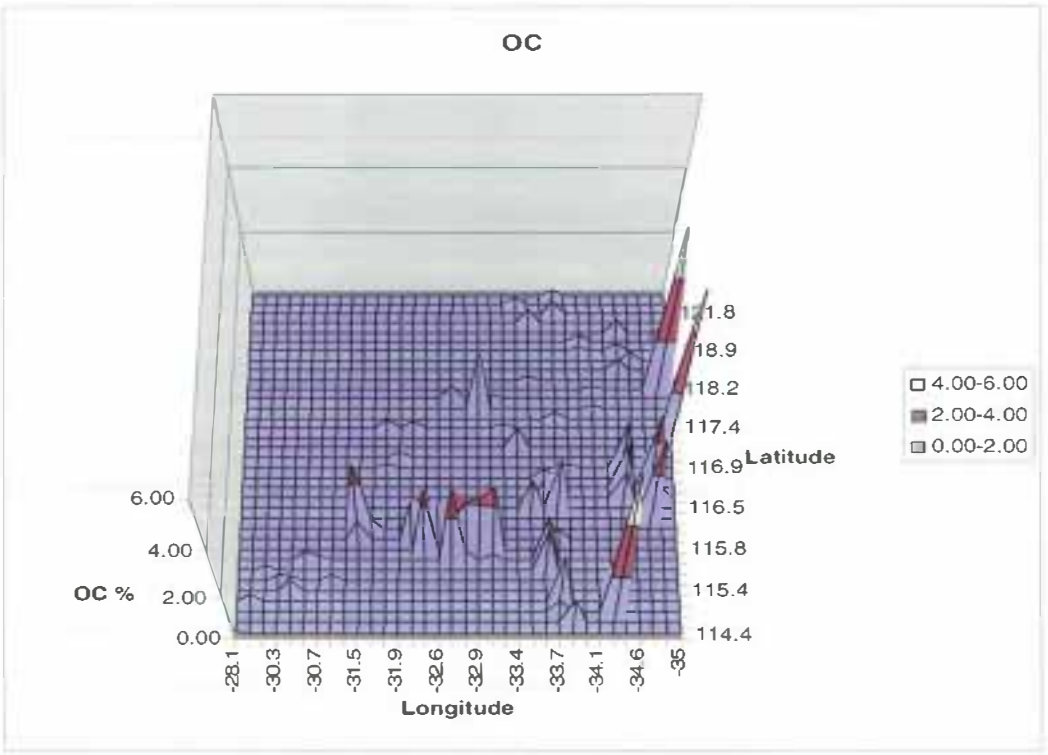
Soil 3 - Pale deep sand



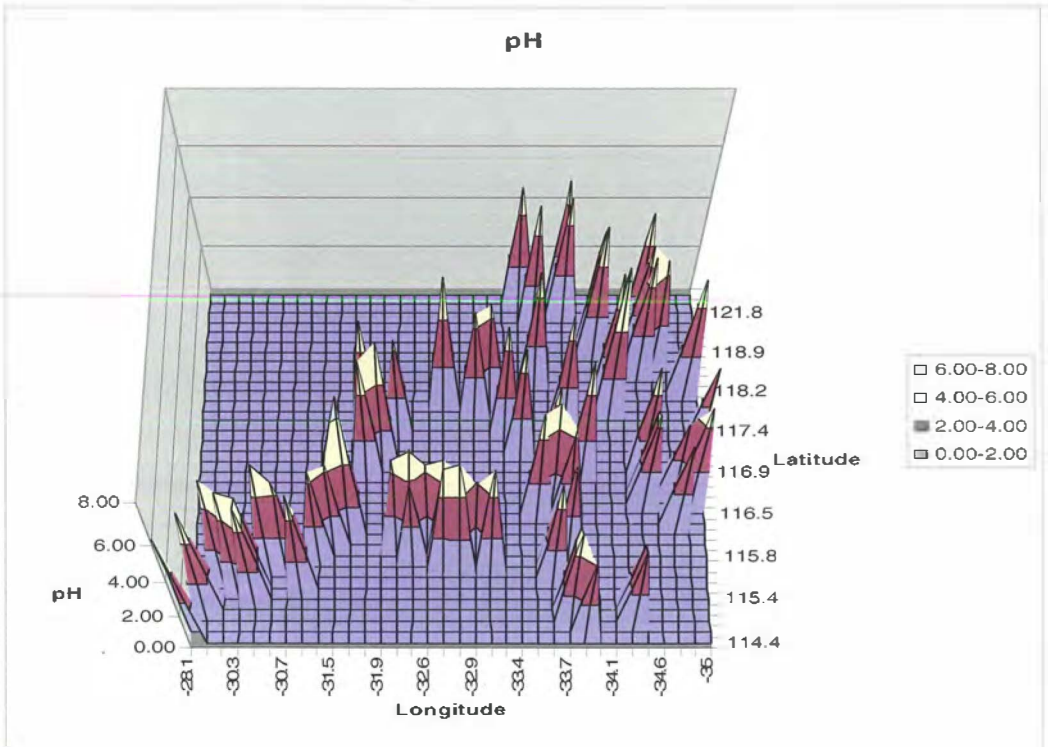
Pale deep sand (Normal data) – Elevation



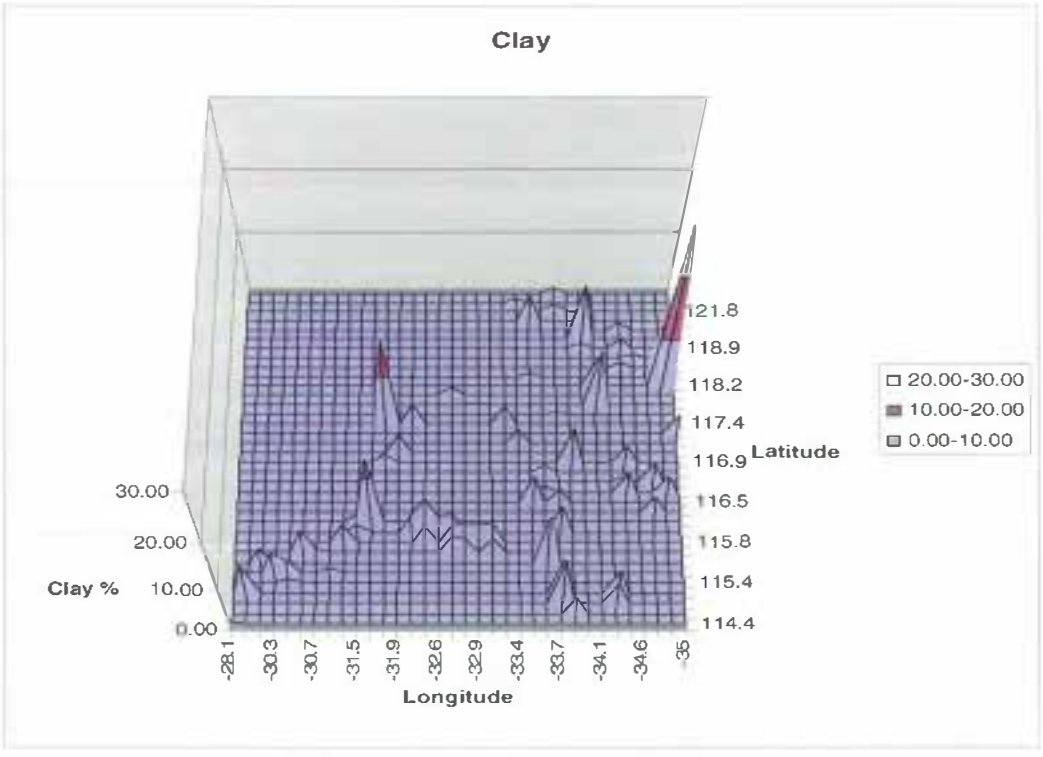
Pale deep sand (Normal data) – CAC03



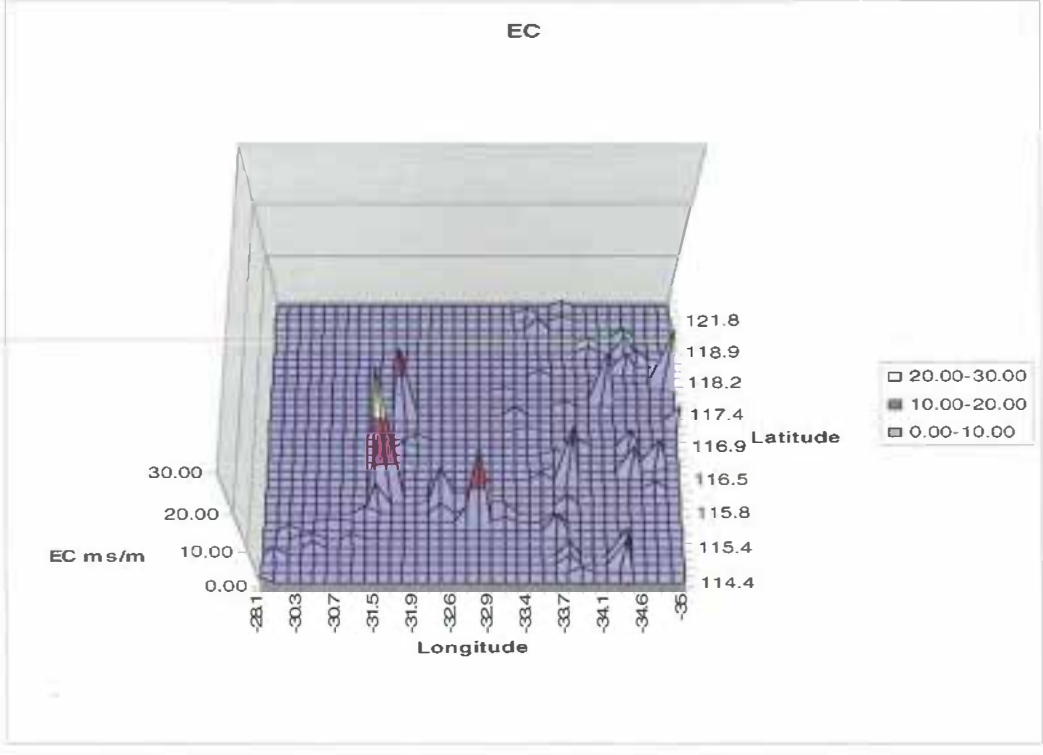
Pale deep sand (Normal data) – OC



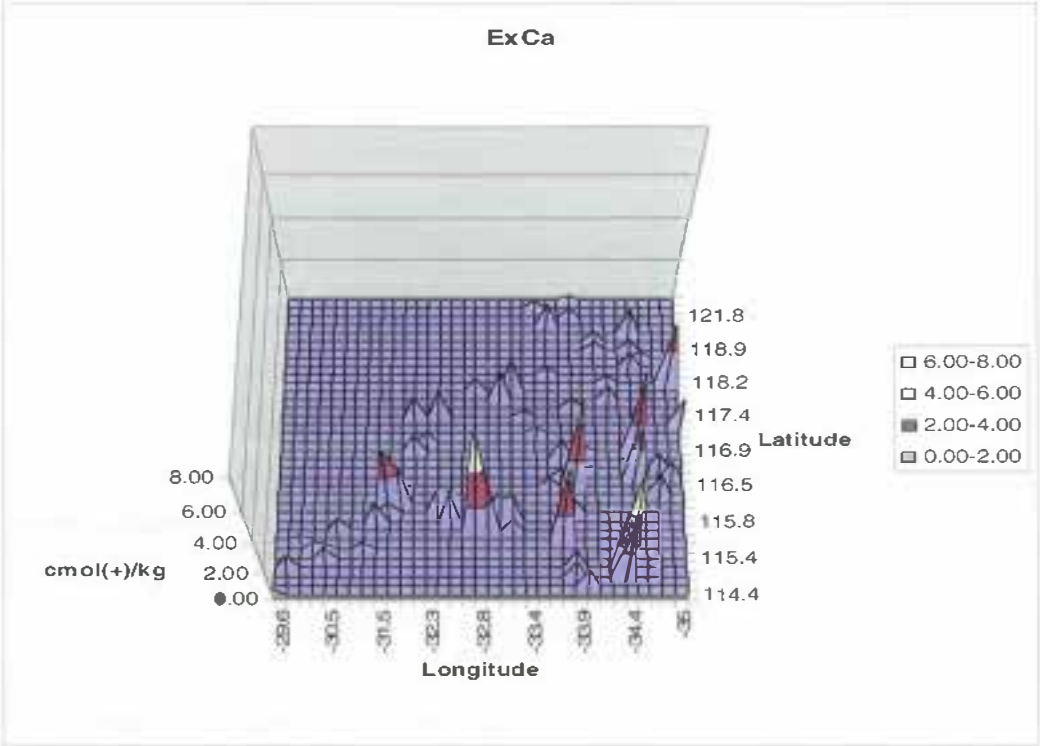
Pale deep sand (Normal data) – pH



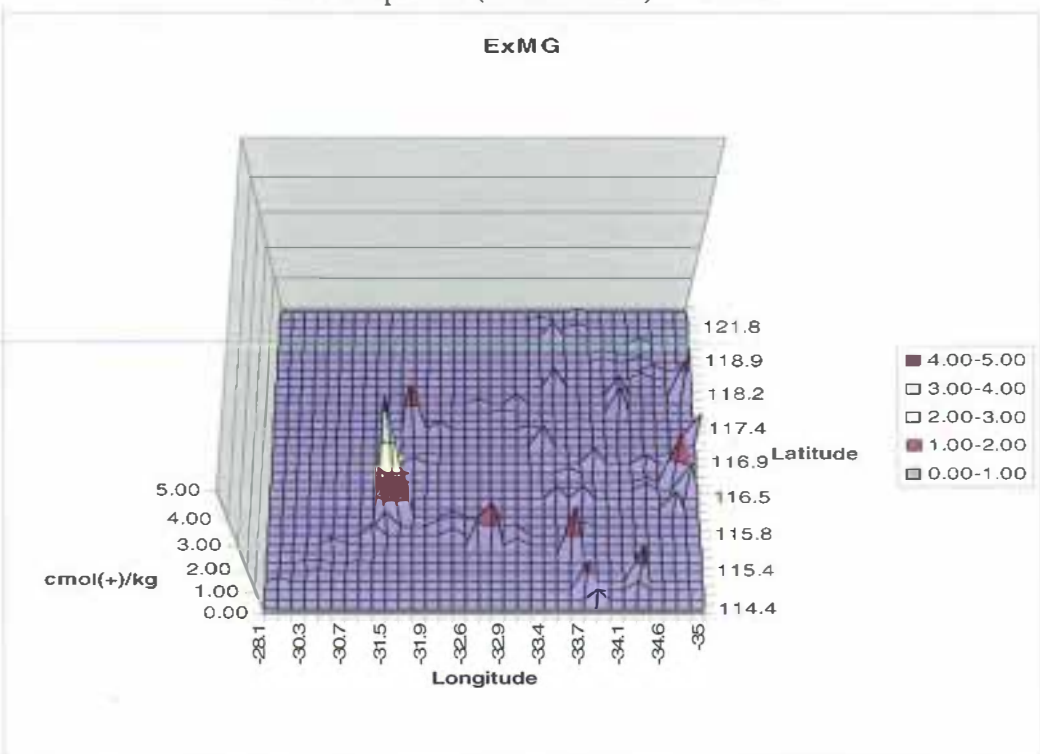
Pale deep sand (Normal data) – Clay



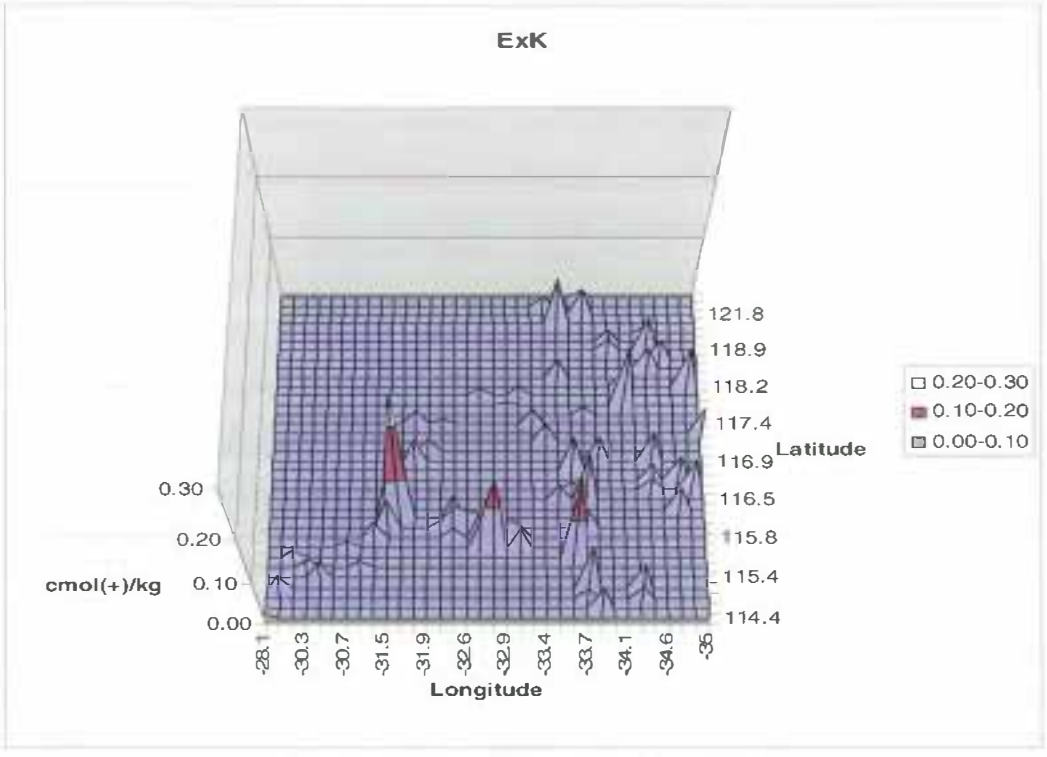
Pale deep sand (Normal data) – EC



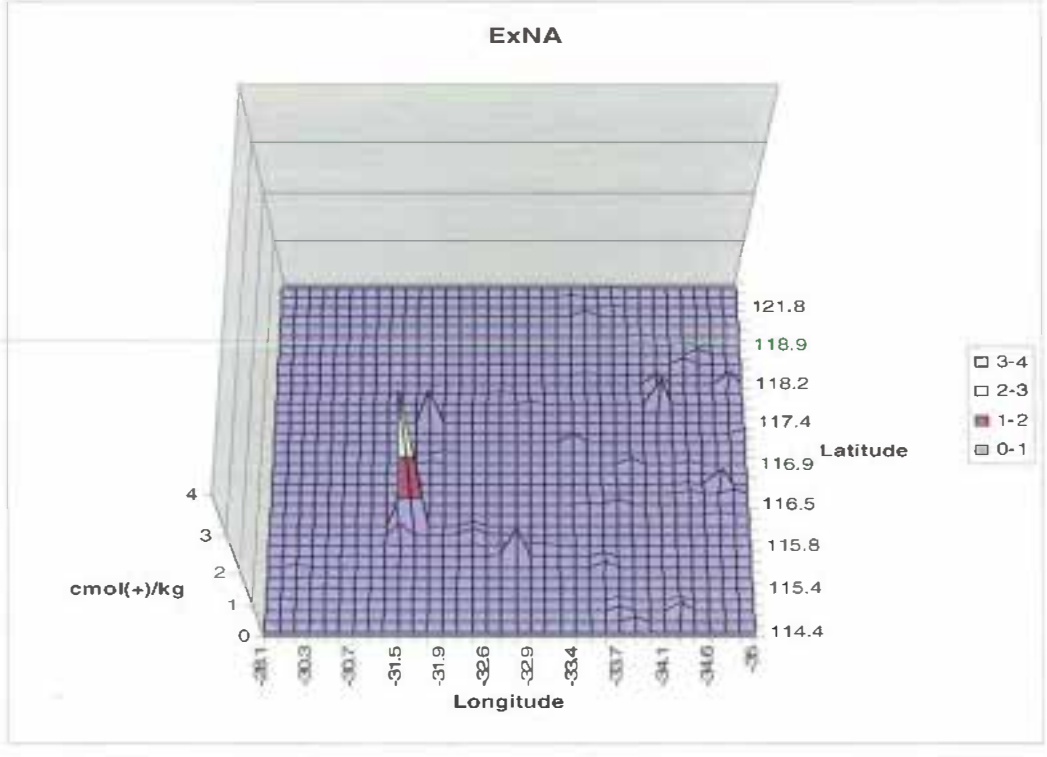
Pale deep sand (Normal data) – ExCA



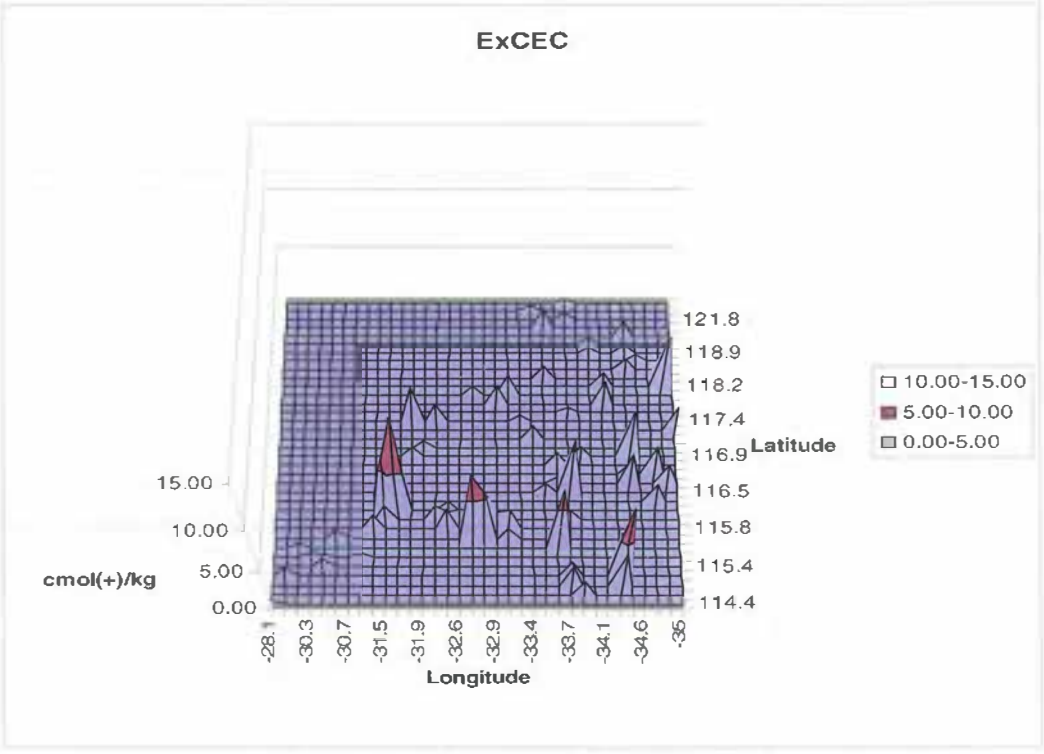
Pale deep sand (Normal data) – ExMG



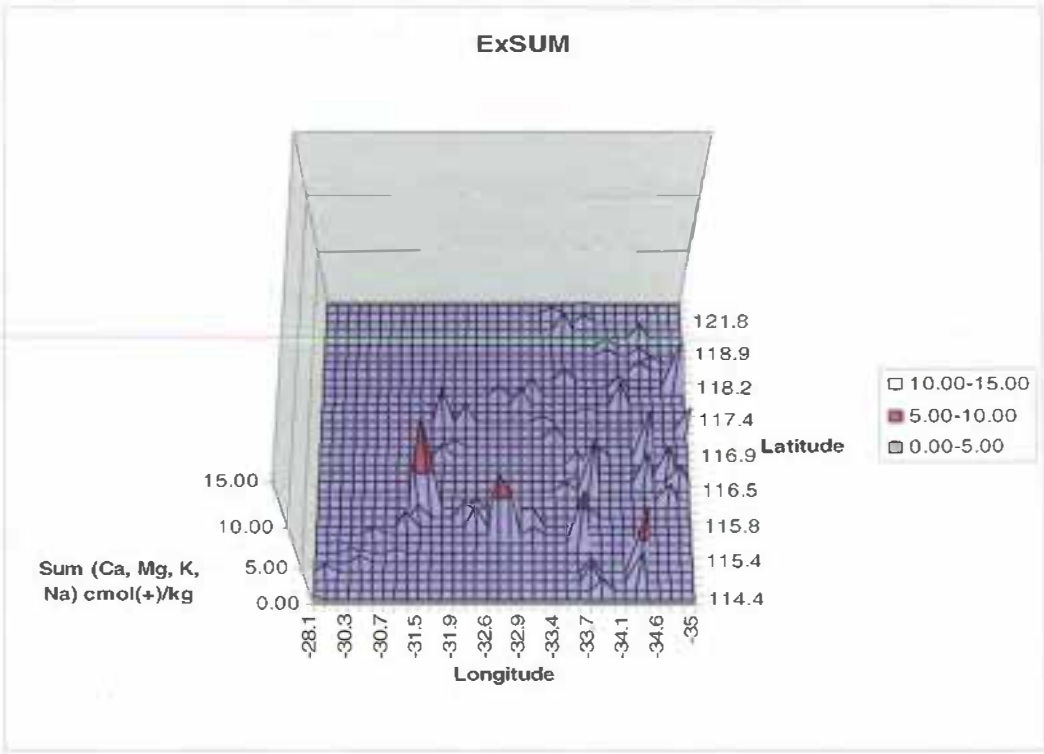
Pale deep sand (Normal data) – ExK



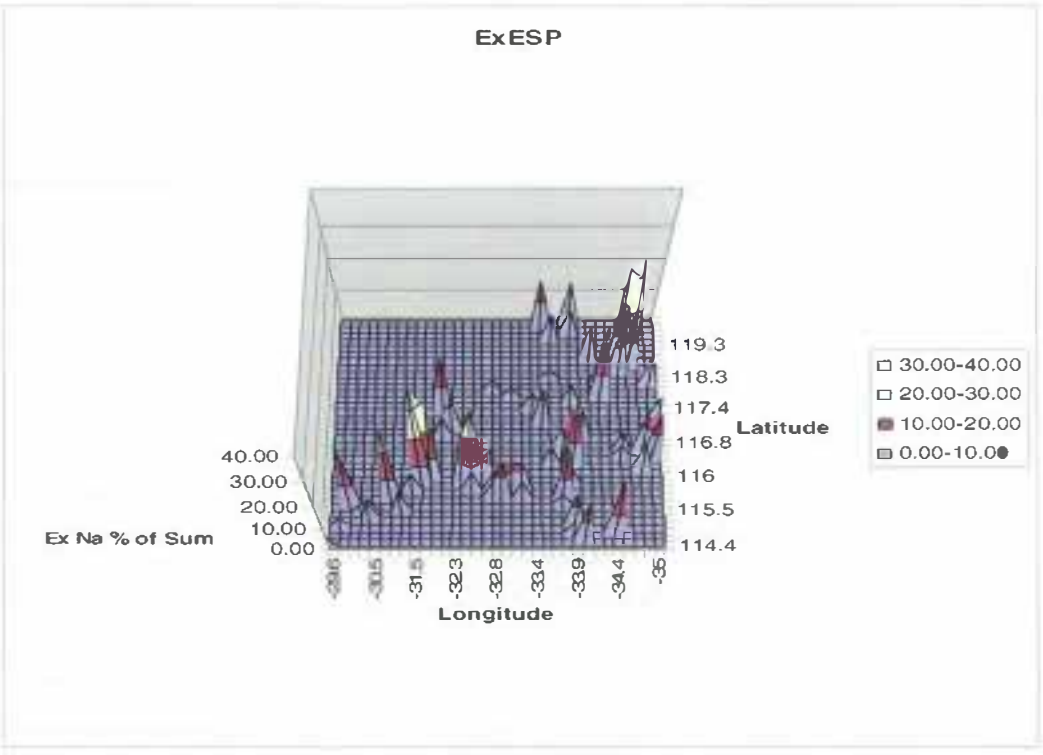
Pale deep sand (Normal data) – ExNA



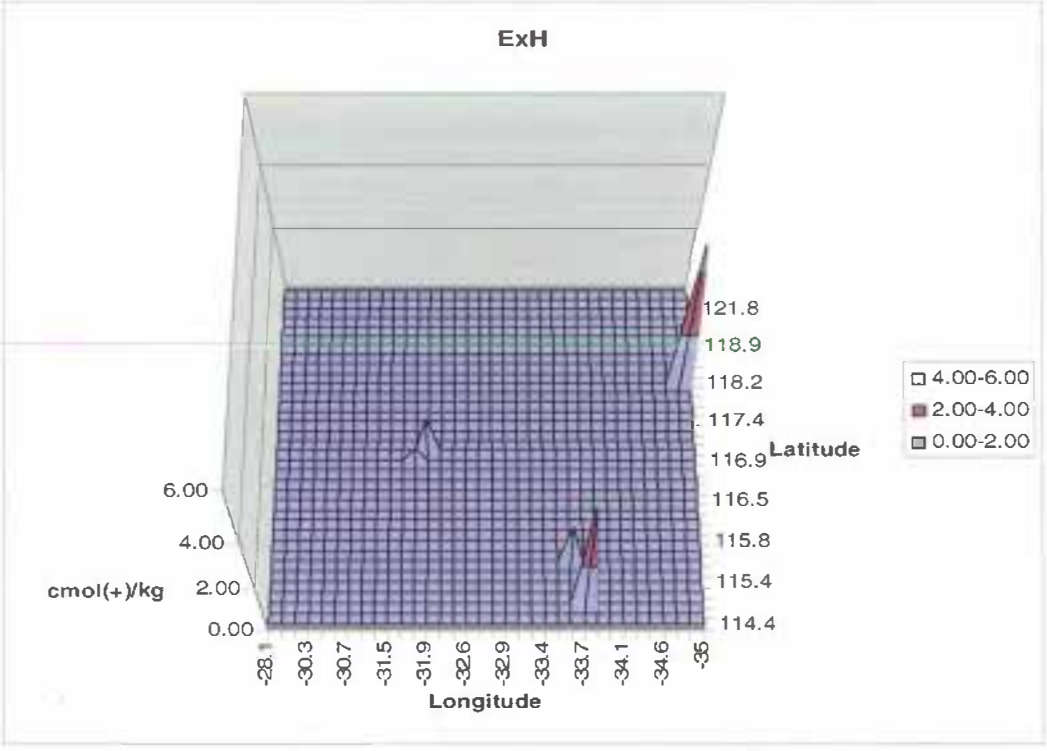
Pale deep sand (Normal data) – ExCEC



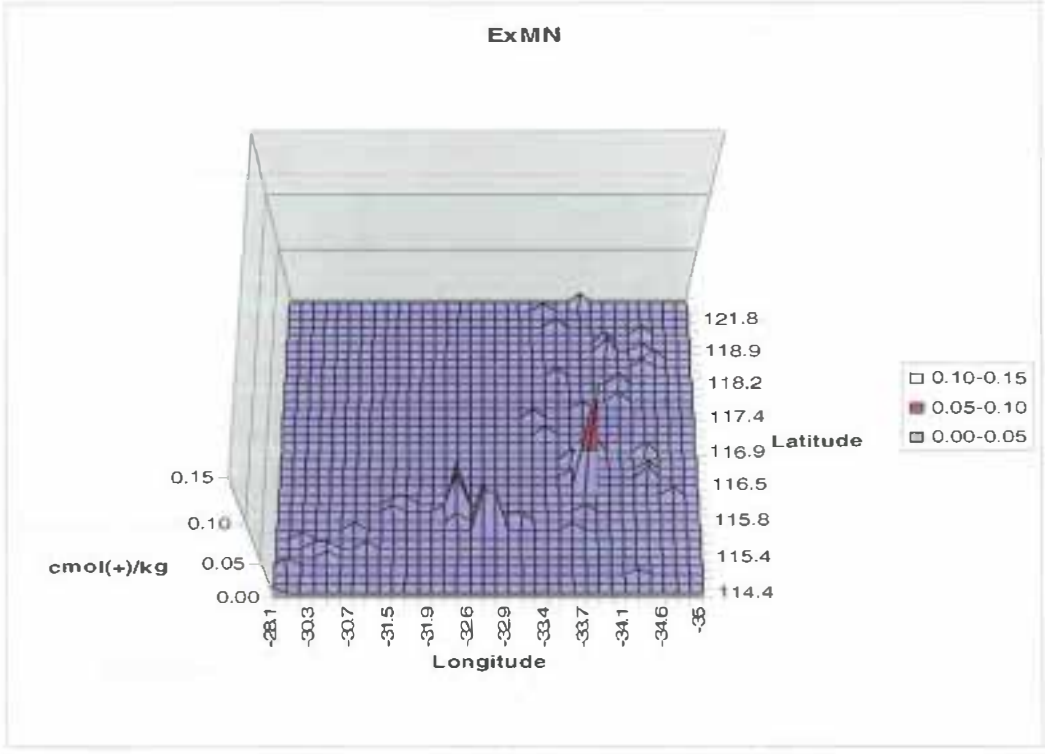
Pale deep sand (Normal data) – ExSUM



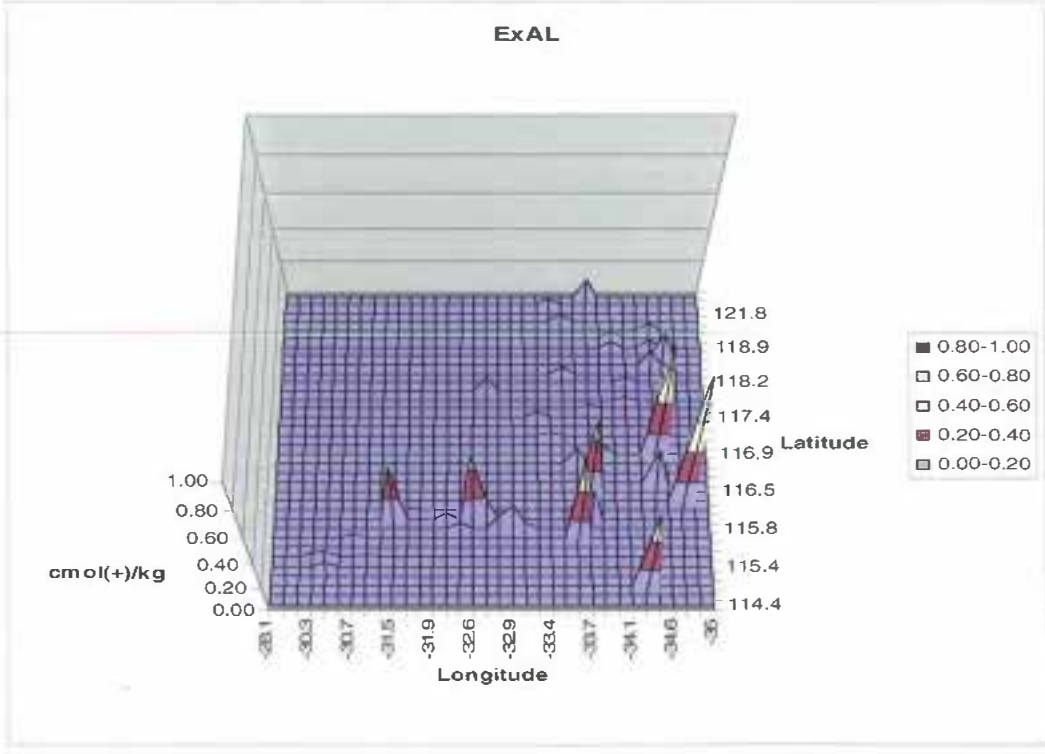
Pale deep sand (Normal data) – ExESP



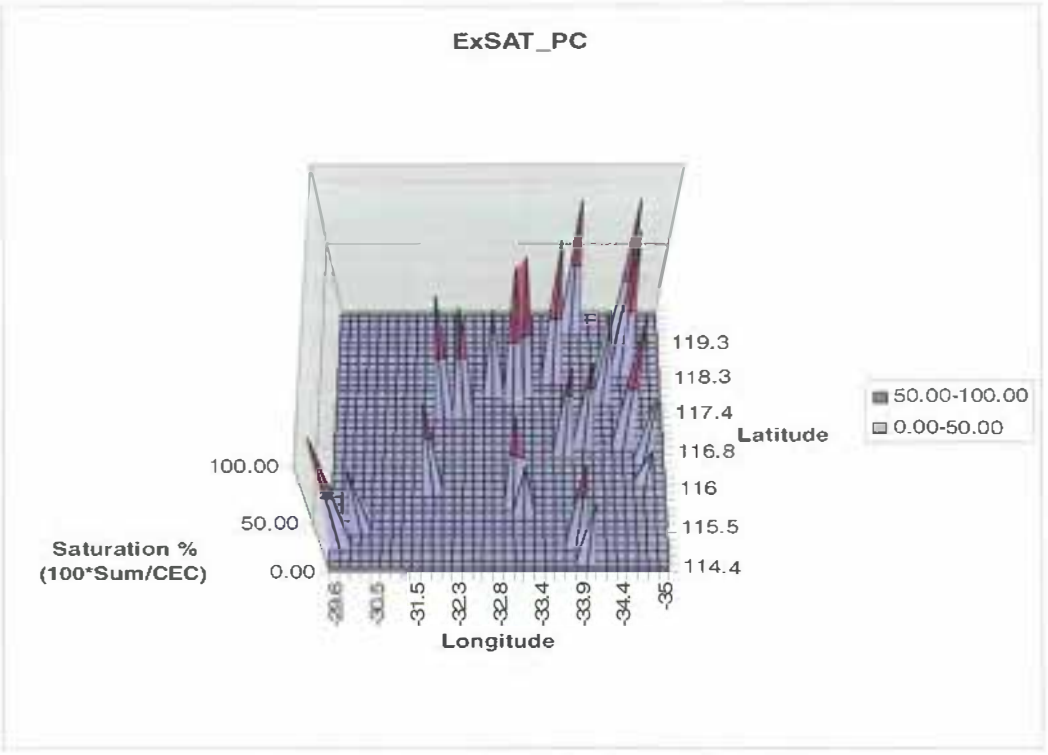
Pale deep sand (Normal data) – ExH



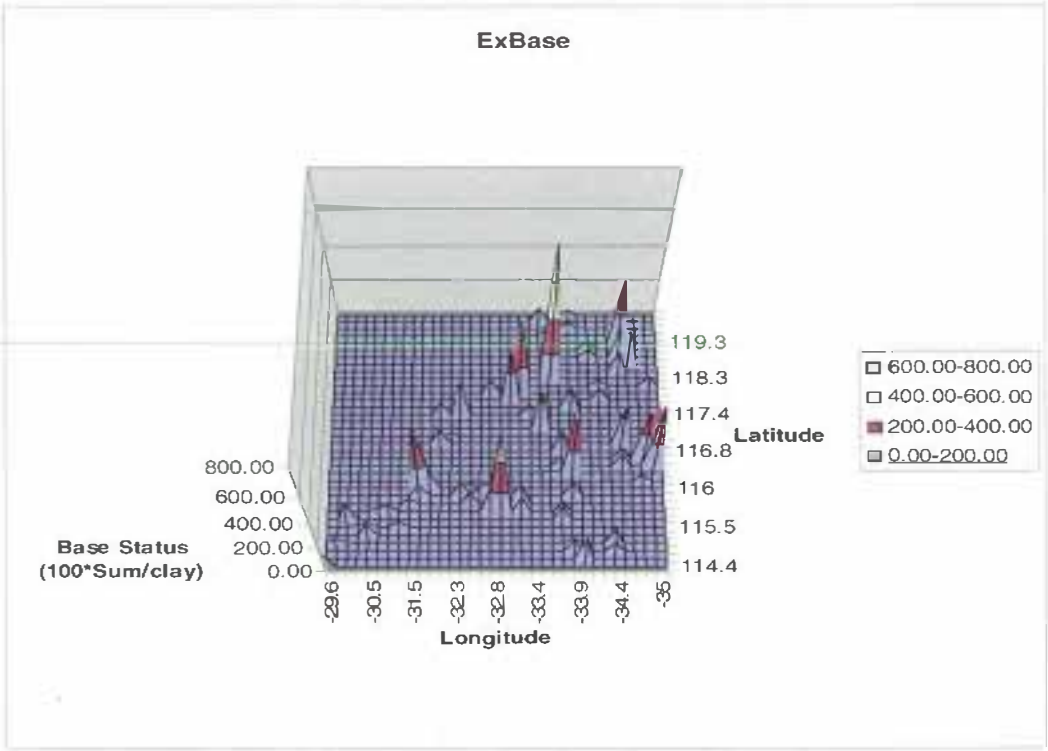
Pale deep sand (Normal data) – ExMN



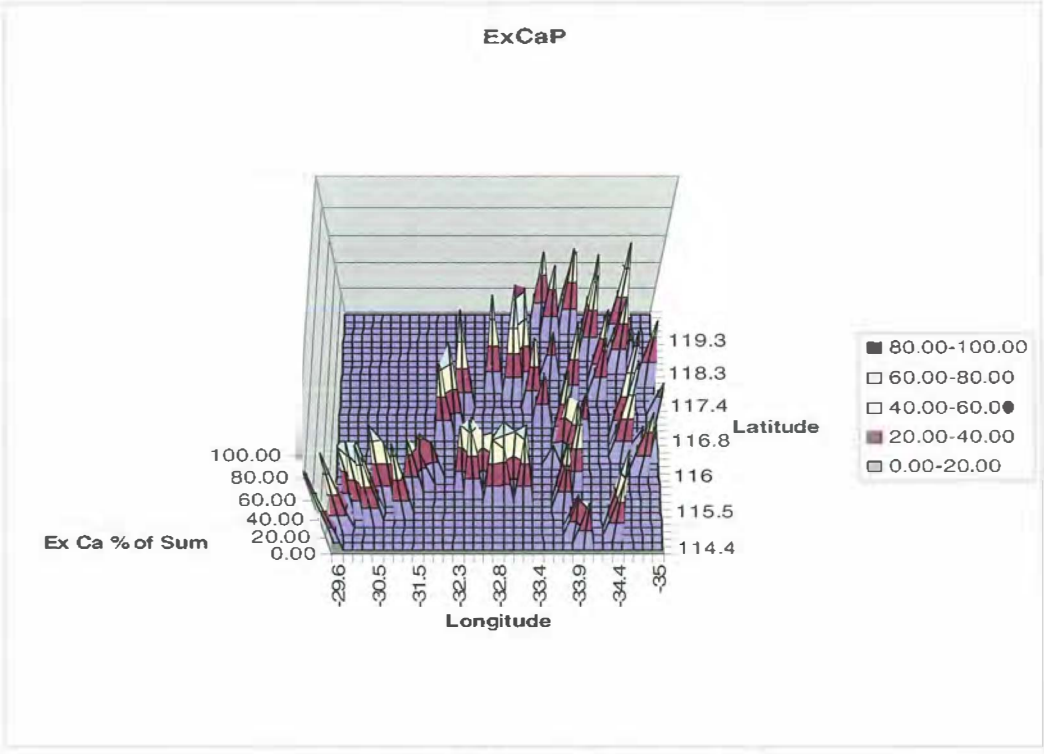
Pale deep sand (Normal data) – ExAL



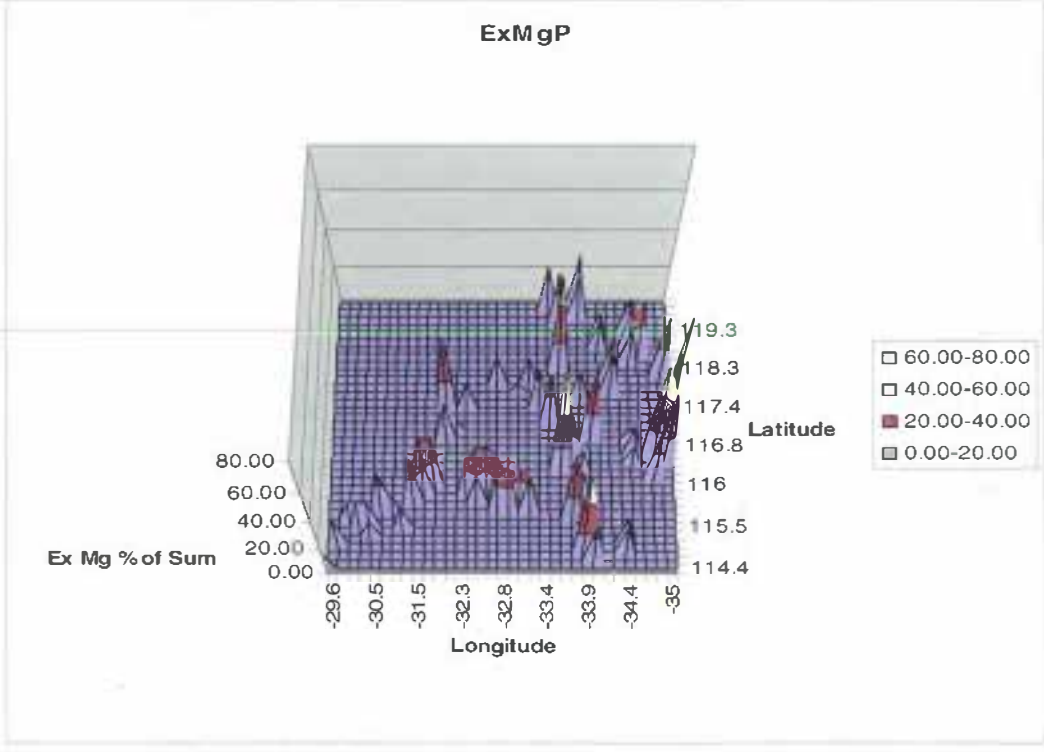
Pale deep sand (Normal data) – ExSAT_PC



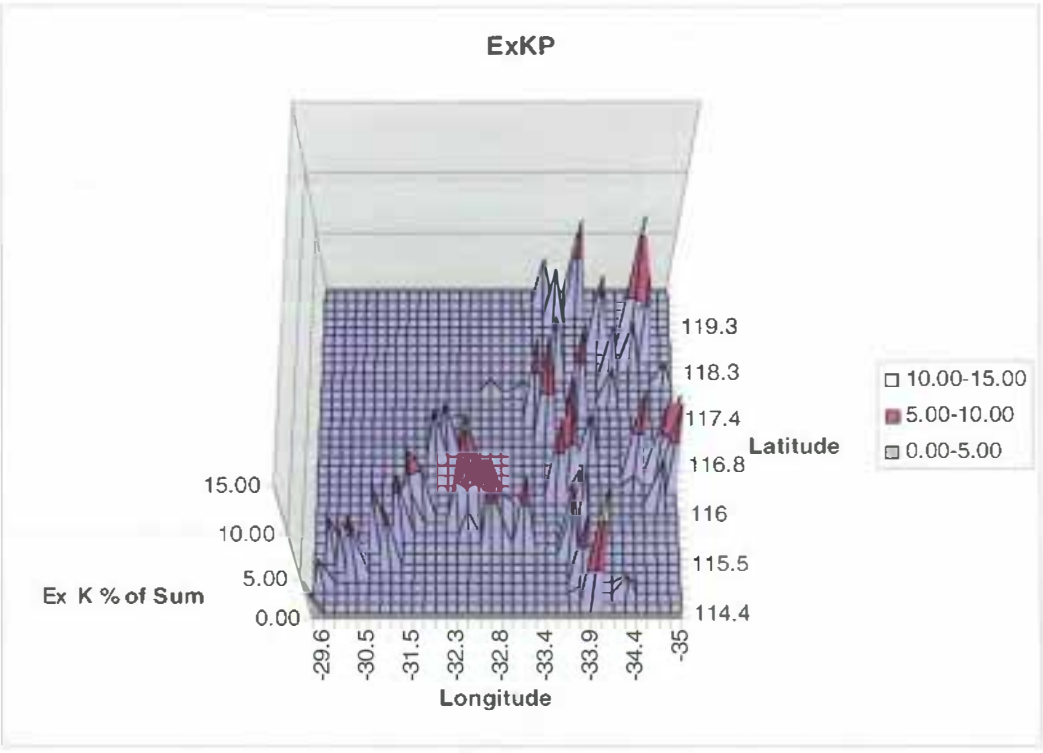
Pale deep sand (Normal data) – ExBASE



Pale deep sand (Normal data) – ExCaP



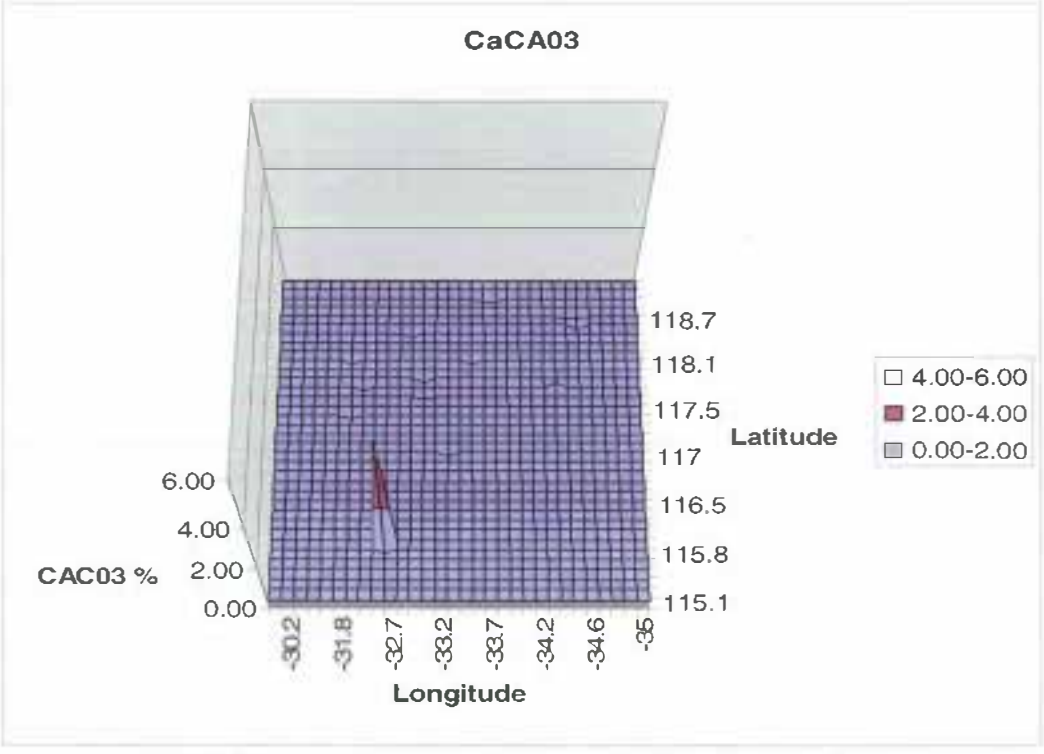
Pale deep sand (Normal data) – ExMgP



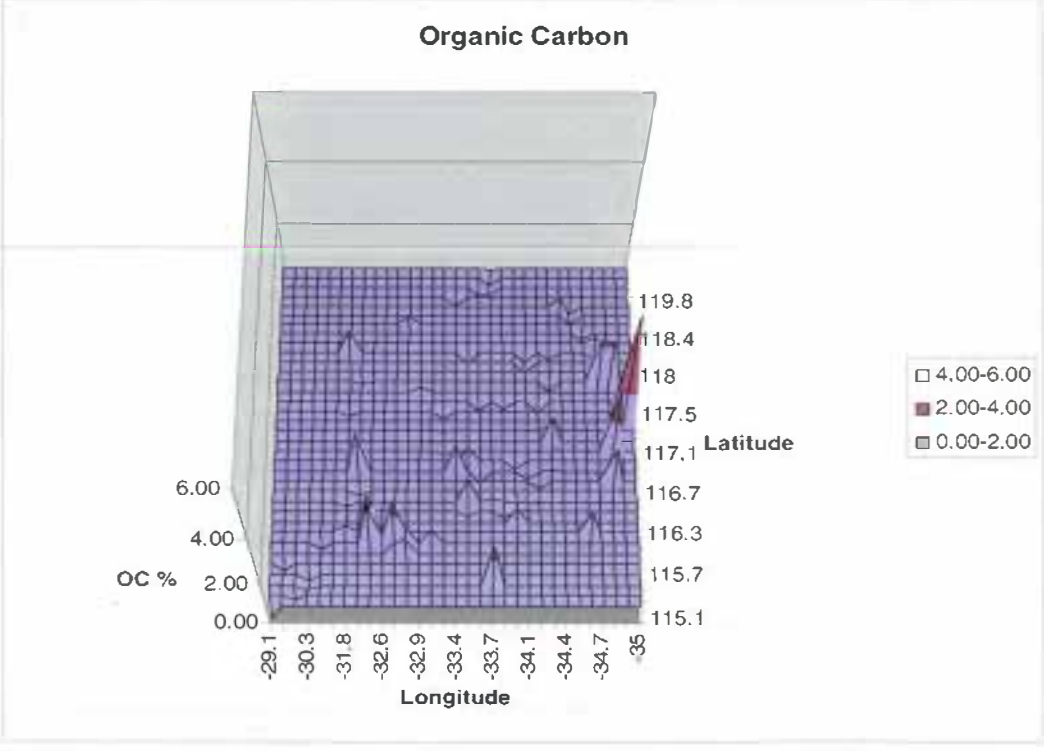
Pale deep sand (Normal data) – ExKP

8.3.5 Stage 5: Standardized data – 3 Main soil types

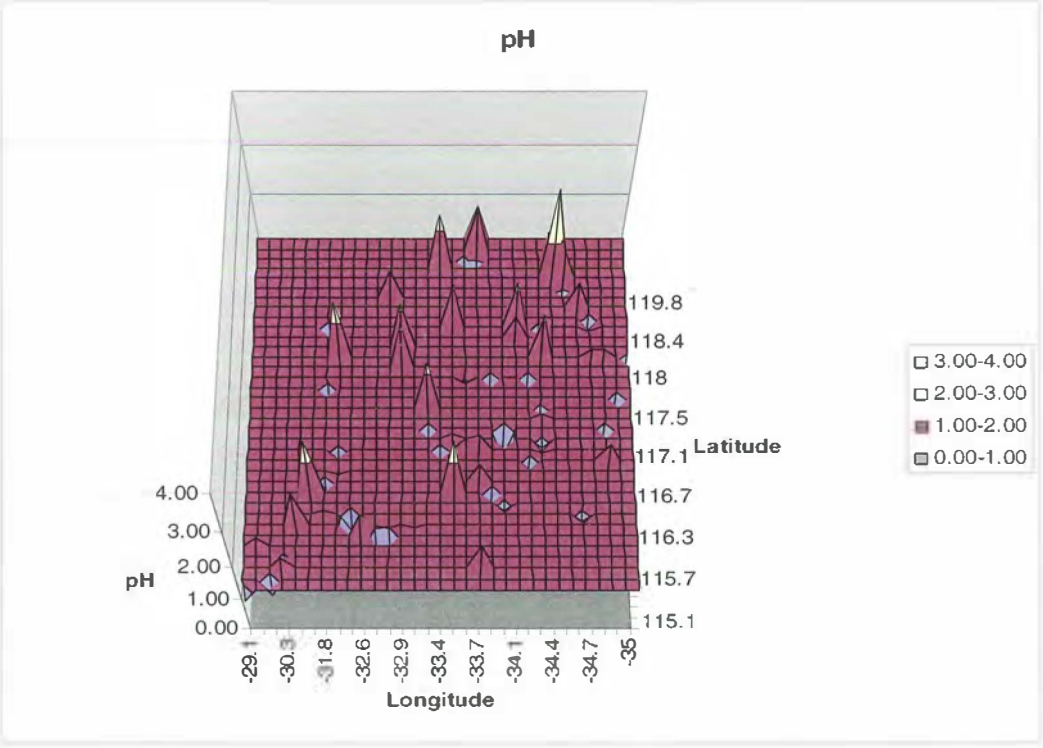
Soil 1: Grey deep sandy duplex



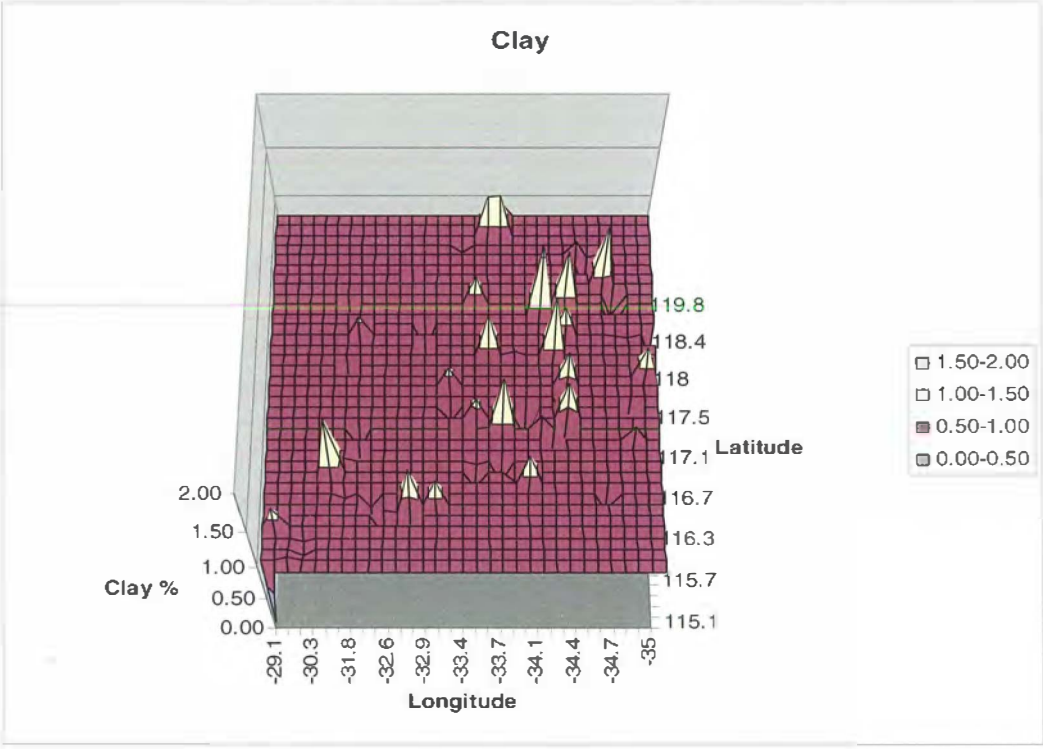
Grey deep sandy duplex (Standardized data) – CAC03 %



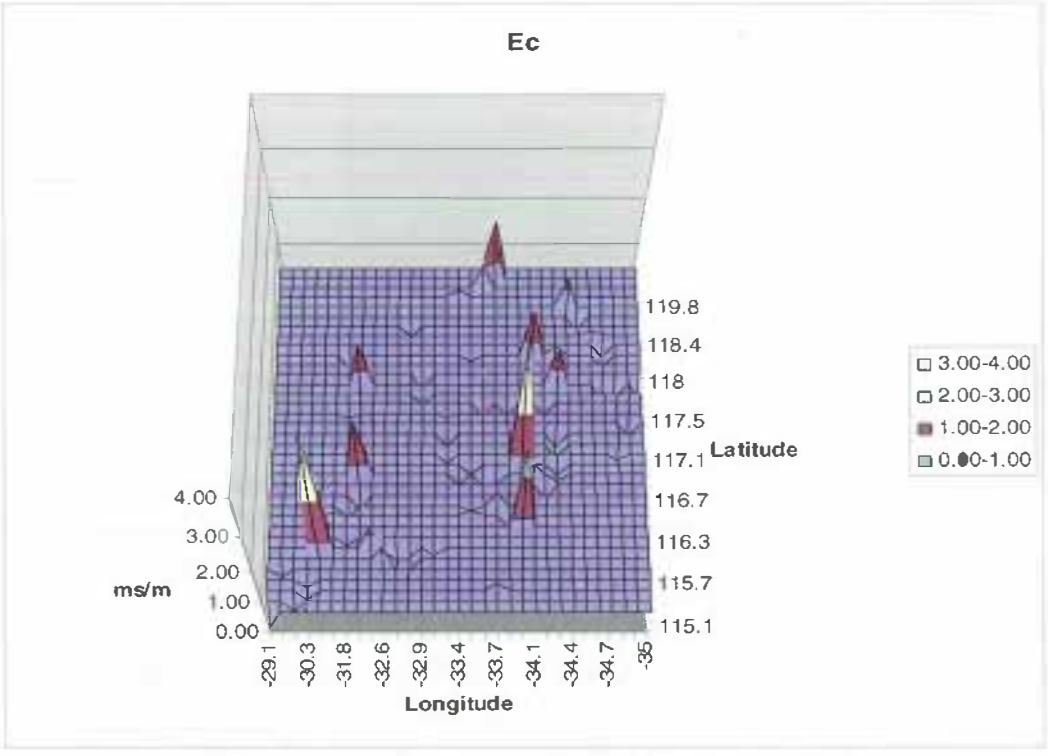
Grey deep sandy duplex (Standardized data) – OC



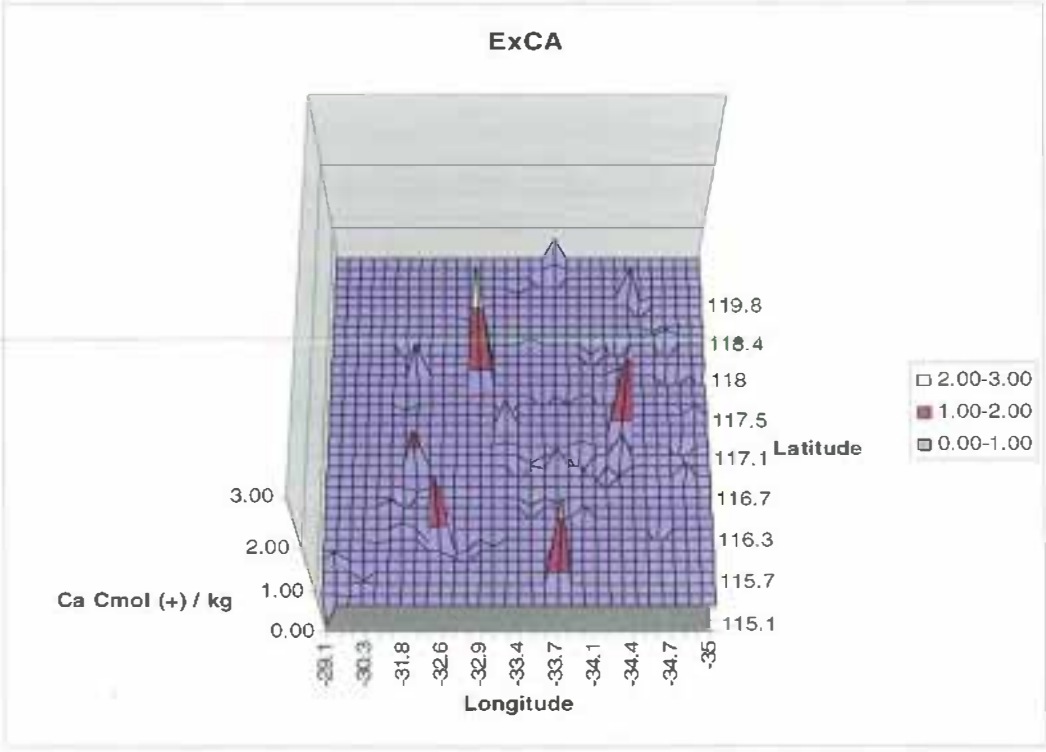
Grey deep sandy duplex (Standardized data) – pH



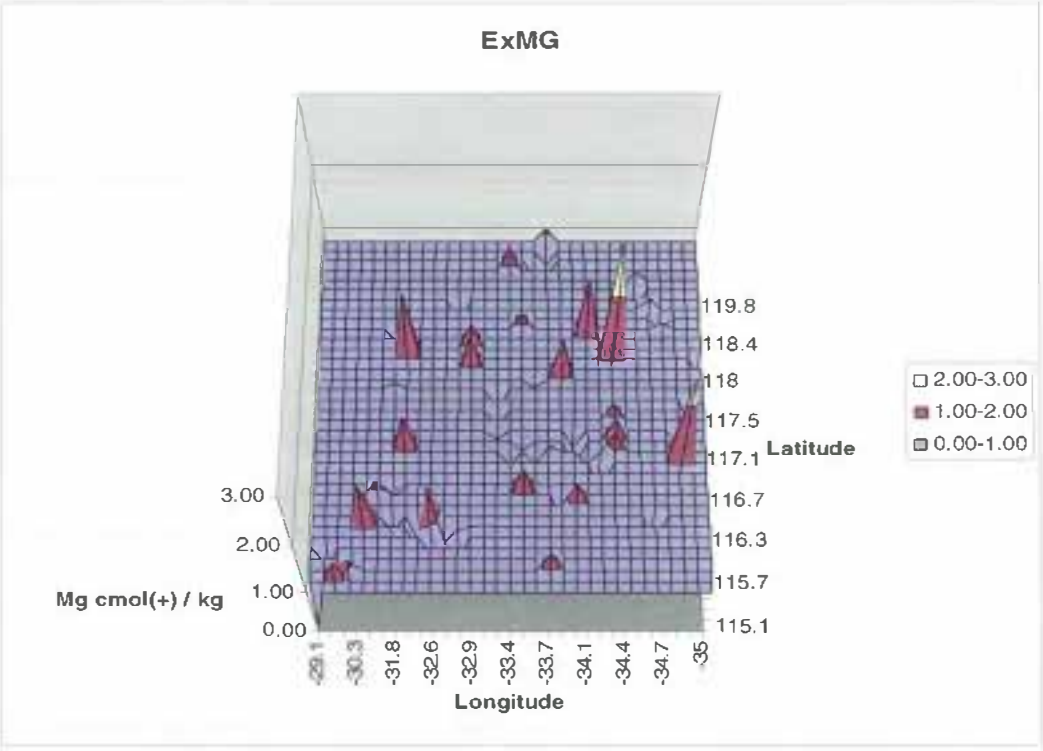
Grey deep sandy duplex (Standardized data) – Clay



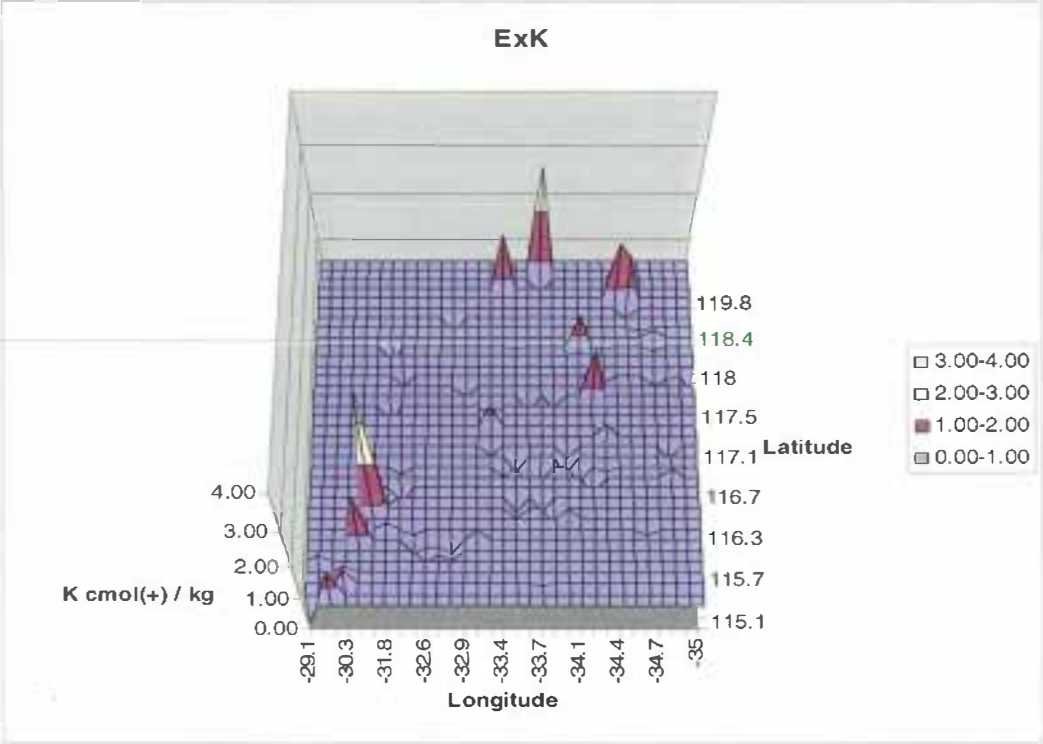
Grey deep sandy duplex (Standardized data) – EC



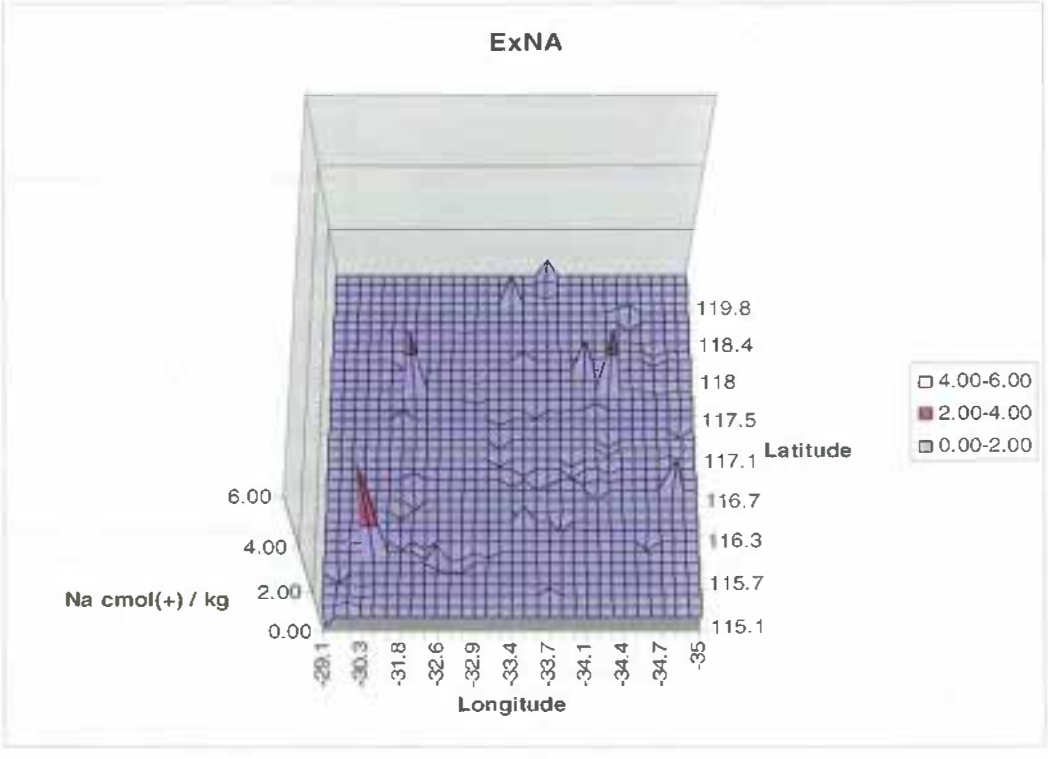
Grey deep sandy duplex (Standardized data) – ExCA



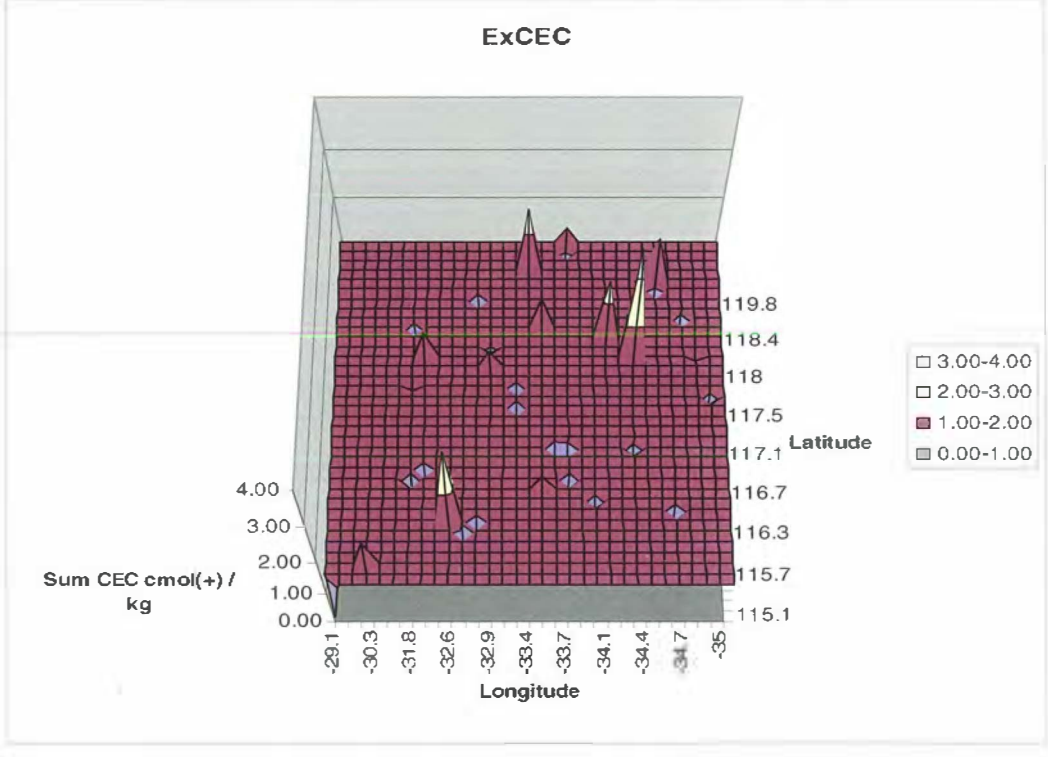
Grey deep sandy duplex (Standardized data) – ExMG



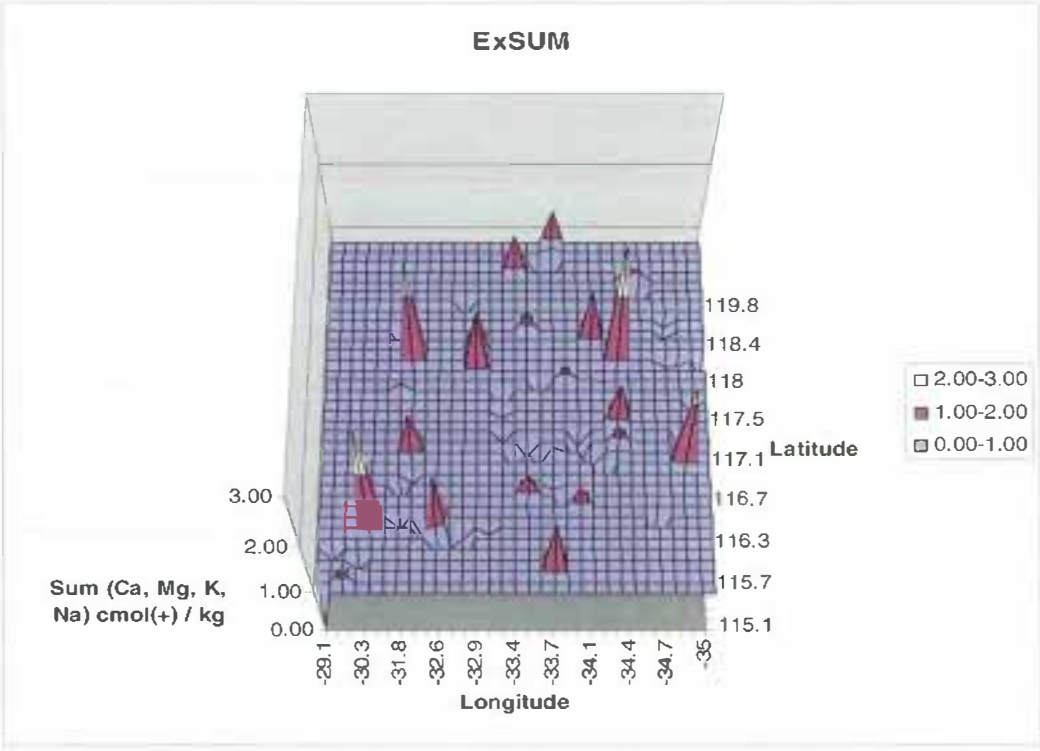
Grey deep sandy duplex (Standardized data) – ExK



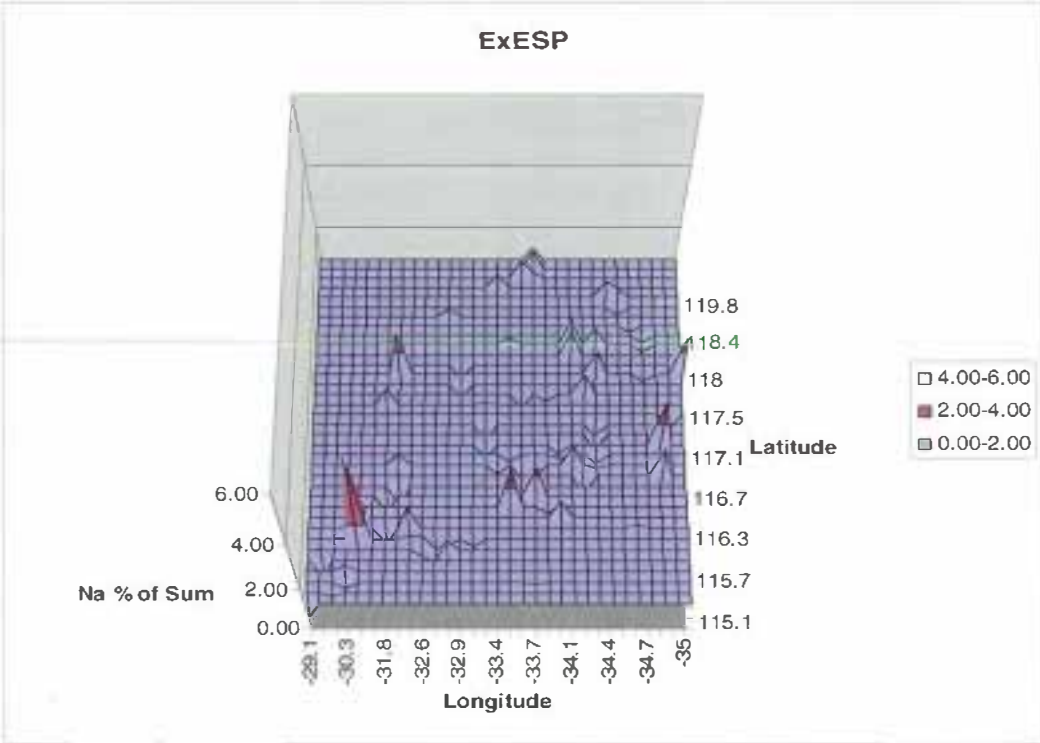
Grey deep sandy duplex (Standardized data) – ExNA



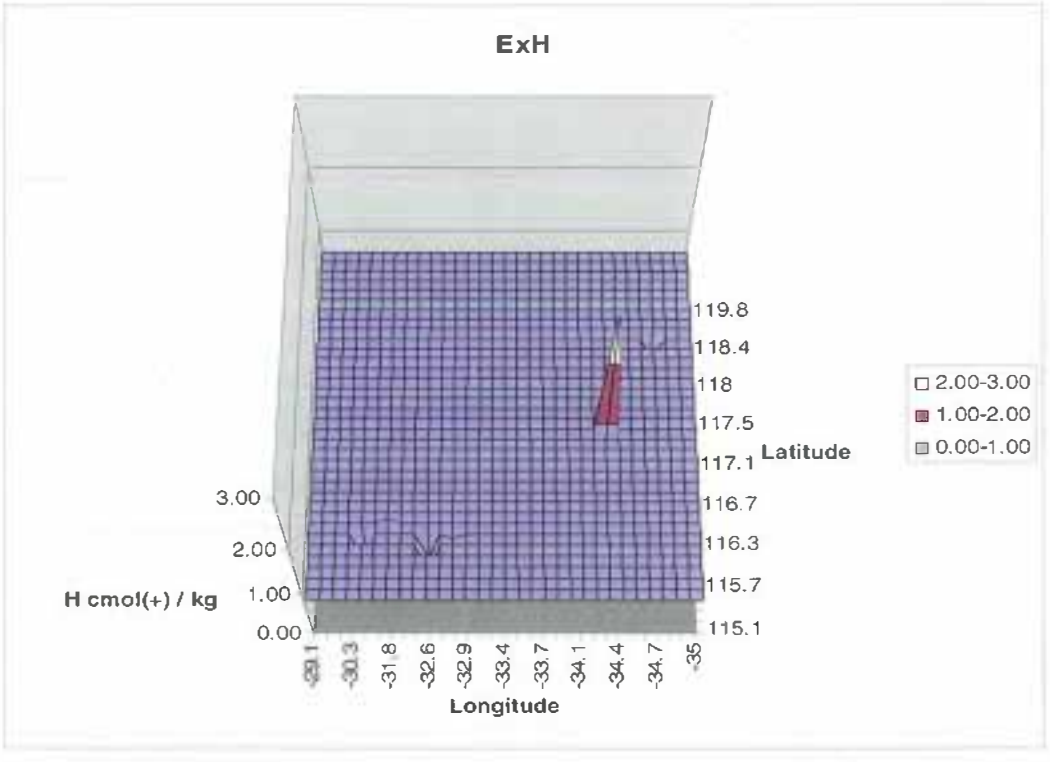
Grey deep sandy duplex (Standardized data) – ExCEC



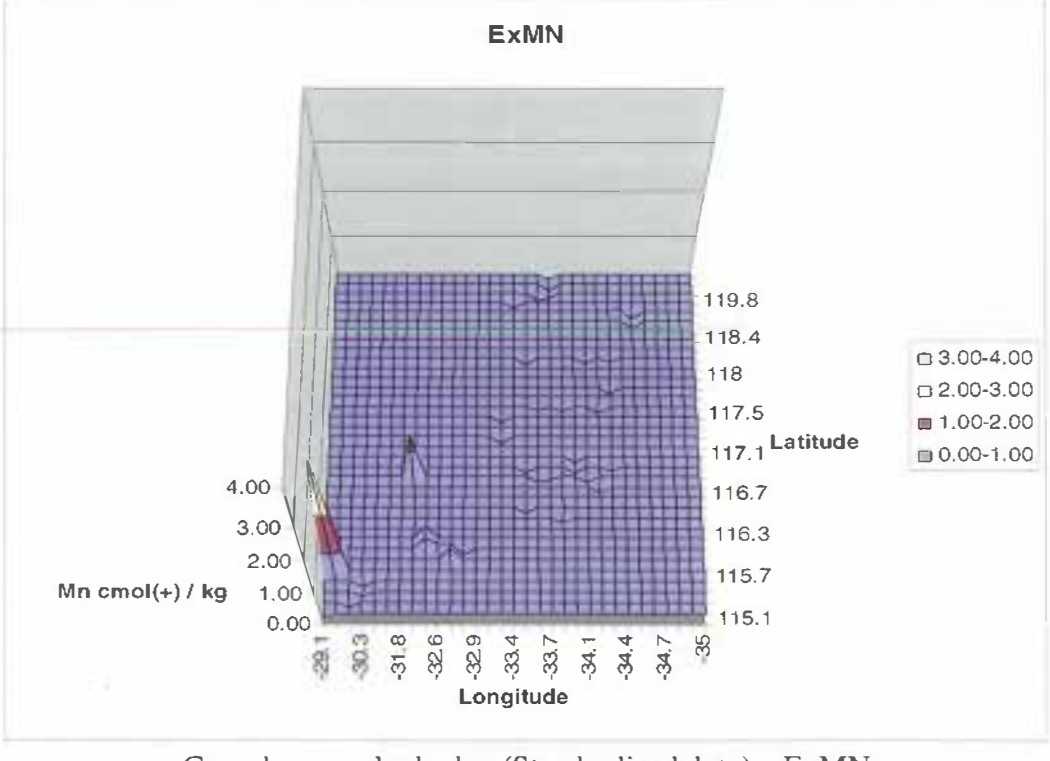
Grey deep sandy duplex (Standardized data) – ExSUM



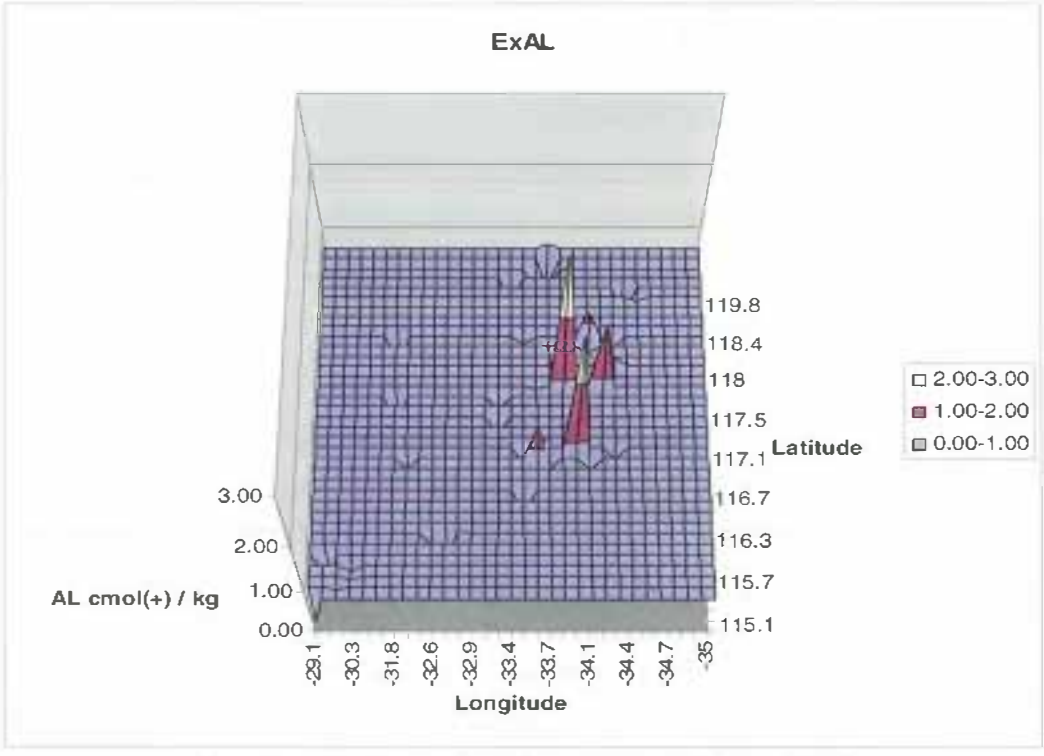
Grey deep sandy duplex (Standardized data) – ExSUM



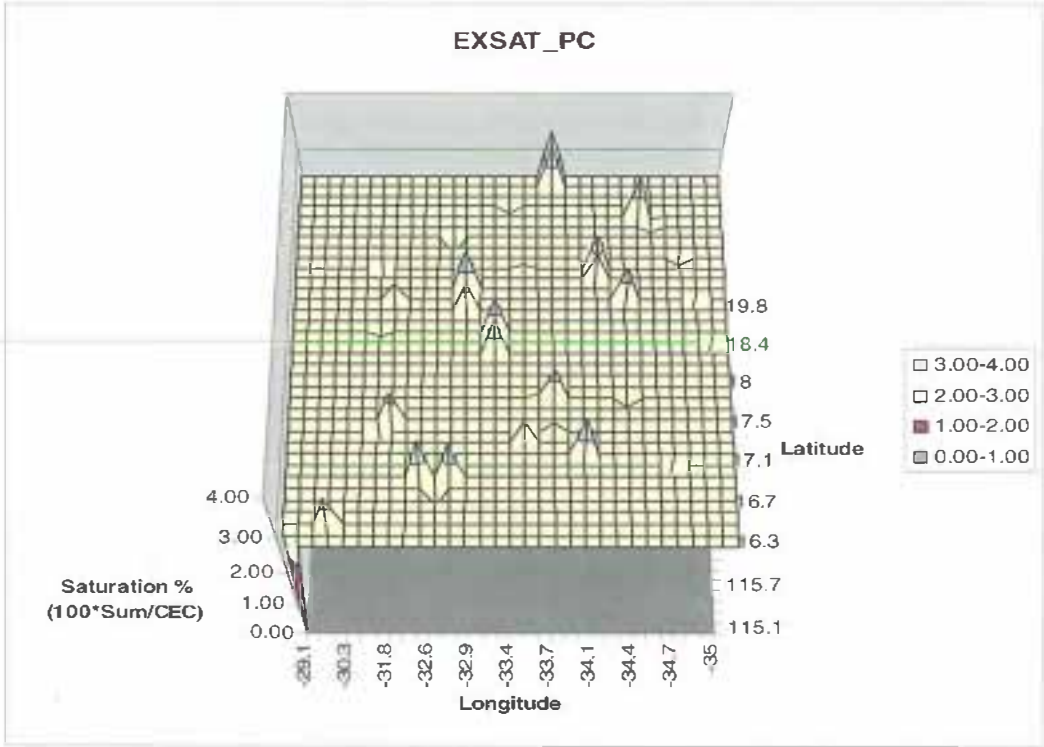
Grey deep sandy duplex (Standardized data) – ExH



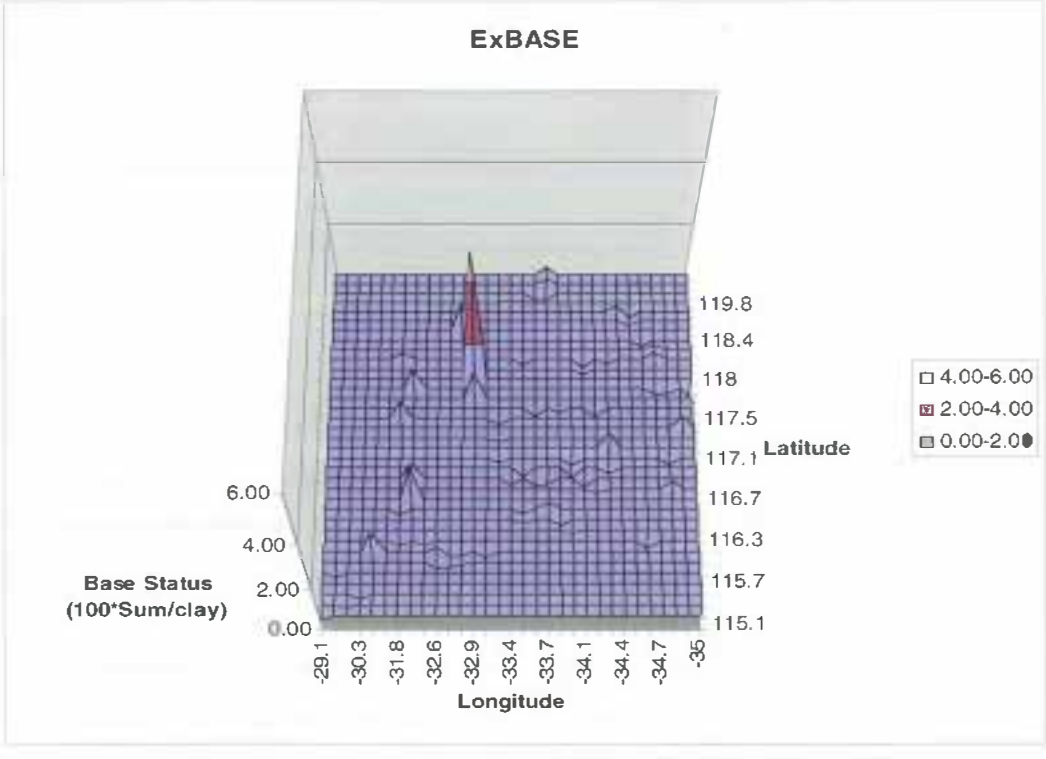
Grey deep sandy duplex (Standardized data) – ExMN



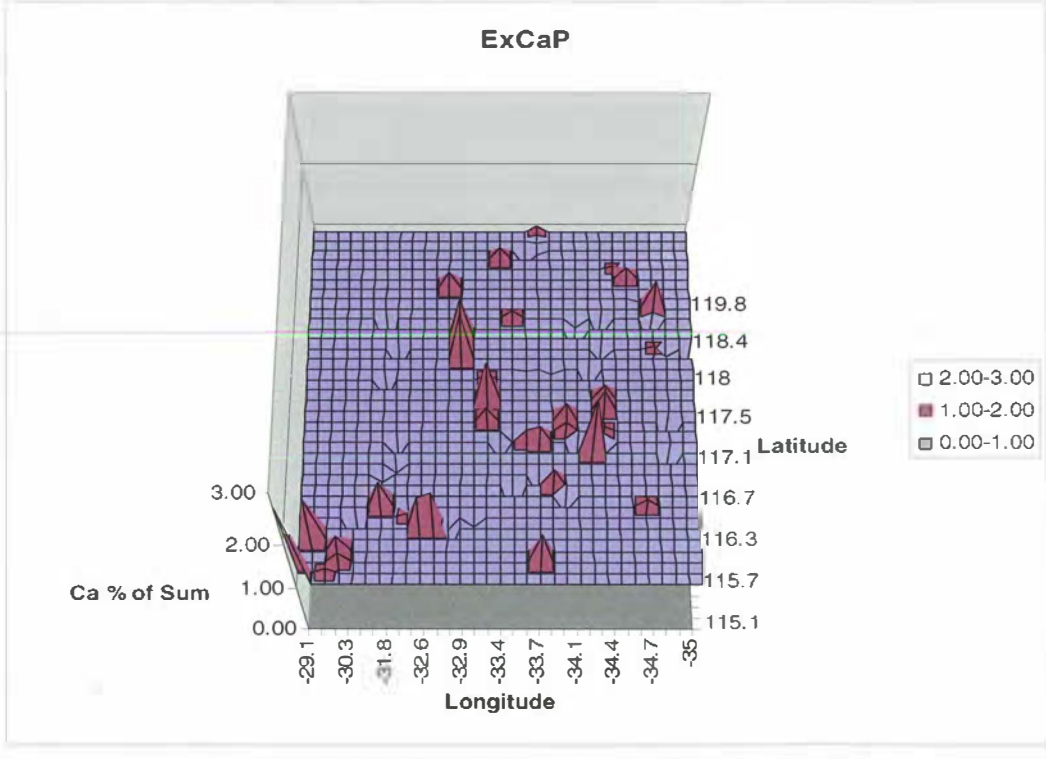
Grey deep sandy duplex (Standardized data) – ExAL



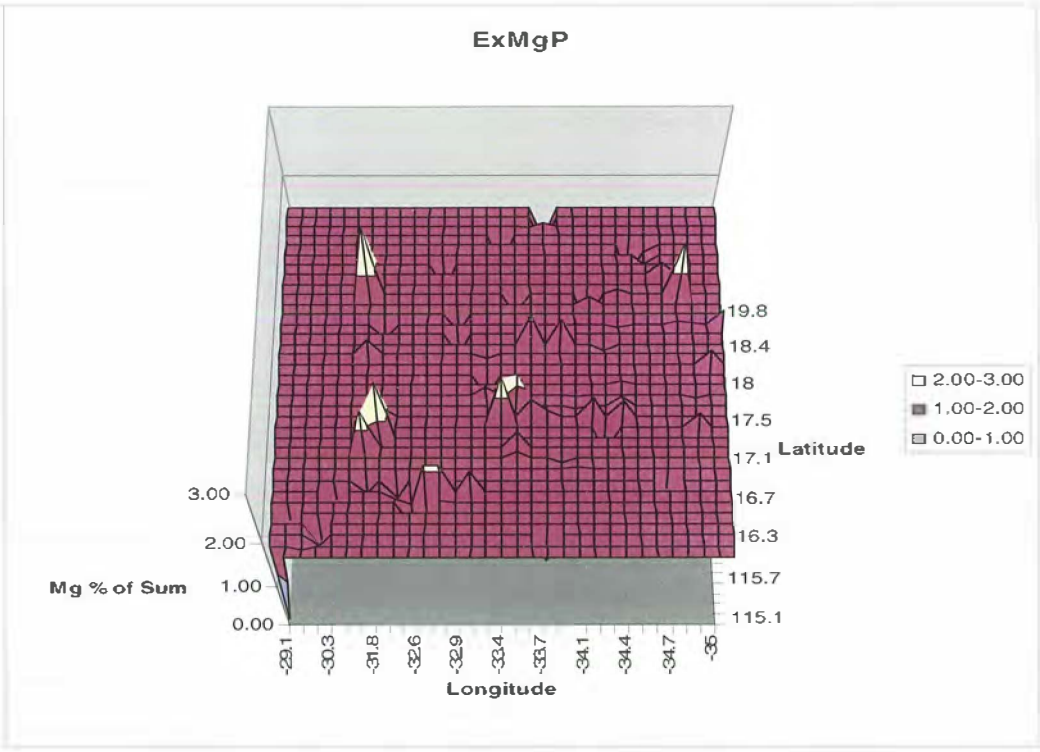
Grey deep sandy duplex (Standardized data) – ExSAT_PC



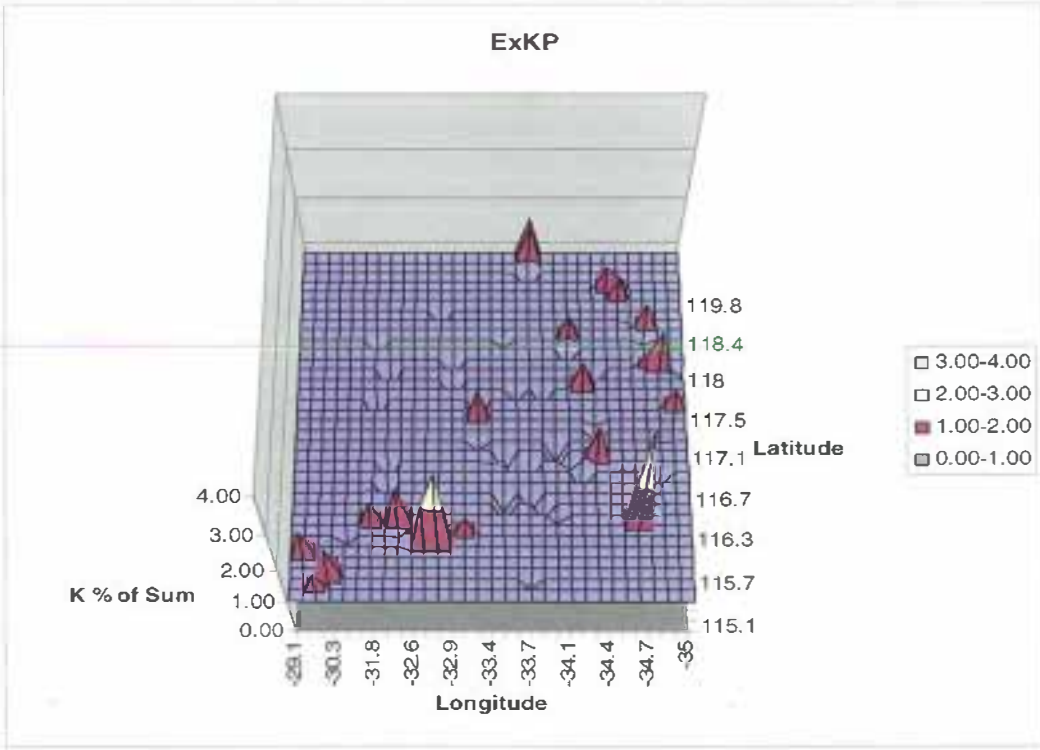
Grey deep sandy duplex (Standardized data) – ExBASE



Grey deep sandy duplex (Standardized data) – ExCaP

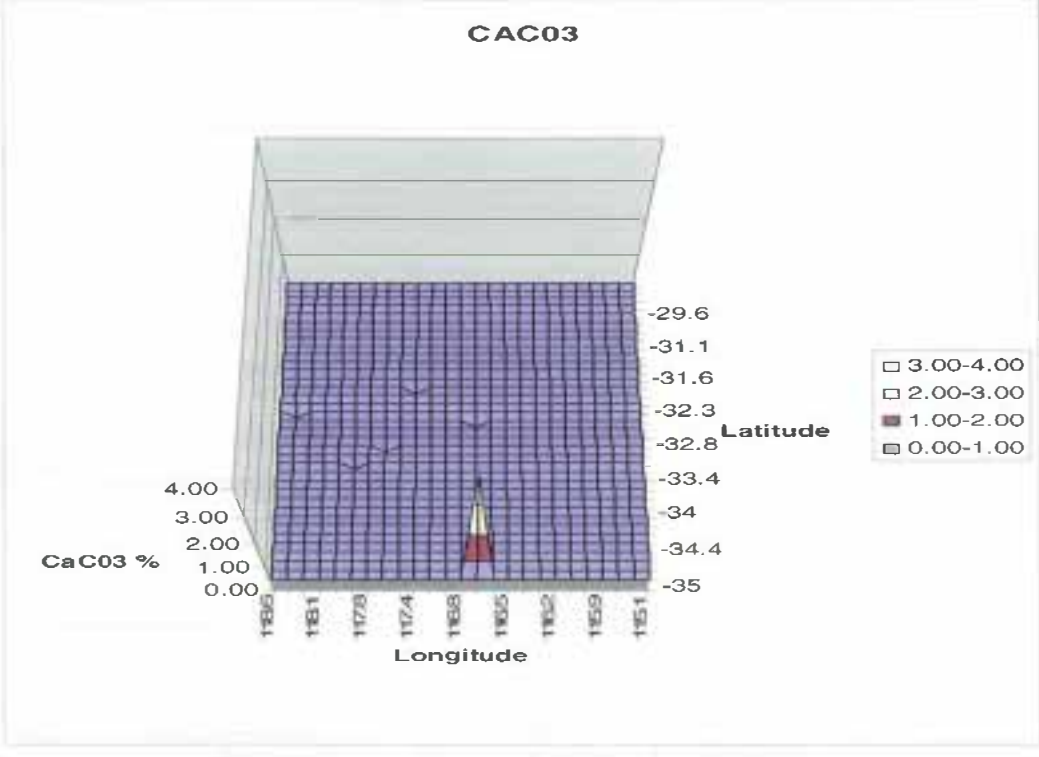


Grey deep sandy duplex (Standardized data) – ExMgP

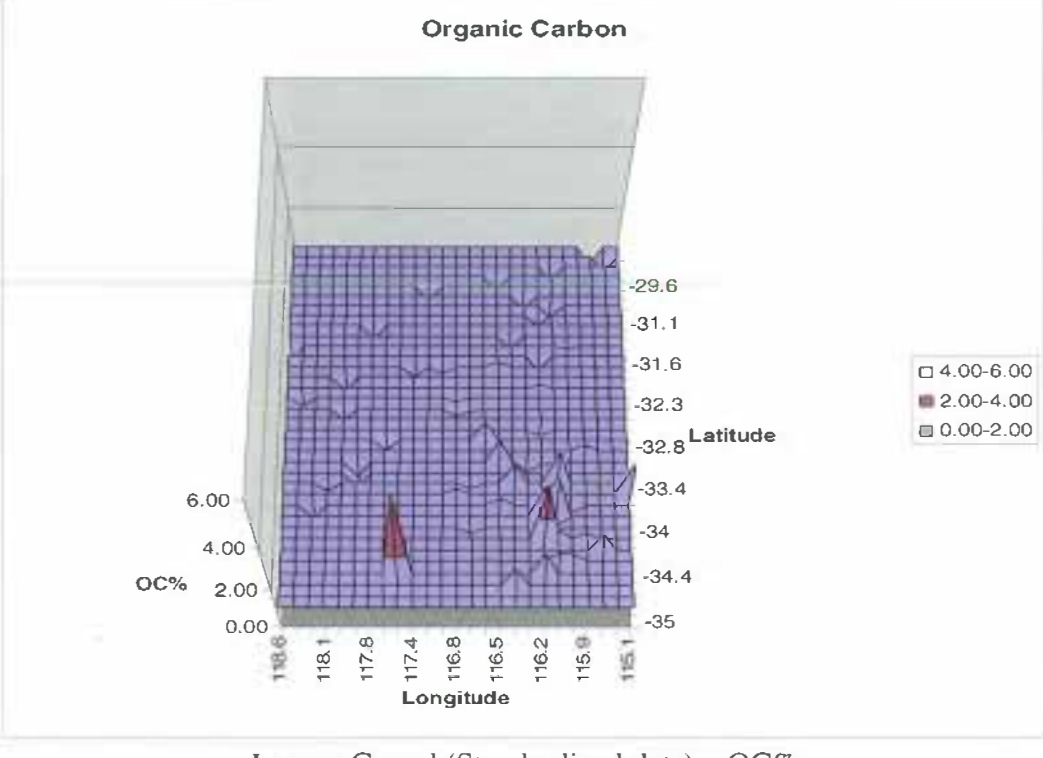


Grey deep sandy duplex (Standardized data) – ExKP

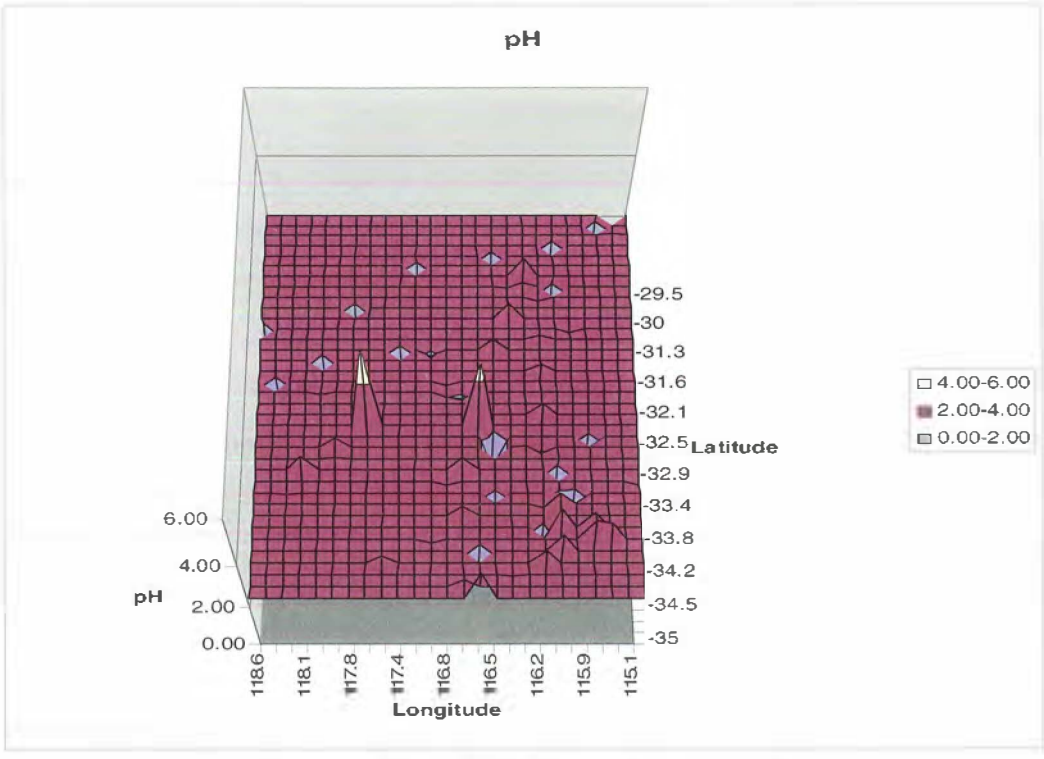
Soil 2 - Loamy Gravel



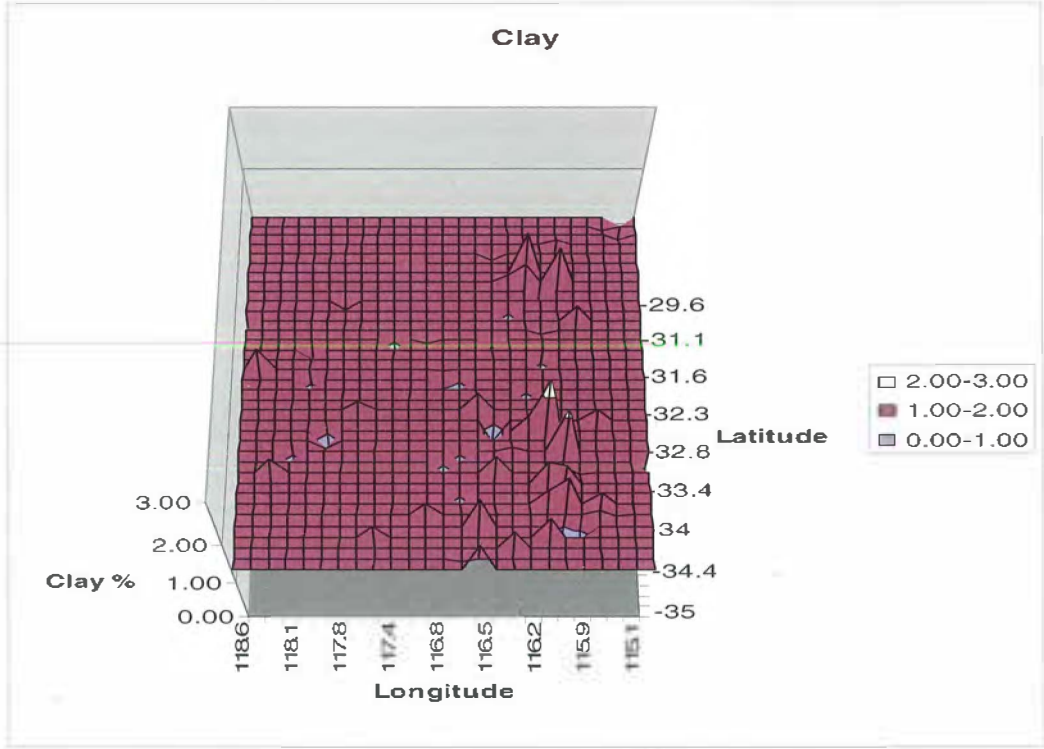
Loamy Gravel (Standardized data) – CAC03 %



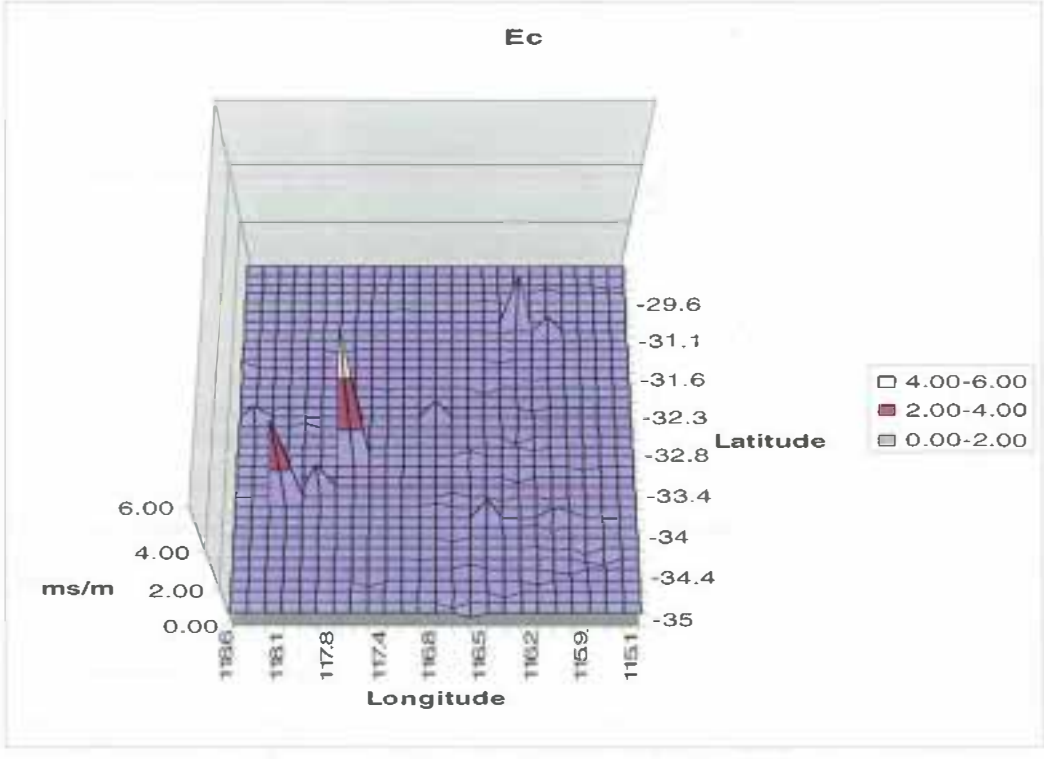
Loamy Gravel (Standardized data) – OC%



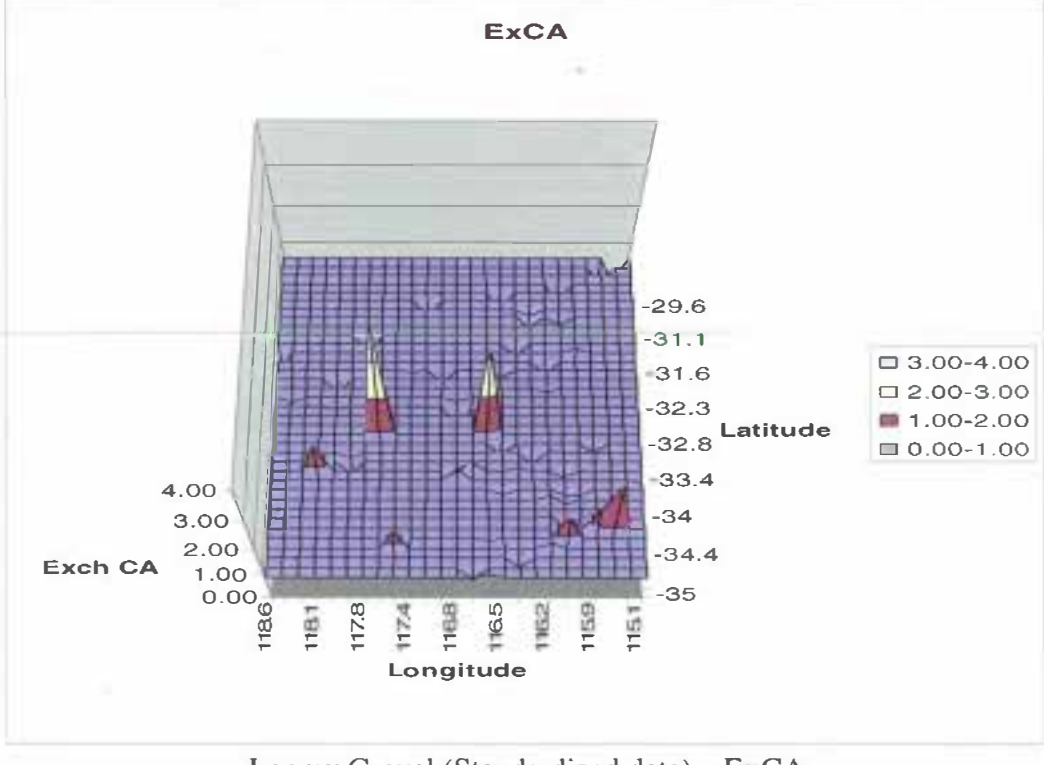
Loamy Gravel (Standardized data) – pH



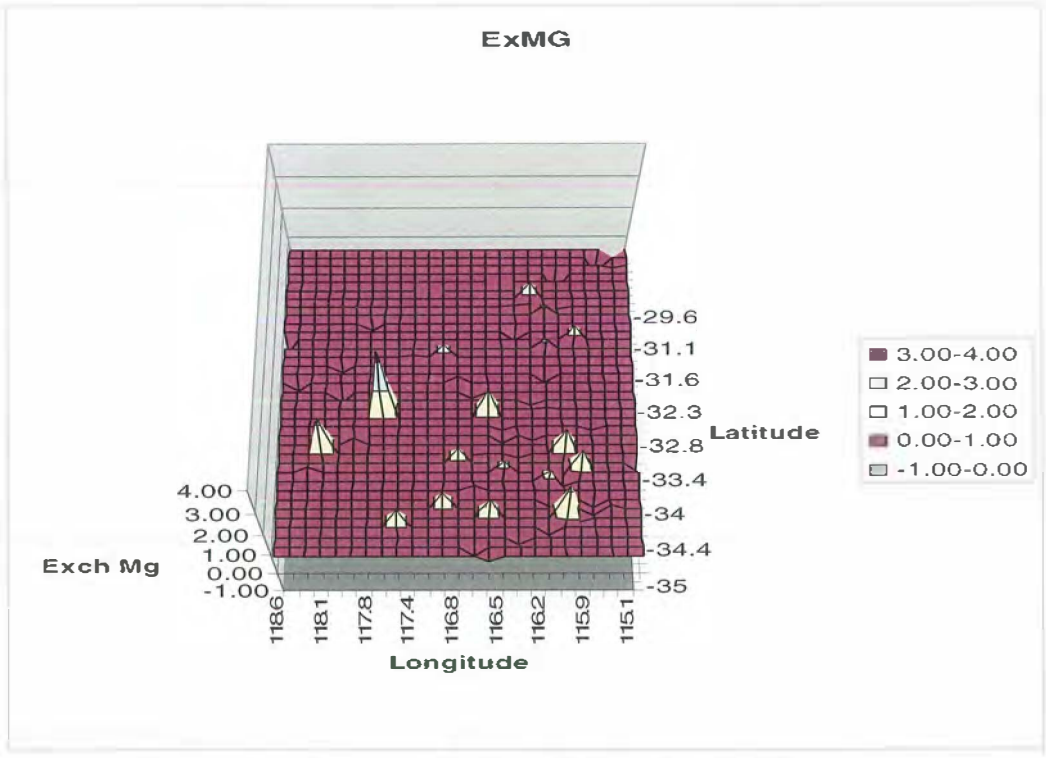
Loamy Gravel (Standardized data) – Clay



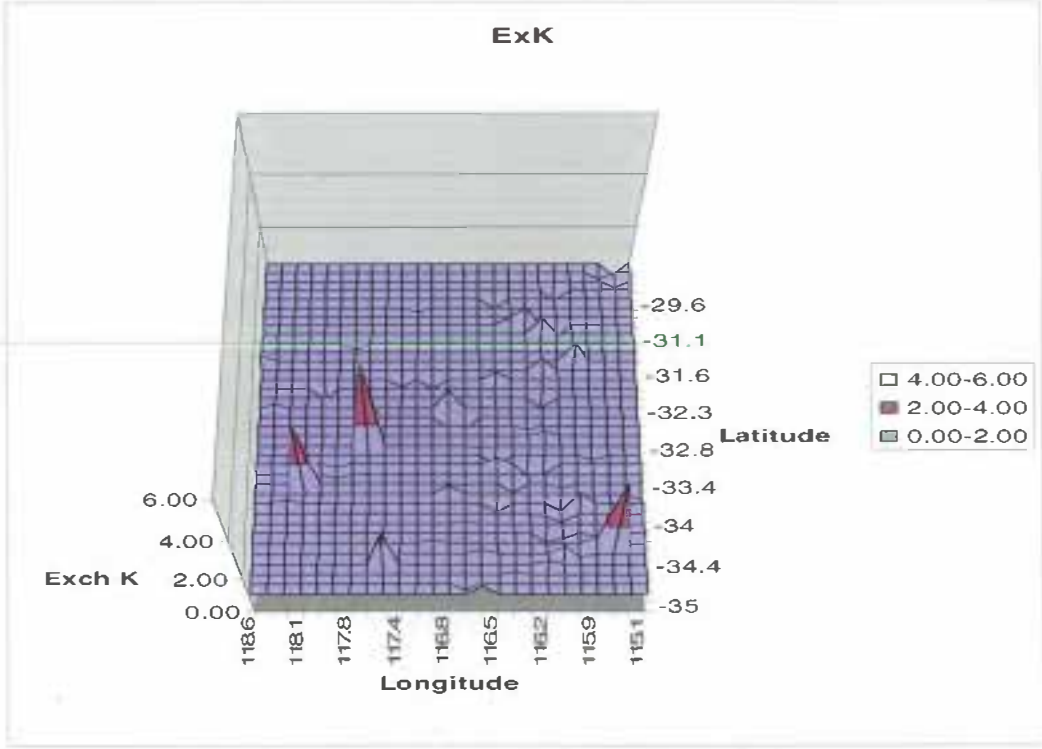
Loamy Gravel (Standardized data) – EC



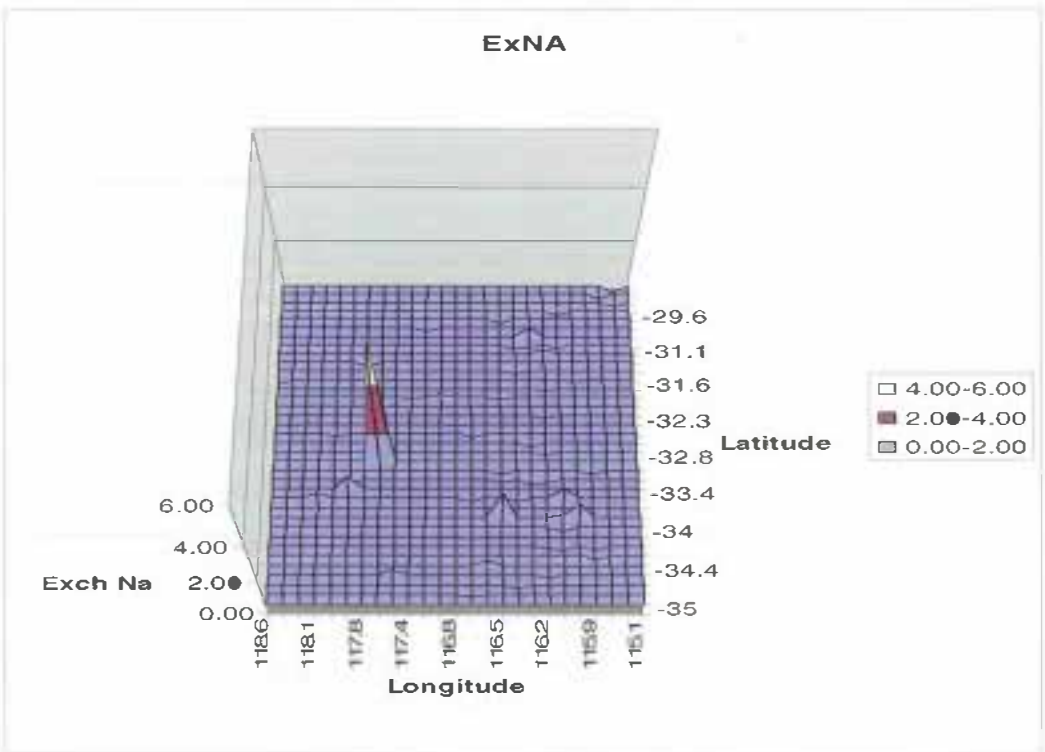
Loamy Gravel (Standardized data) – ExCA



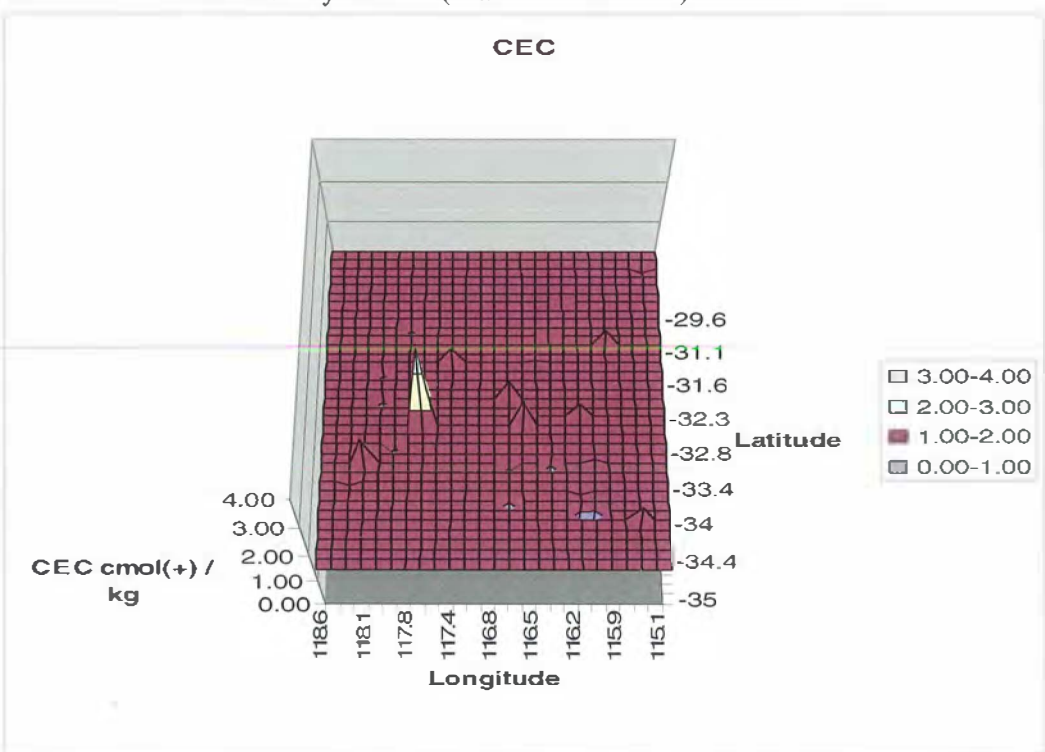
Loamy Gravel (Standardized data) – ExMG



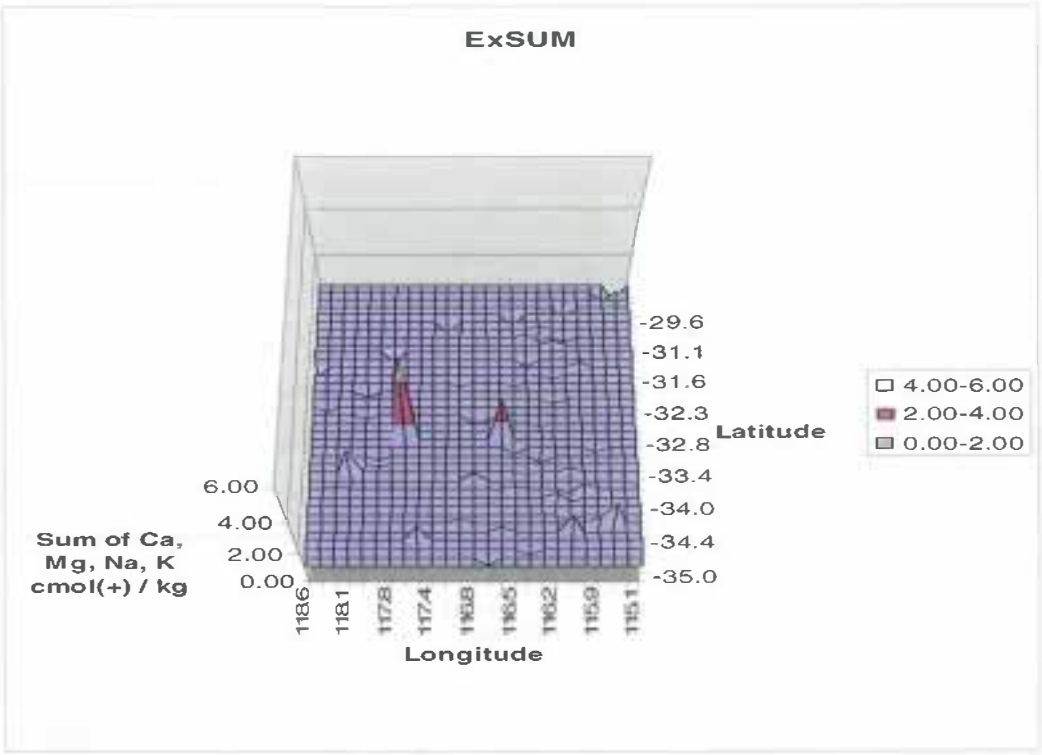
Loamy Gravel (Standardized data) – ExK



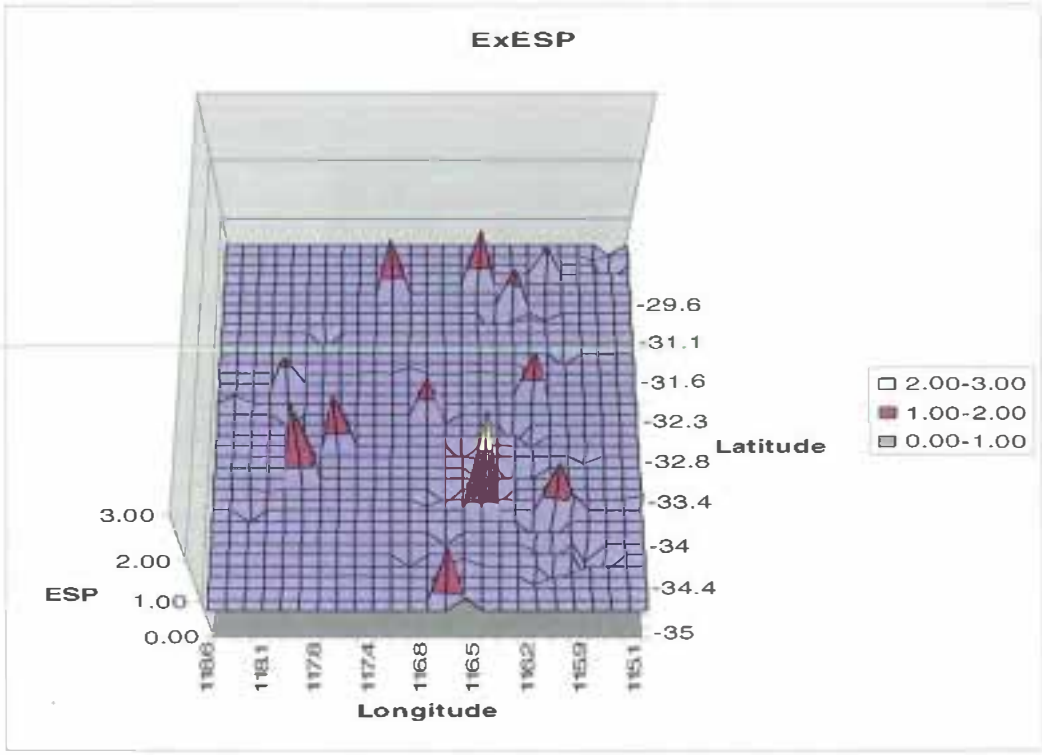
Loamy Gravel (Standardized data) – ExNA



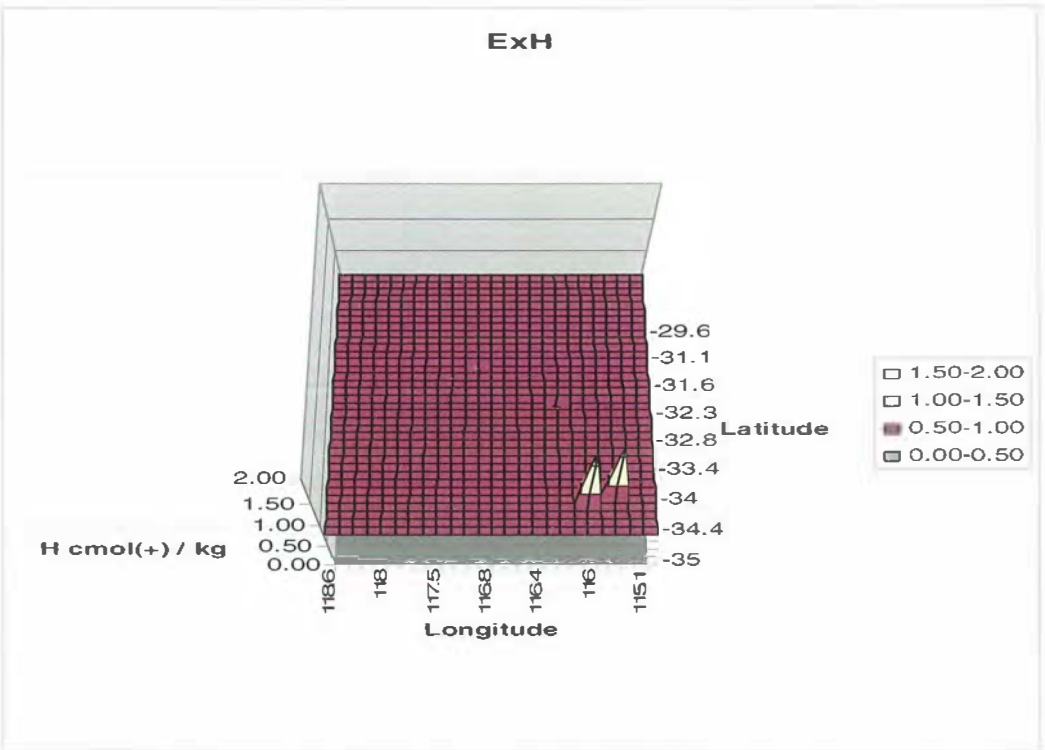
Loamy Gravel (Standardized data) – ExCEC



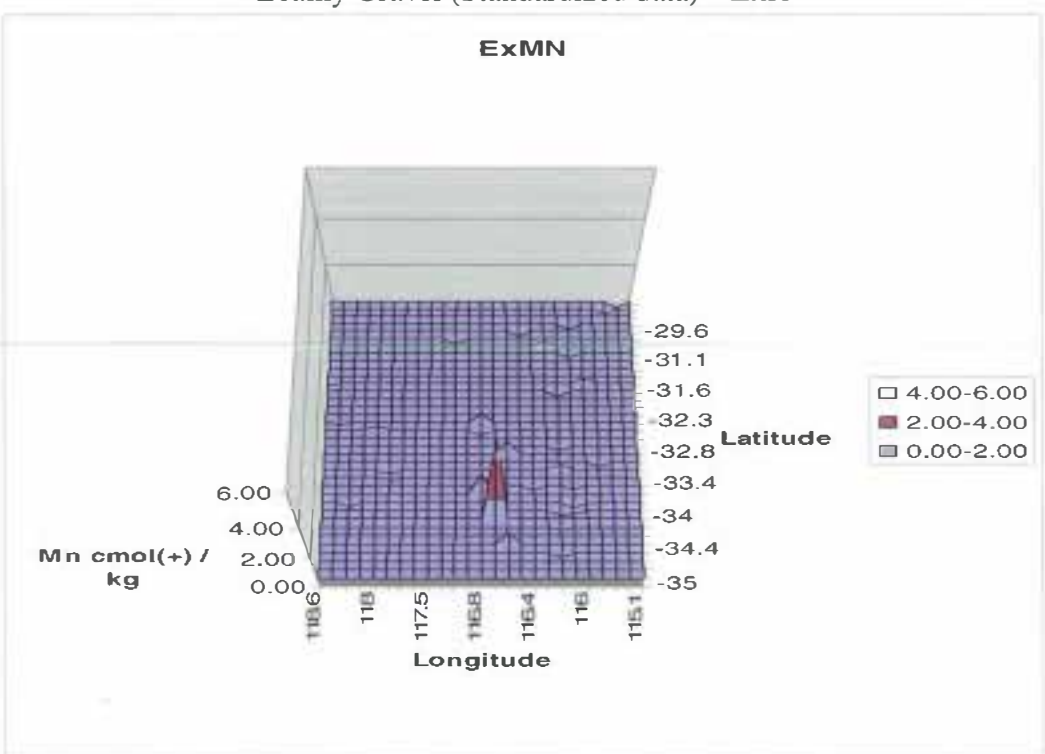
Loamy Gravel (Standardized data) – ExSUM



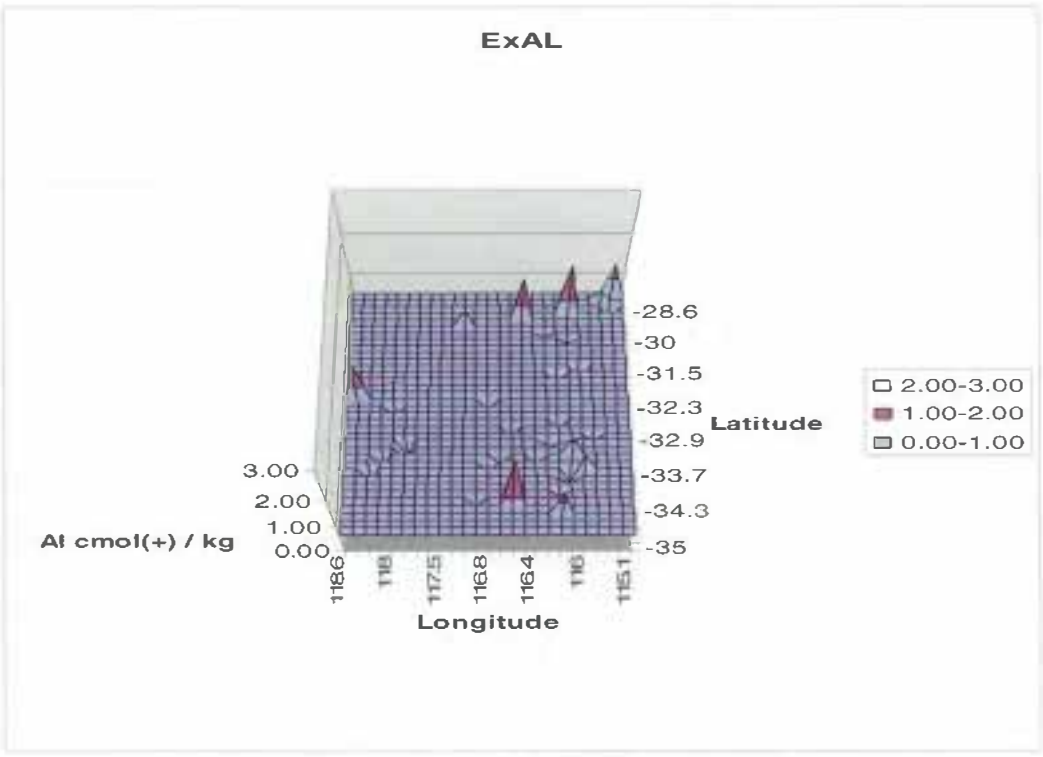
Loamy Gravel (Standardized data) – ExESP



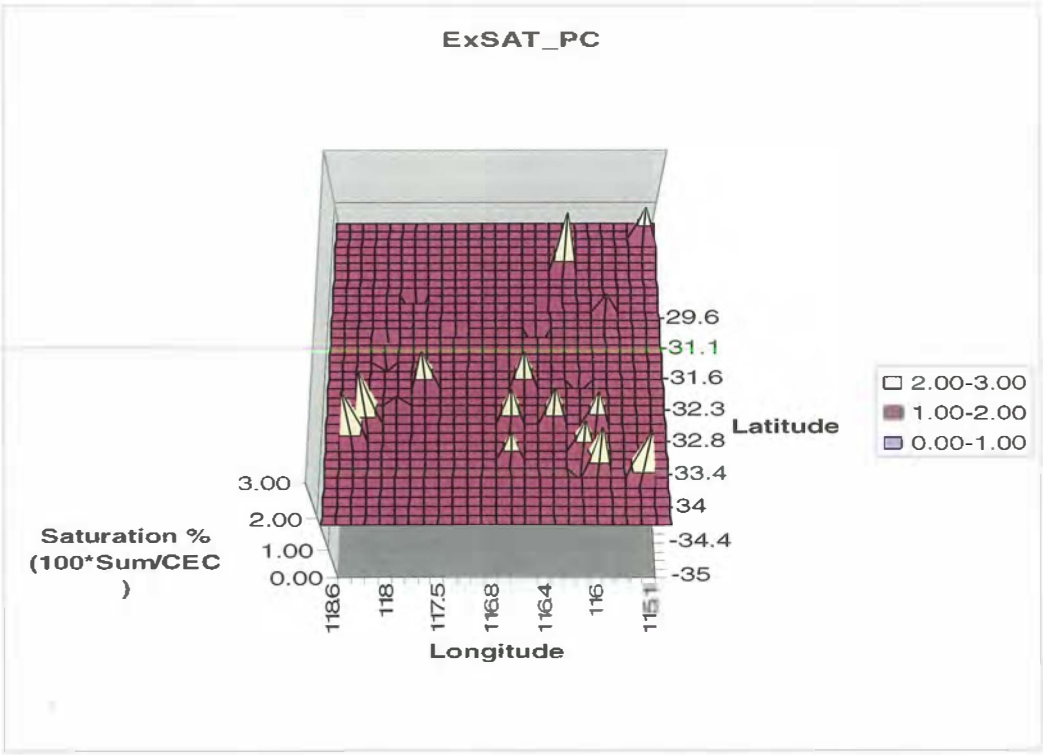
Loamy Gravel (Standardized data) – ExH



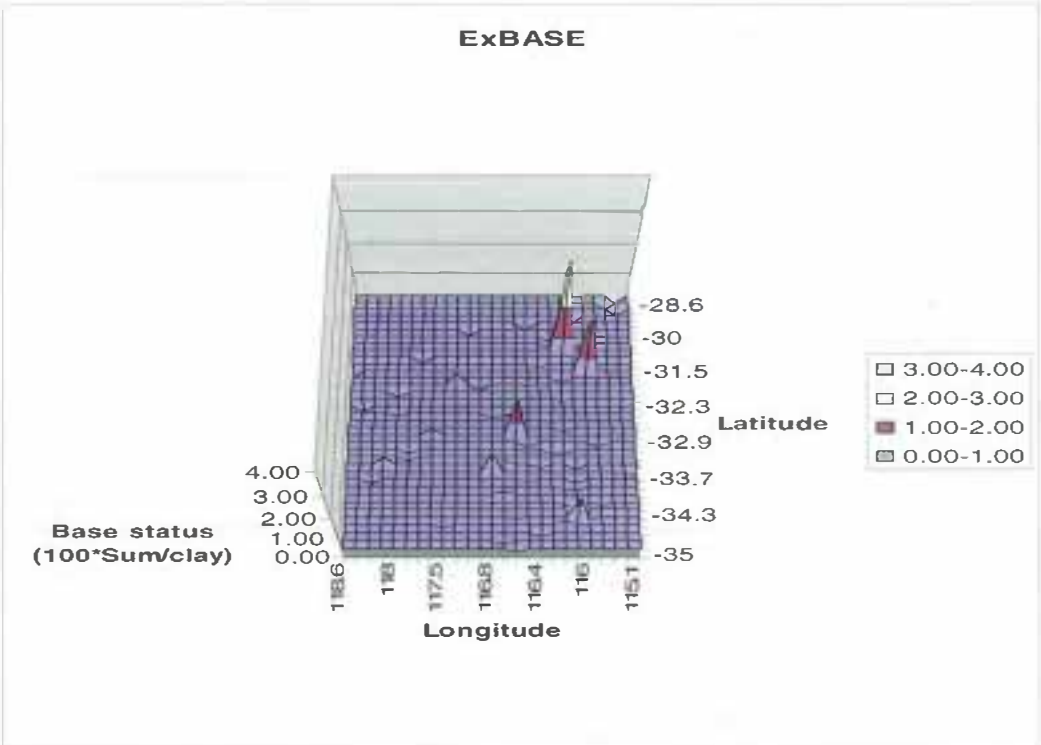
Loamy Gravel (Standardized data) – ExMN



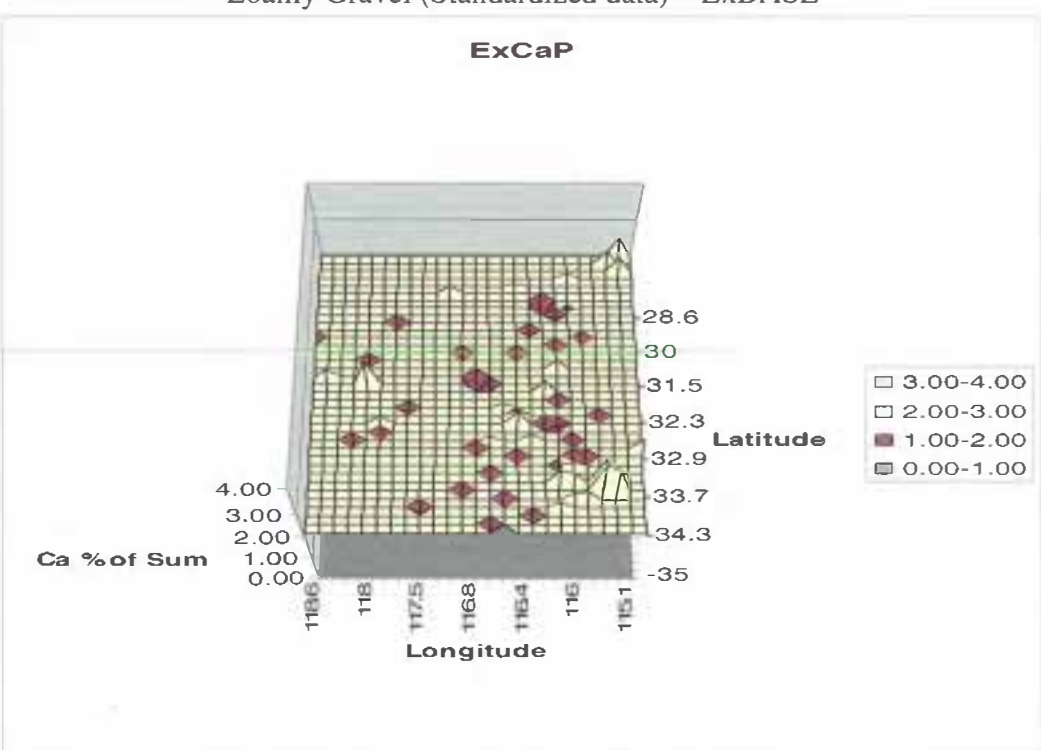
Loamy Gravel (Standardized data) – ExAL



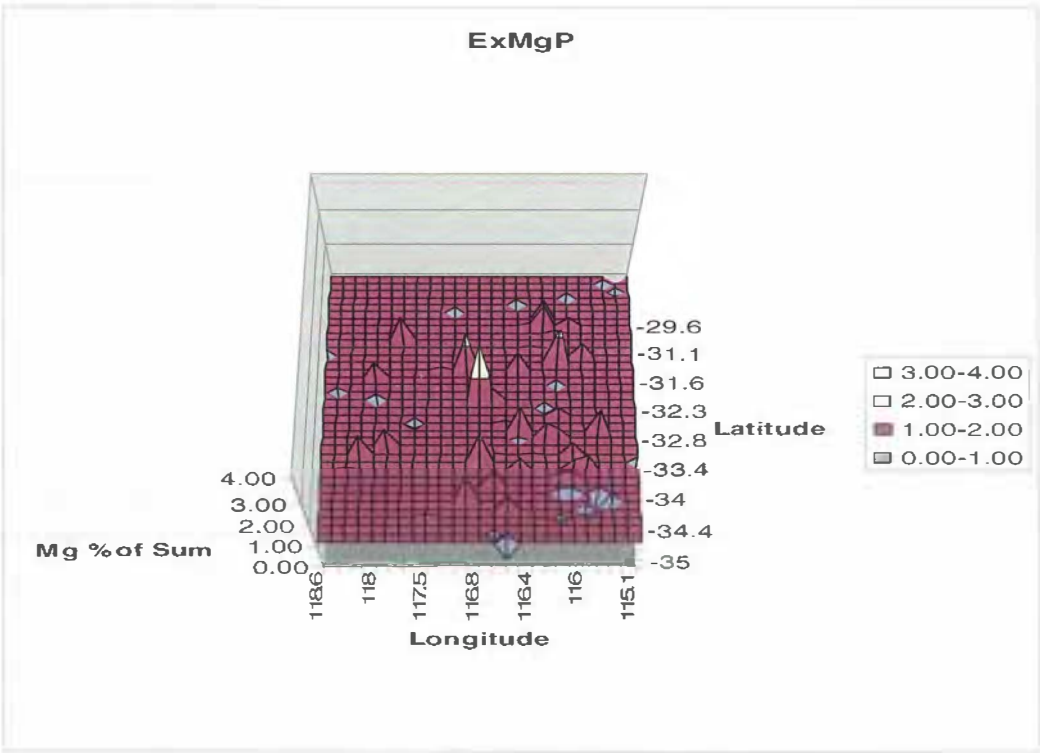
Loamy Gravel (Standardized data) – ExSAT_PC



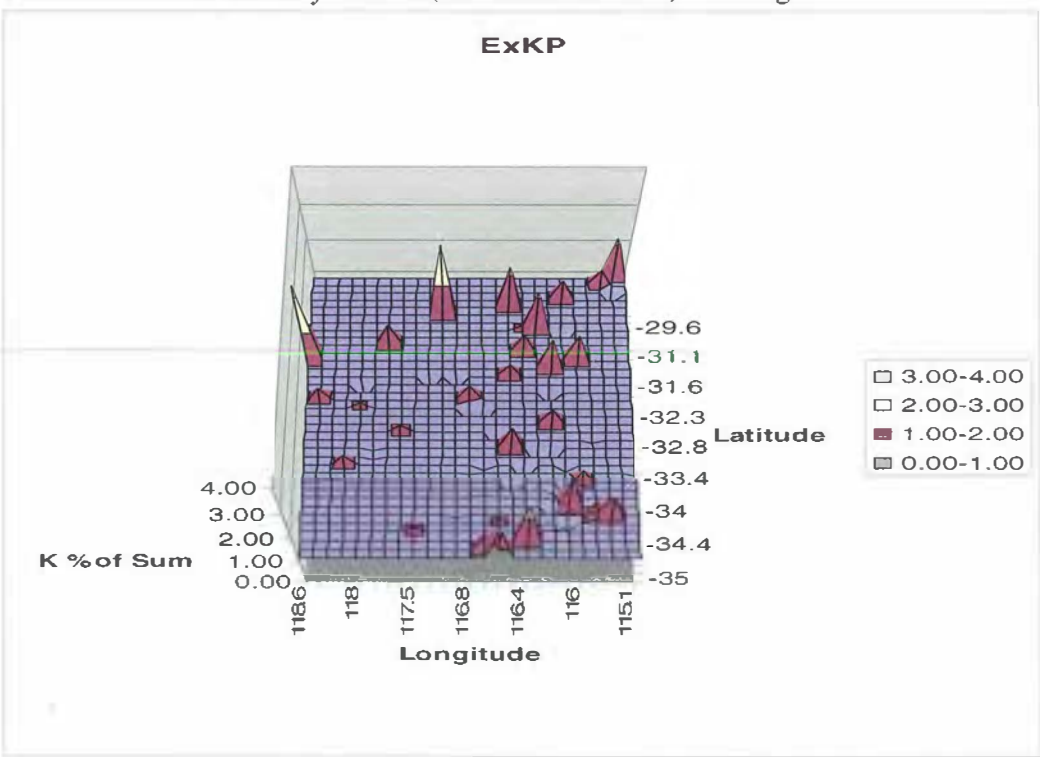
Loamy Gravel (Standardized data) – ExBASE



Loamy Gravel (Standardized data) – ExCaP

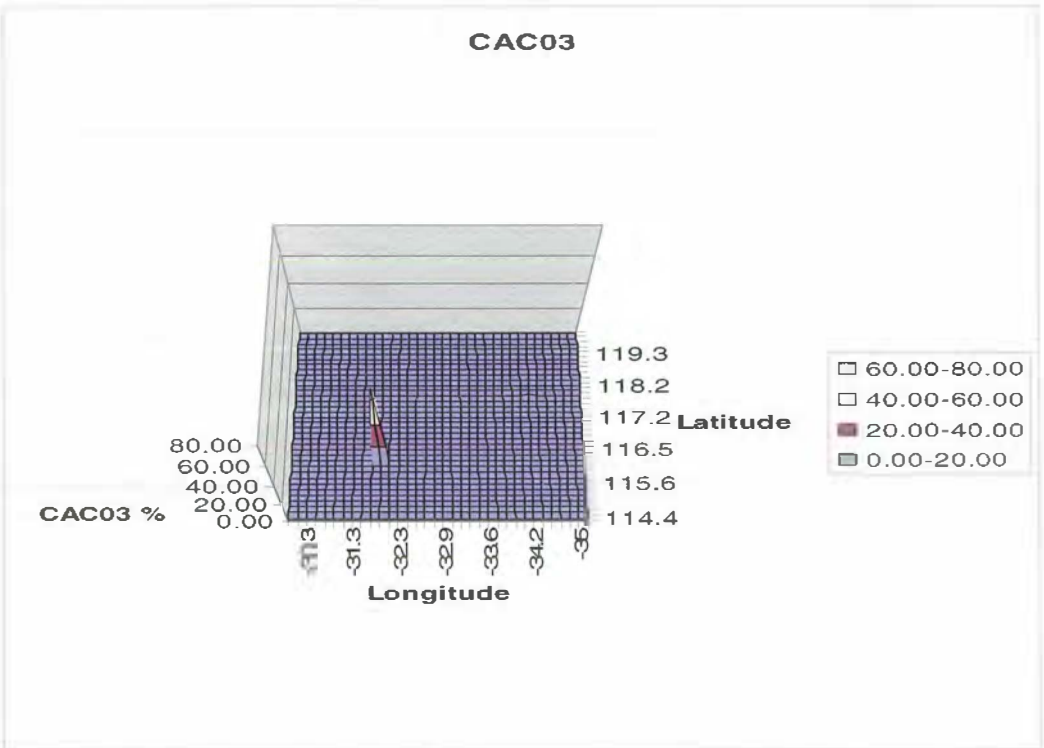


Loamy Gravel (Standardized data) – ExMgP

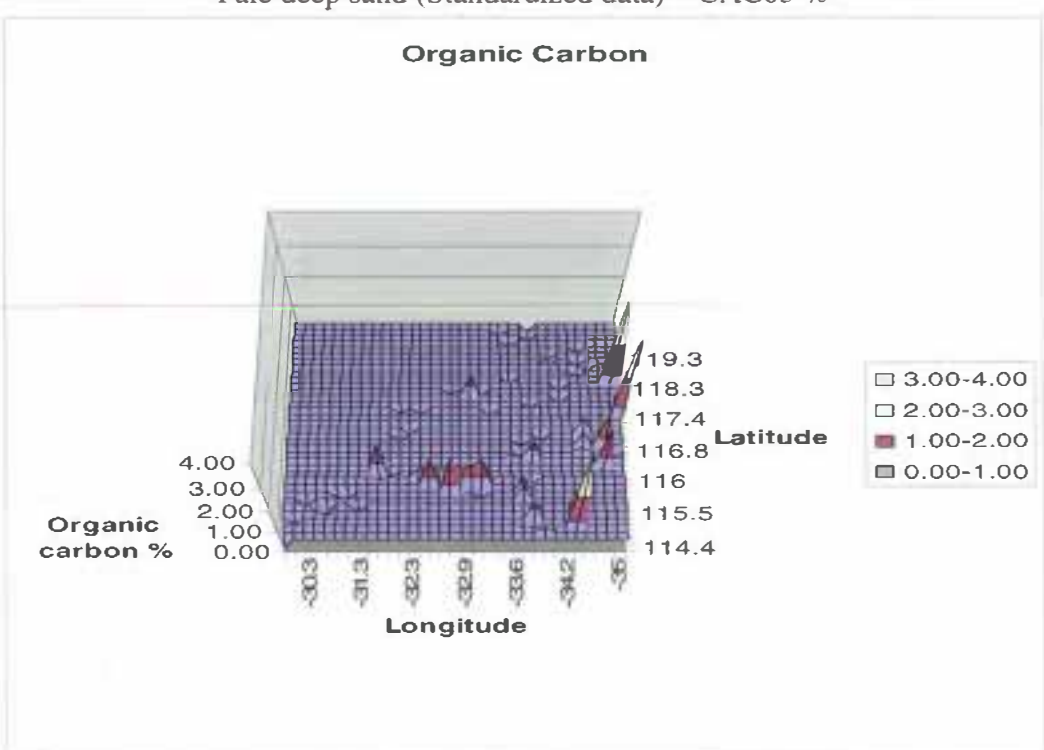


Loamy Gravel (Standardized data) – ExKP

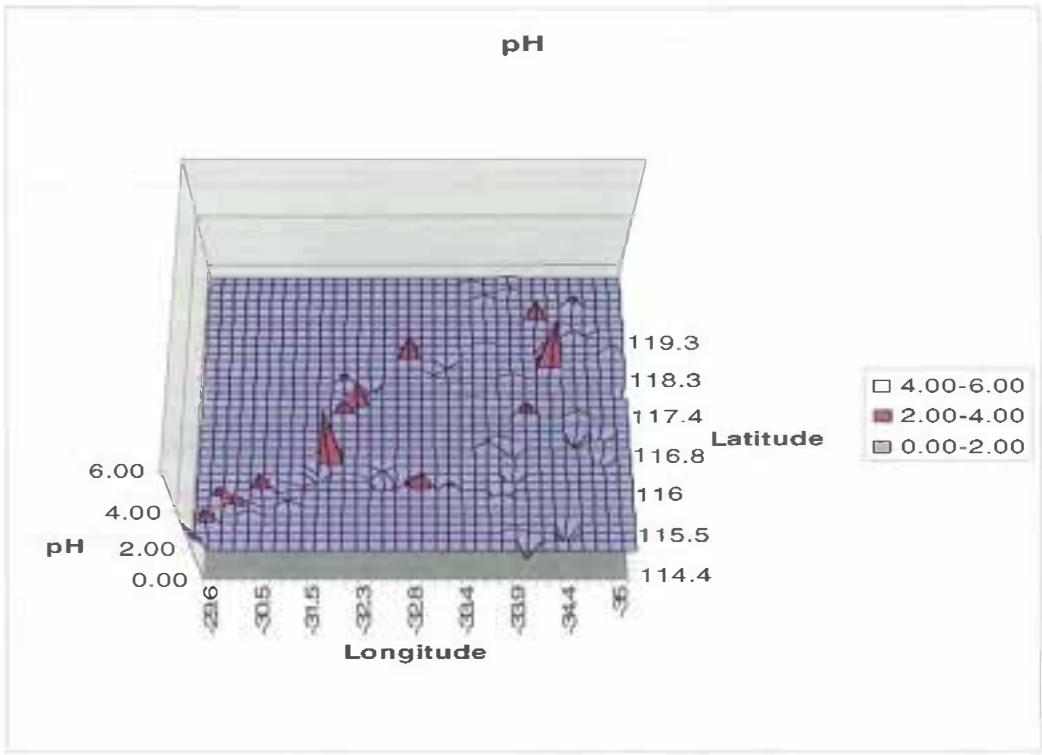
Soil 3 - Pale deep sand



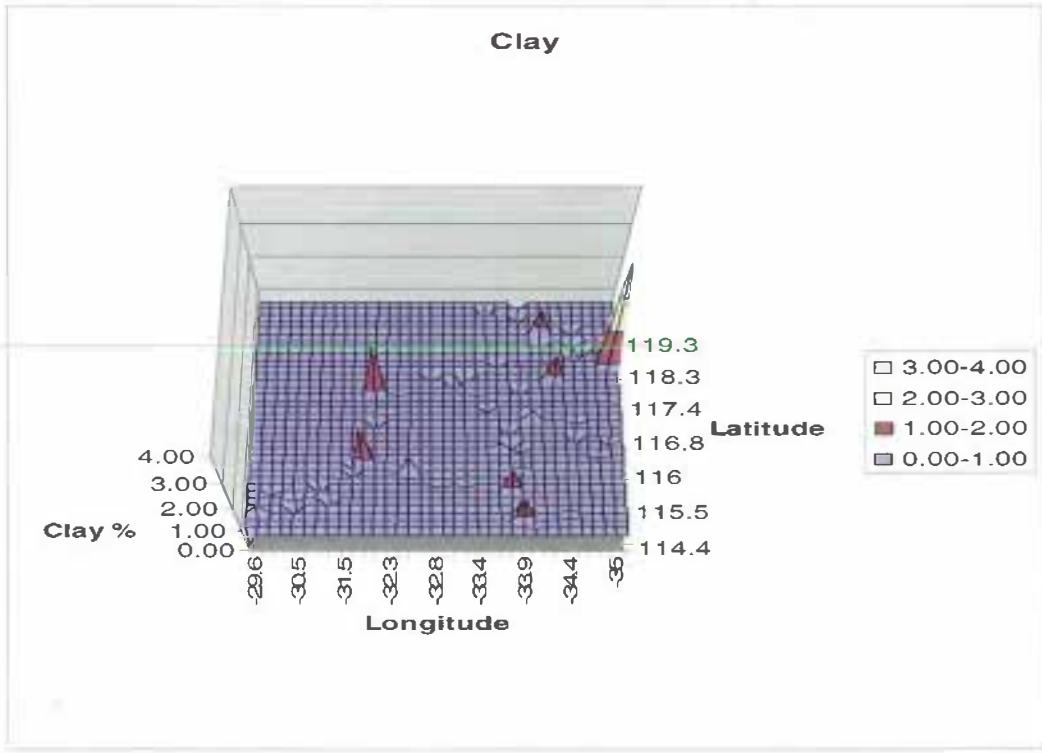
Pale deep sand (Standardized data) – CAC03 %



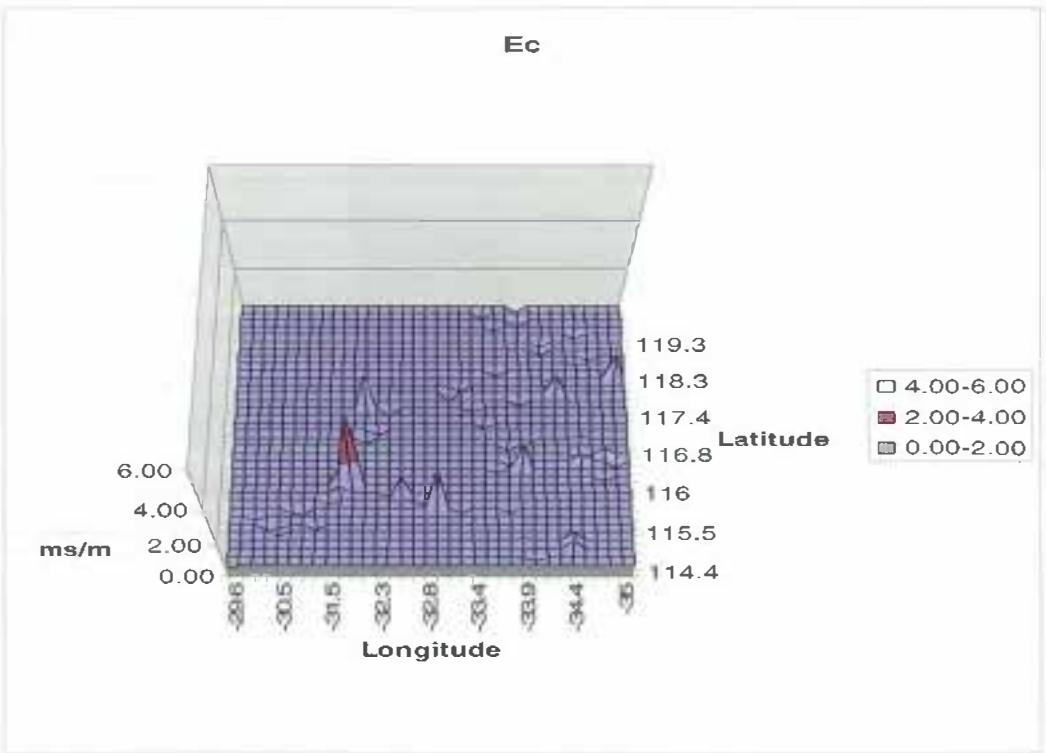
Pale deep sand (Standardized data) – OC %



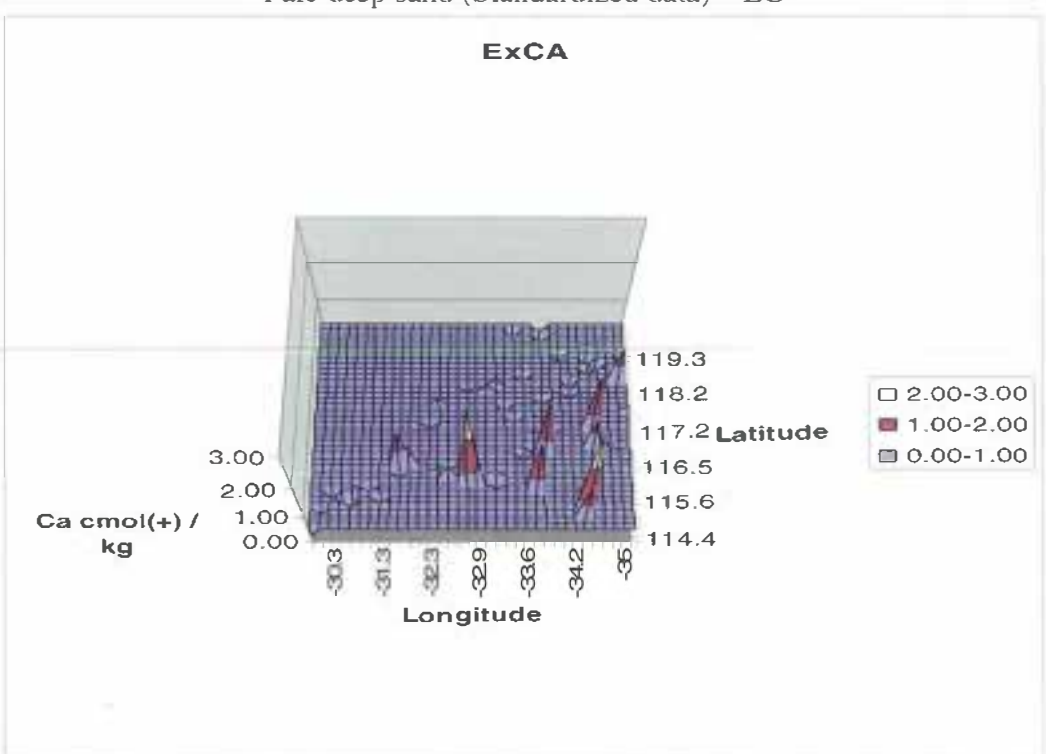
Pale deep sand (Standardized data) – pH



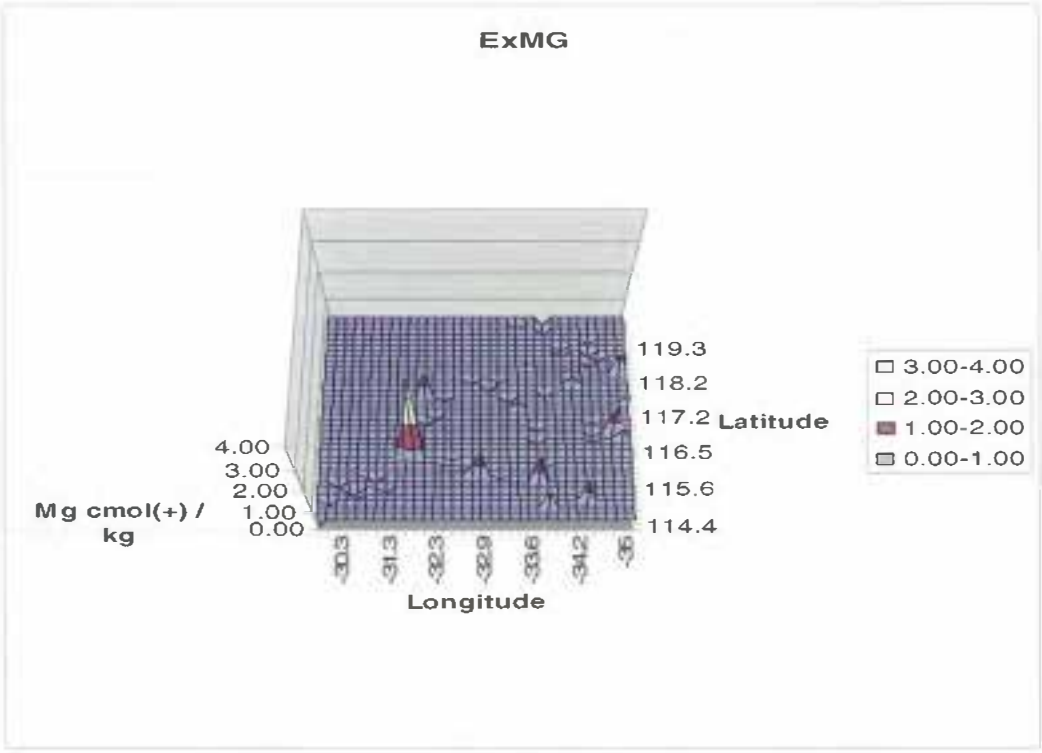
Pale deep sand (Standardized data) – Clay



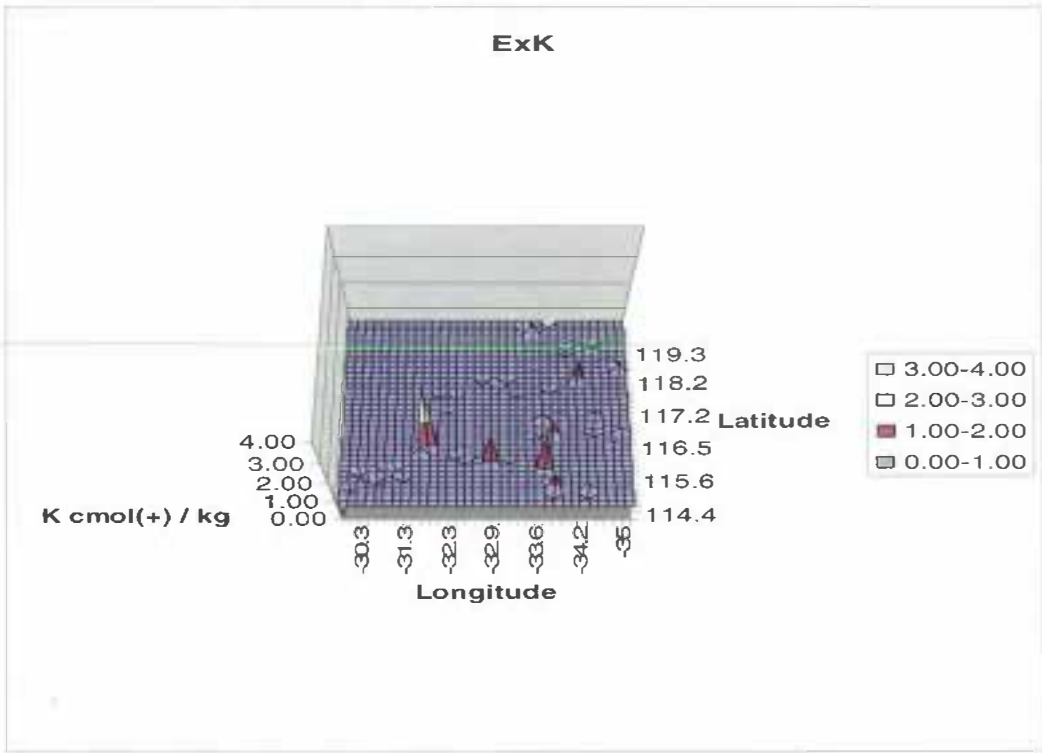
Pale deep sand (Standardized data) – EC



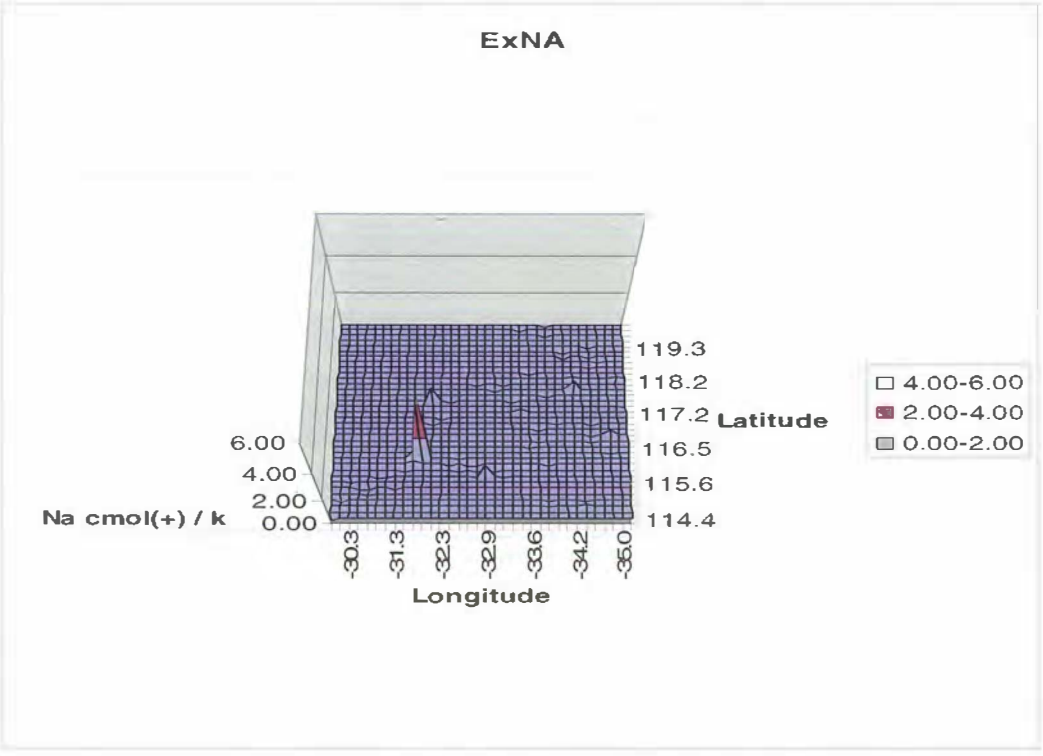
Pale deep sand (Standardized data) – ExCA



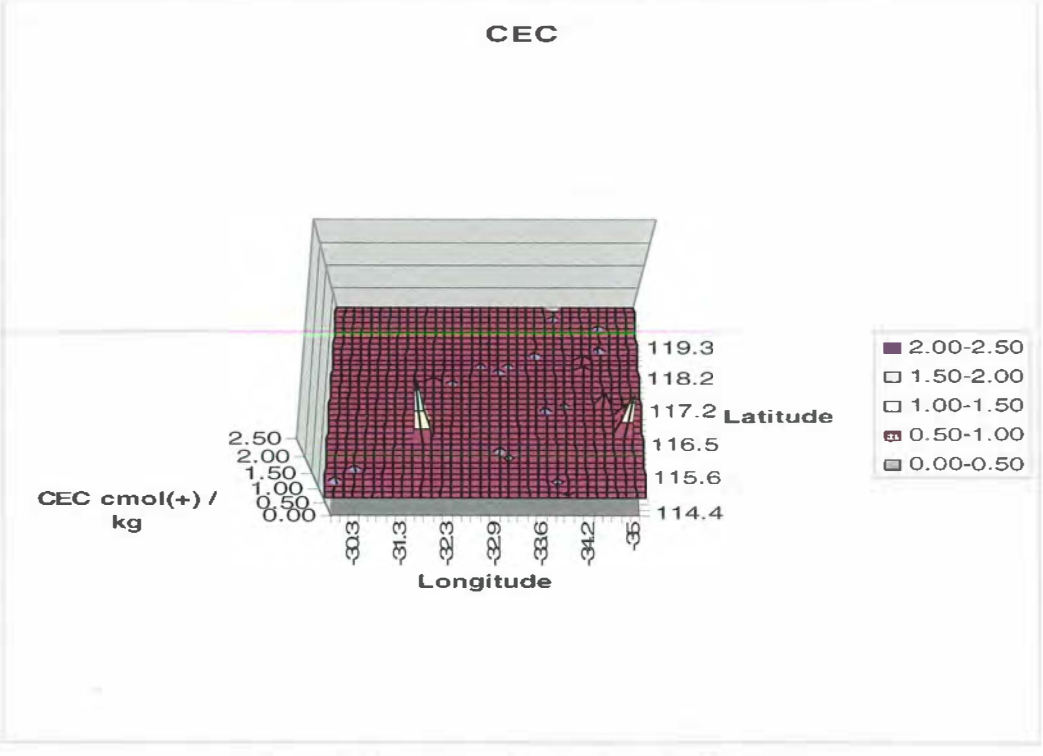
Pale deep sand (Standardized data) – ExMG



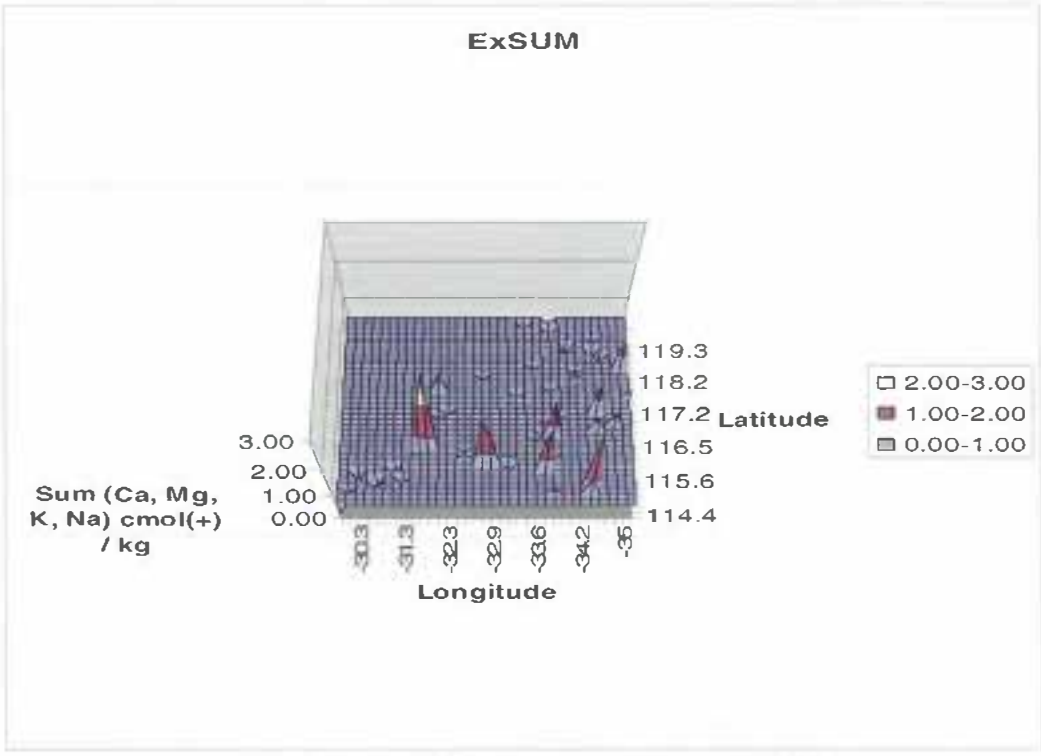
Pale deep sand (Standardized data) – ExK



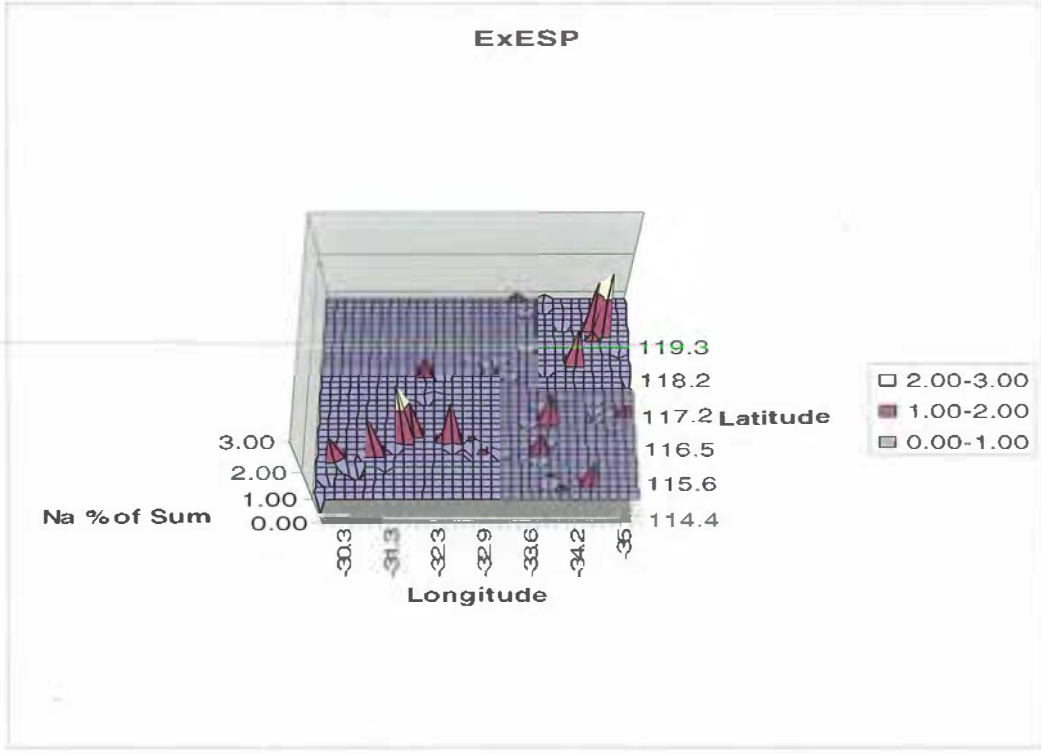
Pale deep sand (Standardized data) – ExNA



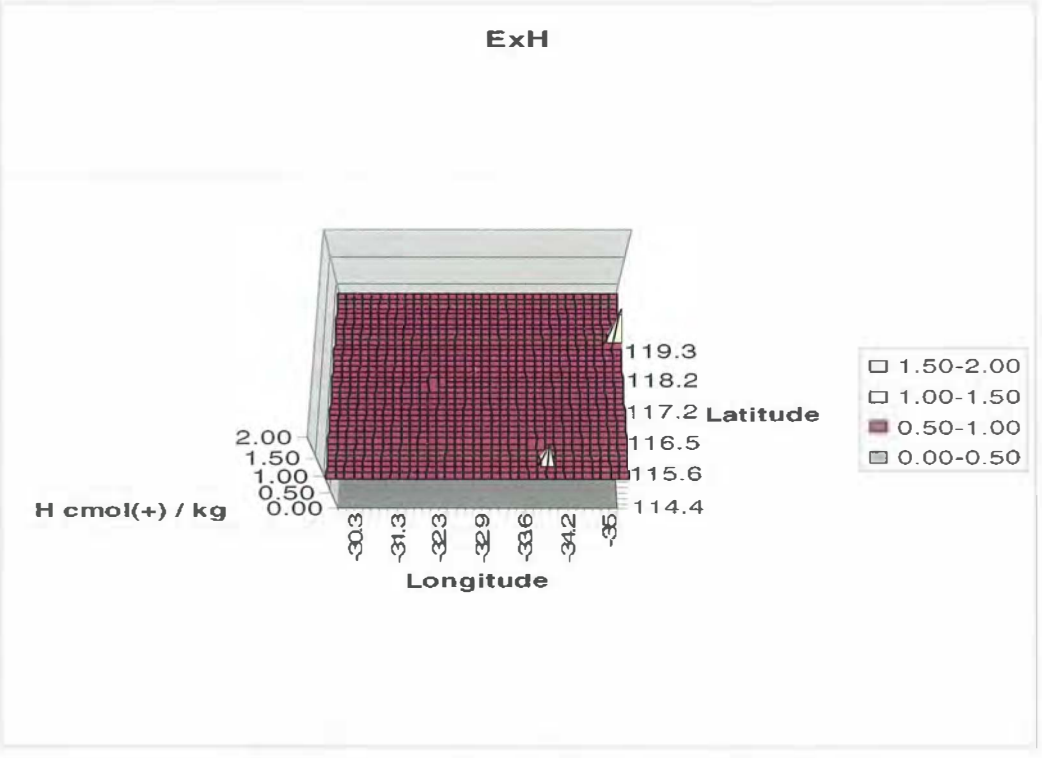
Pale deep sand (Standardized data) – ExCEC



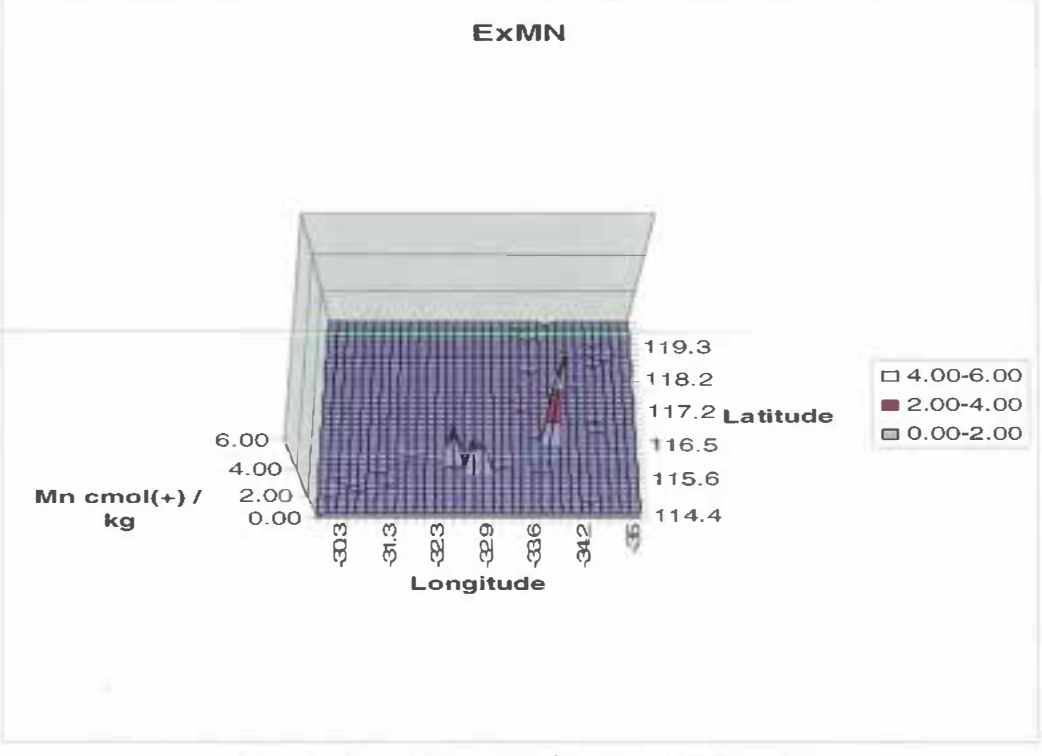
Pale deep sand (Standardized data) – ExSUM



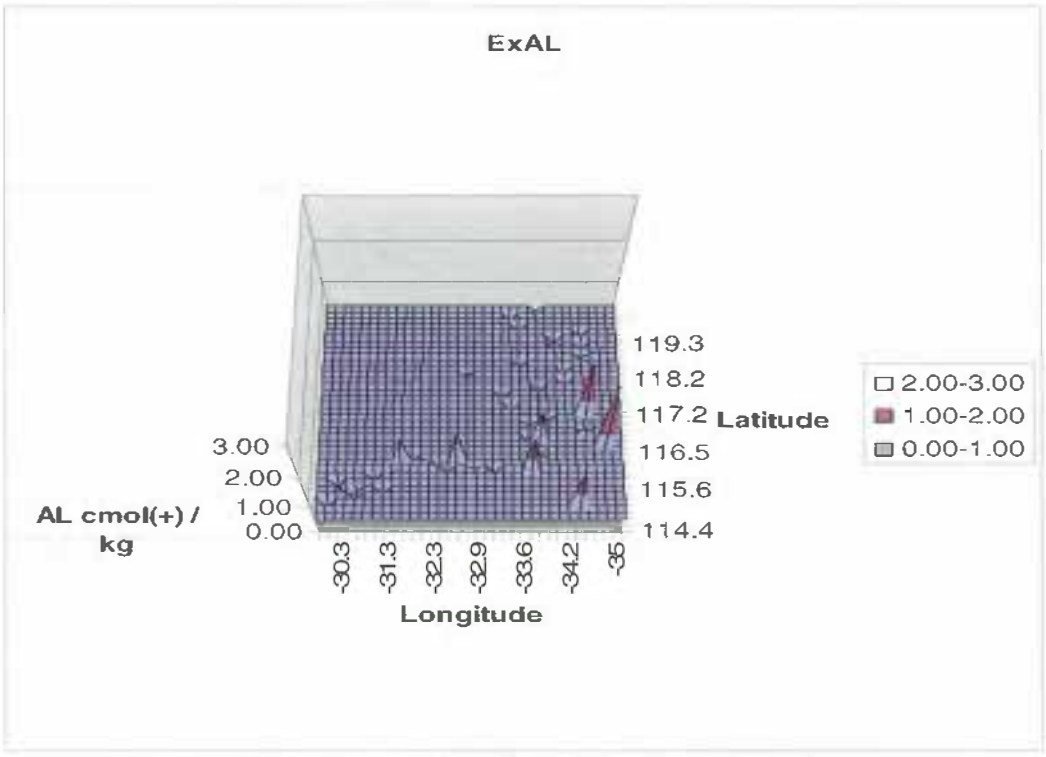
Pale deep sand (Standardized data) – ExESP



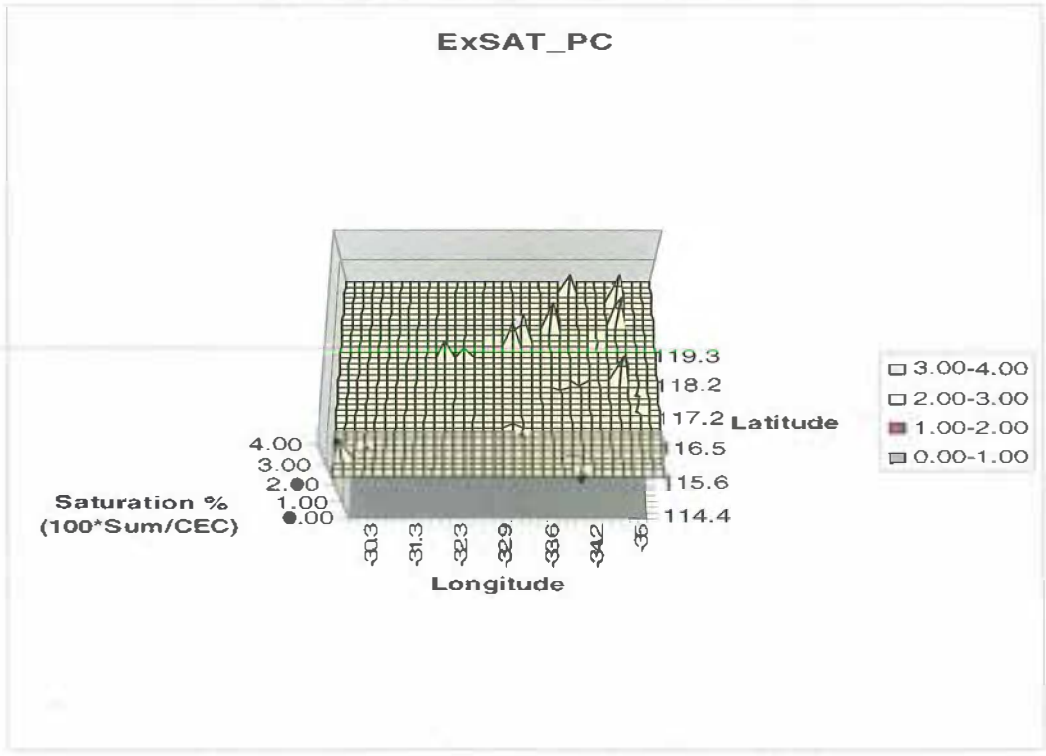
Pale deep sand (Standardized data) – ExH



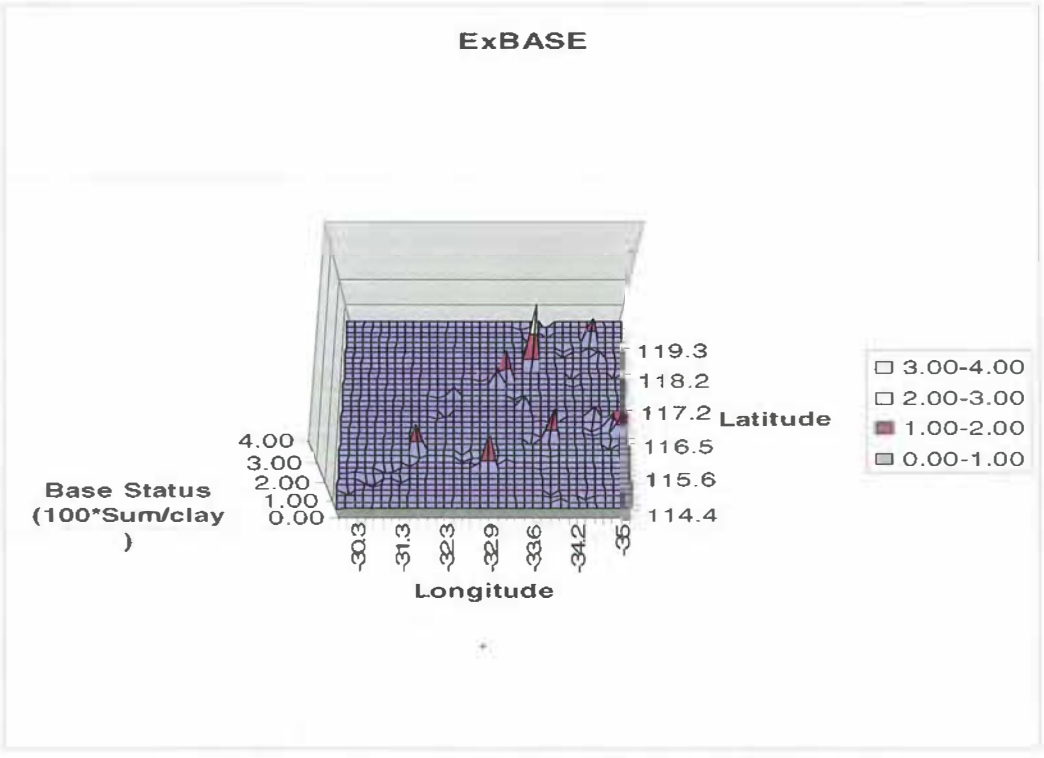
Pale deep sand (Standardized data) – ExMN



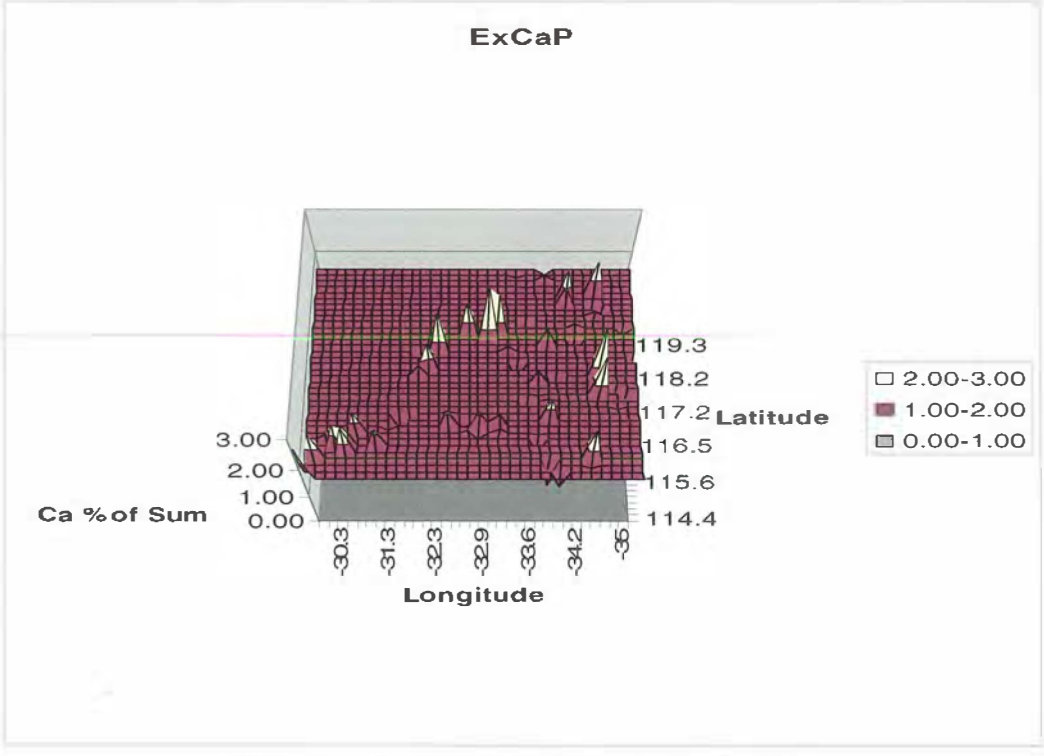
Pale deep sand (Standardized data) – ExAL



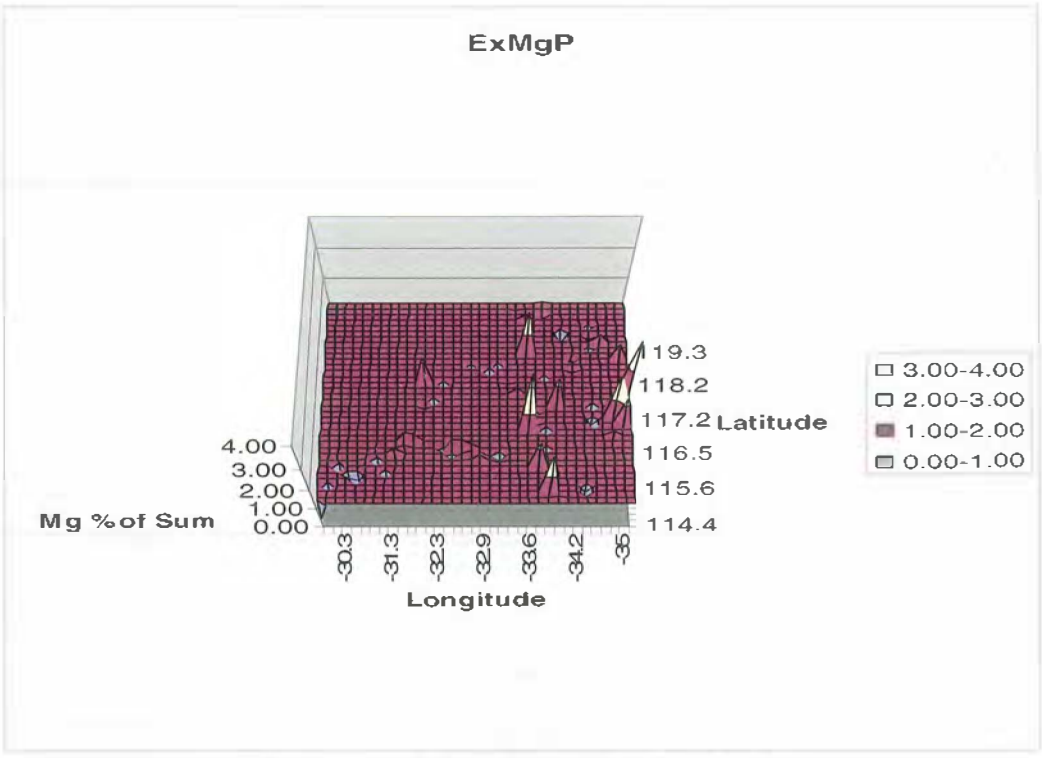
Pale deep sand (Standardized data) – ExSAT_PC



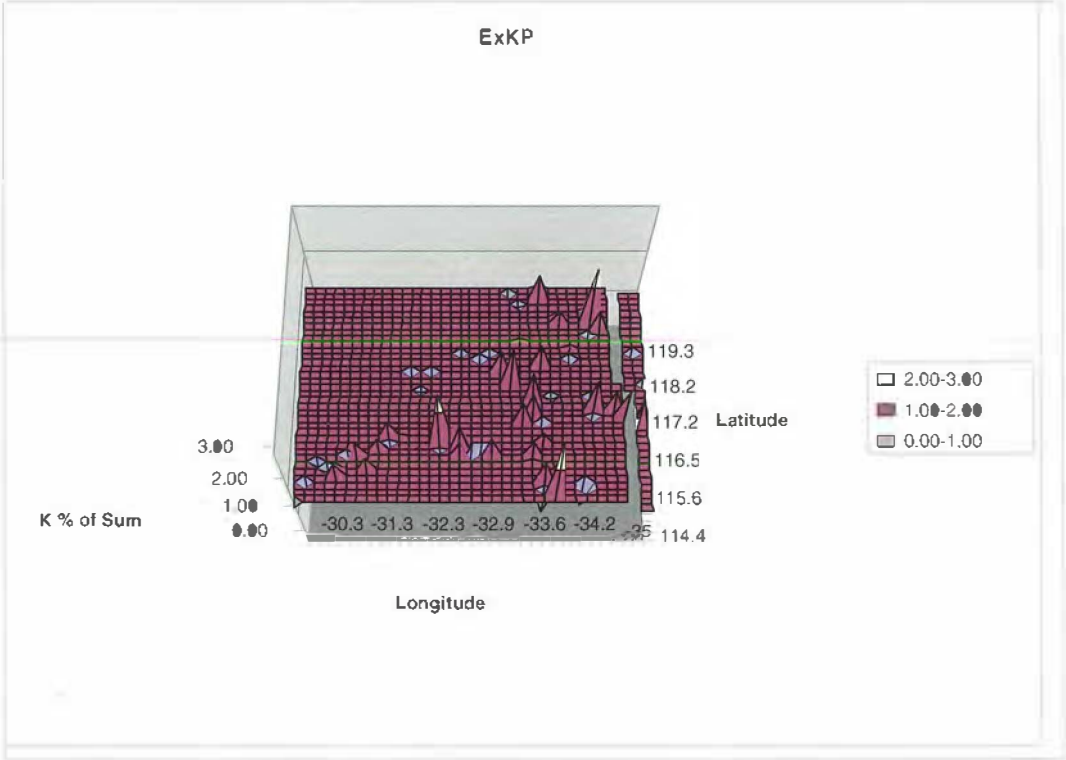
Pale deep sand (Standardized data) – ExBASE



Pale deep sand (Standardized data) – ExCaP



Pale deep sand (Standardized data) – ExMgP



Pale deep sand (Standardized data) – ExKP

8.4 Stage 6: Data mining – all soil types and three main soils

8.4.1 Data mining

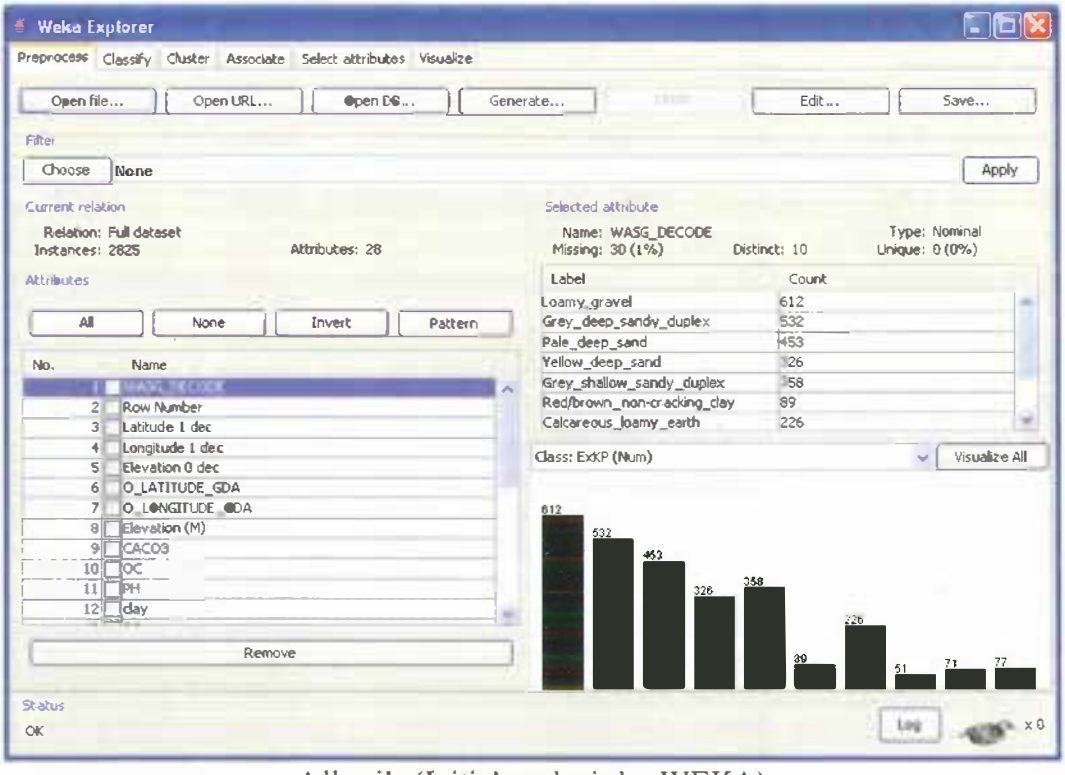
Table 3: EM clusters, all traits

Trait name	0	1
Row Number	Mean = 2472.8852 StdDev = 446.4977	Mean = 1762.1303 StdDev = 251.8598
Latitude 1 dec	Mean = -32.6095 StdDev = 1.5048	Mean = -33.2014 StdDev = 0.9406
Longitude 1 dec	Mean = 118.1097 StdDev = 1.3339	Mean = 116.1299 StdDev = 1.0682
Elevation 0 dec	Mean = 251.7012 StdDev = 91.266	Mean = 34.8555 StdDev = 28.2815
O_LATITUDE_GDA	Mean = -32.6081 StdDev = 1.5102	Mean = -33.2047 StdDev = 0.9425
O_LONGITUDE_GDA	Mean = 118.1047 StdDev = 1.3358	Mean = 116.1347 StdDev = 1.0654
Elevation (M)	Mean = 251.6632 StdDev = 91.31	Mean = 34.8681 StdDev = 28.2912
CACO3	Mean = 6.6707 StdDev = 10.1739	Mean = 4.5948 StdDev = 0.5996
OC	Mean = 0.7949 StdDev = 0.6918	Mean = 1.3672 StdDev = 1.2519
PH	Mean = 7.3324 StdDev = 1.2027	Mean = 4.7107 StdDev = 0.7443
clay	Mean = 36.3729 StdDev = 14.2696	Mean = 4.3644 StdDev = 3.0747
EC	Mean = 81.003 StdDev = 116.7387	Mean = 9.4558 StdDev = 9.1395
ExCA	Mean = 5.6769 StdDev = 5.8022	Mean = 1.8936 StdDev = 1.965
ExMG	Mean = 5.8662 StdDev = 3.7478	Mean = 1.1885 StdDev = 1.0409
ExK	Mean = 0.8893 StdDev = 0.8857	Mean = 0.1553 StdDev = 0.1317
ExNA	Mean = 3.3174 StdDev = 3.837	Mean = 0.4842 StdDev = 0.4726
ExCEC	Mean = 15.8523 StdDev = 7.734	Mean = 10.8582 StdDev = 0.765
ExSUM	Mean = 15.7444 StdDev = 9.2264	Mean = 3.7209 StdDev = 3.0155
ExESP	Mean = 19.3353 StdDev = 15.4593	Mean = 13.1548 StdDev = 12.1574
ExH	Mean = 4.372 StdDev = 0.0542	Mean = 4.3756 StdDev = 0.0002
ExMN	Mean = 0.0228 StdDev = 0.0542	Mean = 0.0221 StdDev = 0.0116
ExAL	Mean = 0.1799 StdDev = 0.1892	Mean = 0.2178 StdDev = 0.3041
ExSAT_PC	Mean = 93.1221 StdDev = 7.6628	Mean = 84.1632 StdDev = 3.4362
ExBASE	Mean = 64.0817 StdDev = 54.3474	Mean = 89.6101 StdDev = 119.7516
ExCaP	Mean = 36.4587 StdDev = 21.5496	Mean = 51.5759 StdDev = 18.7696
ExMgP	Mean = 38.2712 StdDev = 12.8669	Mean = 30.0874 StdDev = 12.4427
ExKP	Mean = 5.879 StdDev = 4.9544	Mean = 5.1496 StdDev = 3.4323

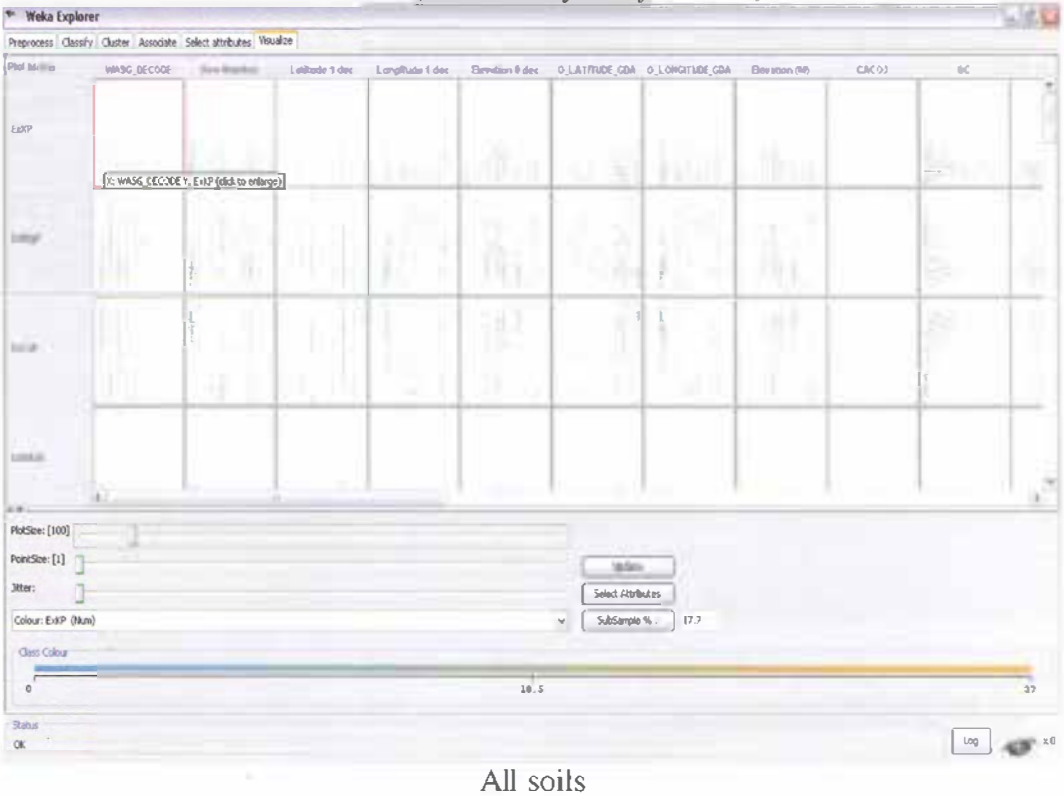
Trait name	2	3
Row Number	Mean = 966.5342 StdDev = 436.2244	Mean = 2175.6498 StdDev = 215.3224
Latitude 1 dec	Mean = -33.3844 StdDev = 0.1922	Mean = -31.2644 StdDev = 2.1574
Longitude 1 dec	Mean = 121.0203 StdDev = 1.6295	Mean = 116.3206 StdDev = 1.5368
Elevation 0 dec	Mean = 200.71 StdDev = 78.9579	Mean = 218.348 StdDev = 90.0006
O_LATITUDE_GDA	Mean = -33.3922 StdDev = 0.184	Mean = -31.2685 StdDev = 2.1614
O_LONGITUDE_GDA	Mean = 121.0108 StdDev = 1.6272	Mean = 116.3182 StdDev = 1.5343
Elevation (M)	Mean = 200.5589 StdDev = 78.9665	Mean = 218.3555 StdDev = 90.0096
CACO3	Mean = 4.2152 StdDev = 4.0859	Mean = 4.396 StdDev = 1.0769
OC	Mean = 0.3163 StdDev = 0.3172	Mean = 0.5001 StdDev = 0.4684
PH	Mean = 7.6645 StdDev = 1.058	Mean = 5.2355 StdDev = 0.6621
clay	Mean = 32.6945 StdDev = 13.4789	Mean = 6.3571 StdDev = 5.3764
EC	Mean = 74.9496 StdDev = 68.6524	Mean = 4.8951 StdDev = 7.6303
ExCA	Mean = 3.4515 StdDev = 4.9585	Mean = 0.8303 StdDev = 0.8625
ExMG	Mean = 5.1671 StdDev = 2.9821	Mean = 0.6339 StdDev = 0.8244
ExK	Mean = 1.4014 StdDev = 0.9926	Mean = 0.0965 StdDev = 0.112
ExNA	Mean = 4.7786 StdDev = 3.5632	Mean = 0.2564 StdDev = 0.4031
ExCEC	Mean = 14.545 StdDev = 6.3112	Mean = 8.9379 StdDev = 3.8546
ExSUM	Mean = 14.8059 StdDev = 8.5194	Mean = 1.8168 StdDev = 2.1142
ExESP	Mean = 29.9349 StdDev = 15.3175	Mean = 10.4342 StdDev = 10.0318
ExH	Mean = 4.341 StdDev = 0.3245	Mean = 4.3052 StdDev = 0.4722
ExMN	Mean = 0.0227 StdDev = 0.0014	Mean = 0.0168 StdDev = 0.0071
ExAL	Mean = 0.1815 StdDev = 0.016	Mean = 0.1445 StdDev = 0.1376
ExSAT_PC	Mean = 90.6496 StdDev = 9.168	Mean = 80.0399 StdDev = 14.7465
ExBASE	Mean = 90.6496 StdDev = 9.168	Mean = 31.1979 StdDev = 27.6376
ExCaP	Mean = 25.7055 StdDev = 20.538	Mean = 53.0512 StdDev = 18.2863
ExMgP	Mean = 34.8423 StdDev = 9.4242	Mean = 29.9027 StdDev = 14.4135
ExKP	Mean = 9.6186 StdDev = 4.97	Mean = 6.6532 StdDev = 4.7559

Trait name	4	5
Row Number	Mean = 325.8539 StdDev = 251.8011	Mean = 1517.4171 StdDev = 422.1281
Latitude 1 dec	Mean = -33.1763 StdDev = 1.3705	Mean = -33.6811 StdDev = 0.9219
Longitude 1 dec	Mean = 116.2466 StdDev = 0.3763	Mean = 117.4722 StdDev = 1.0287
Elevation 0 dec	Mean = 220.6804 StdDev = 72.4535	Mean = 223.1119 StdDev = 101.3594
O_LATITUDE_GDA	Mean = -33.181 StdDev = 1.3718	Mean = -33.6847 StdDev = 0.9221
O_LONGITUDE_GDA	Mean = 116.2419 StdDev = 0.3825	Mean = 117.4735 StdDev = 1.0268
Elevation (M)	Mean = 220.6826 StdDev = 72.5041	Mean = 223.0987 StdDev = 101.3459
CACO3	Mean = 4.6731 StdDev = 4.1363	Mean = 4.509 StdDev = 0.8576
OC	Mean = 1.9384 StdDev = 2.1976	Mean = 1.0244 StdDev = 1.0824
PH	Mean = 5.4371 StdDev = 0.622	Mean = 5.2929 StdDev = 0.6971
clay	Mean = 21.7505 StdDev = 15.6136	Mean = 33.9314 StdDev = 19.4208
EC	Mean = 5.761 StdDev = 7.3761	Mean = 36.7121 StdDev = 50.8344
ExCA	Mean = 2.9112 StdDev = 3.2815	Mean = 2.1683 StdDev = 1.6833
ExMG	Mean = 1.6481 StdDev = 1.8717	Mean = 2.9421 StdDev = 1.715
ExK	Mean = 0.2185 StdDev = 0.2337	Mean = 0.2642 StdDev = 0.1773
ExNA	Mean = 0.3799 StdDev = 0.6184	Mean = 1.3278 StdDev = 1.0825
ExCEC	Mean = 10.6577 StdDev = 1.4176	Mean = 10.7797 StdDev = 0.7924
ExSUM	Mean = 5.1527 StdDev = 4.5524	Mean = 6.6925 StdDev = 3.1751
ExESP	Mean = 6.5027 StdDev = 5.8488	Mean = 16.8824 StdDev = 11.3667
ExH	Mean = 4.4897 StdDev = 1.7028	Mean = 4.3491 StdDev = 0.2372
ExMN	Mean = 0.0211 StdDev = 0.0102	Mean = 0.0431 StdDev = 0.1661
ExAL	Mean = 0.223 StdDev = 0.3084	Mean = 0.1823 StdDev = 0.1328
ExSAT_PC	Mean = 84.5311 StdDev = 4.0794	Mean = 84.9924 StdDev = 2.5256
ExBASE	Mean = 75.4069 StdDev = 216.6157	Mean = 51.5585 StdDev = 47.588
ExCaP	Mean = 55.6577 StdDev = 20.7032	Mean = 36.1725 StdDev = 18.7083
ExMgP	Mean = 33.0807 StdDev = 18.9606	Mean = 42.8389 StdDev = 13.9619
ExKP	Mean = 4.7639 StdDev = 4.0959	Mean = 4.0933 StdDev = 2.2107

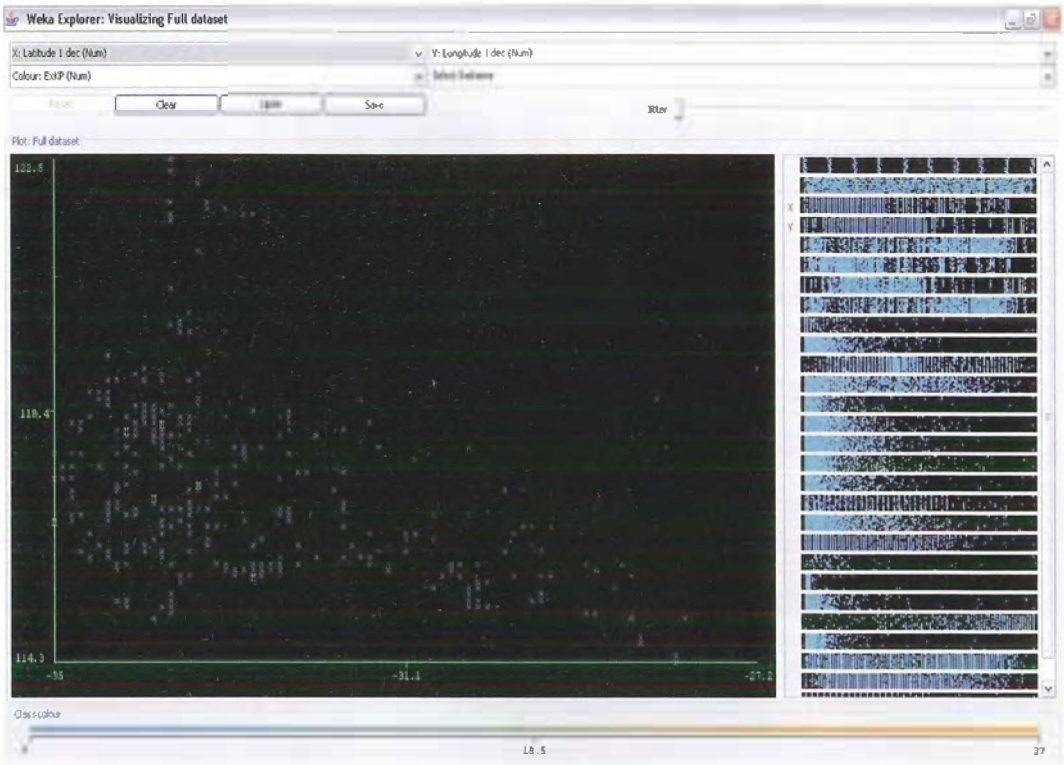
Trait name	6	7
Row Number	Mean = 1566.6261 StdDev = 481.0761	Mean = 871.176 StdDev = 250.0067
Latitude 1 dec	Mean = -32.8142 StdDev = 1.4286	Mean = -33.1501 StdDev = 1.1899
Longitude 1 dec	Mean = 117.3873 StdDev = 1.3427	Mean = 116.9388 StdDev = 0.9234
Elevation 0 dec	Mean = 270.2125 StdDev = 65.0335	Mean = 196.0337 StdDev = 118.8473
O_LATITUDE_GDA	Mean = -32.8198 StdDev = 1.4244	Mean = -33.1463 StdDev = 1.1882
O_LONGITUDE_GDA	Mean = 117.383 StdDev = 1.3403	Mean = 116.9395 StdDev = 0.9222
Elevation (M)	Mean = 270.1914 StdDev = 65.0514	Mean = 196.0411 StdDev = 118.8289
CACO3	Mean = 3.6118 StdDev = 1.9183	Mean = 4.1536 StdDev = 1.4535
OC	Mean = 0.8156 StdDev = 0.6348	Mean = 0.885 StdDev = 0.6918
PH	Mean = 5.042 StdDev = 0.5215	Mean = 5.2371 StdDev = 0.7614
clay	Mean = 3.2291 StdDev = 2.9257	Mean = 20.964 StdDev = 17.6558
EC	Mean = 4.4514 StdDev = 5.4187	Mean = 12.9183 StdDev = 15.7283
ExCA	Mean = 1.5385 StdDev = 1.2875	Mean = 1.3617 StdDev = 1.1414
ExMG	Mean = 0.6125 StdDev = 0.6223	Mean = 2.1078 StdDev = 1.7423
ExK	Mean = 0.0843 StdDev = 0.0933	Mean = 0.1609 StdDev = 0.1363
ExNA	Mean = 0.172 StdDev = 0.2726	Mean = 0.8298 StdDev = 1.1158
ExCEC	Mean = 8.9949 StdDev = 3.6309	Mean = 9.4205 StdDev = 3.1781
ExSUM	Mean = 2.4076 StdDev = 1.8759	Mean = 4.4633 StdDev = 3.1671
ExESP	Mean = 6.1905 StdDev = 5.1367	Mean = 14.886 StdDev = 11.2322
ExH	Mean = 4.3325 StdDev = 0.4844	Mean = 4.3561 StdDev = 0.234
ExMN	Mean = 0.018 StdDev = 0.0074	Mean = 0.0214 StdDev = 0.0084
ExAL	Mean = 0.1289 StdDev = 0.0722	Mean = 0.1701 StdDev = 0.1007
ExSAT_PC	Mean = 84.8604 StdDev = 7.7694	Mean = 80.1273 StdDev = 15.2846
ExBASE	Mean = 140.1633 StdDev = 226.4264	Mean = 41.2066 StdDev = 41.3964
ExCaP	Mean = 65.0337 StdDev = 17.7227	Mean = 35.1337 StdDev = 19.8522
ExMgP	Mean = 24.8739 StdDev = 15.3936	Mean = 45.2888 StdDev = 18.5868
ExKP	Mean = 3.9282 StdDev = 3.0879	Mean = 4.6641 StdDev = 4.2391



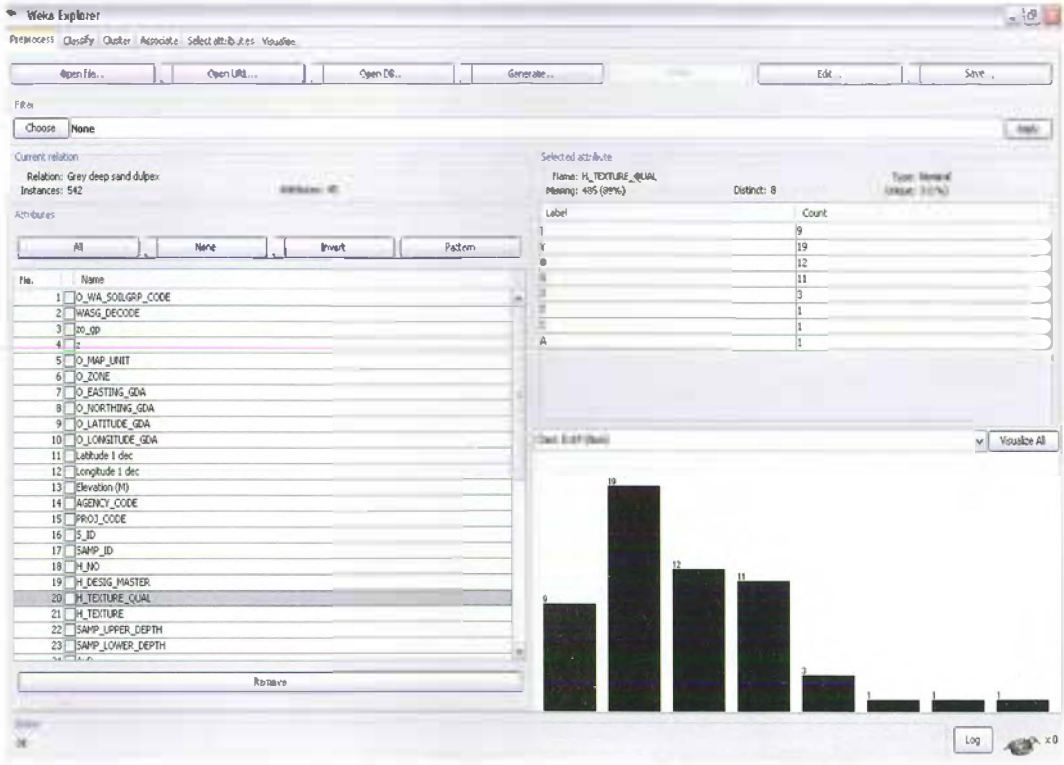
All soils (Initial analysis by WEKA)



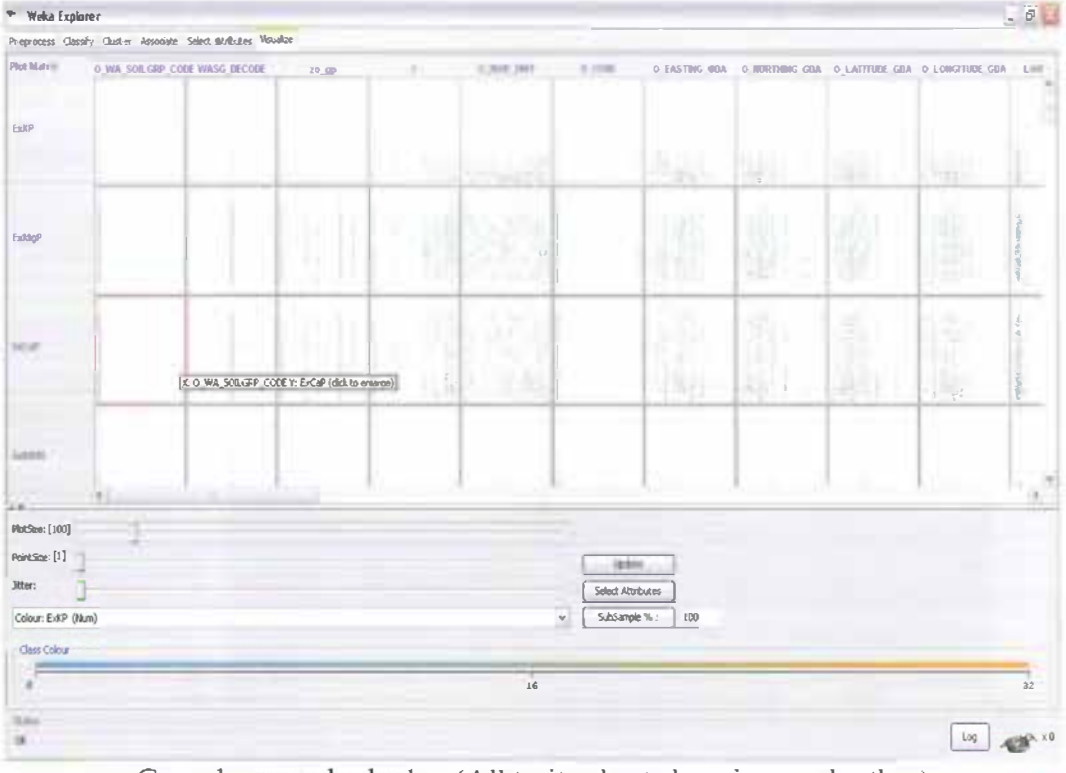
All soils



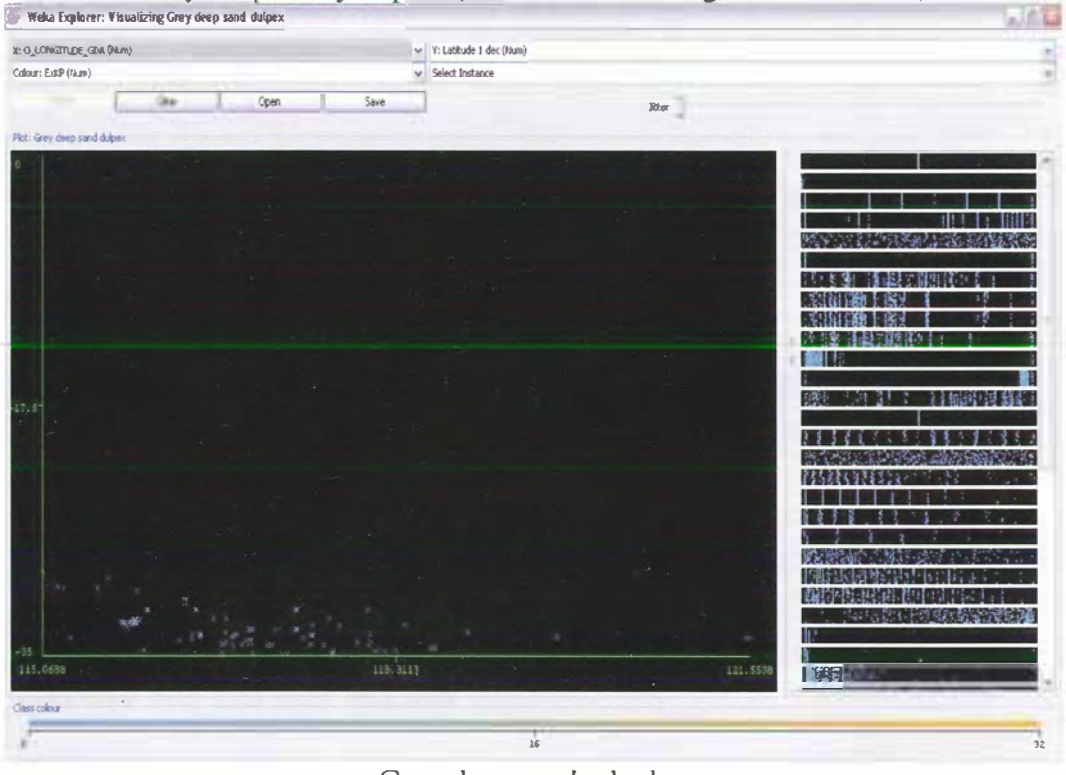
All soils



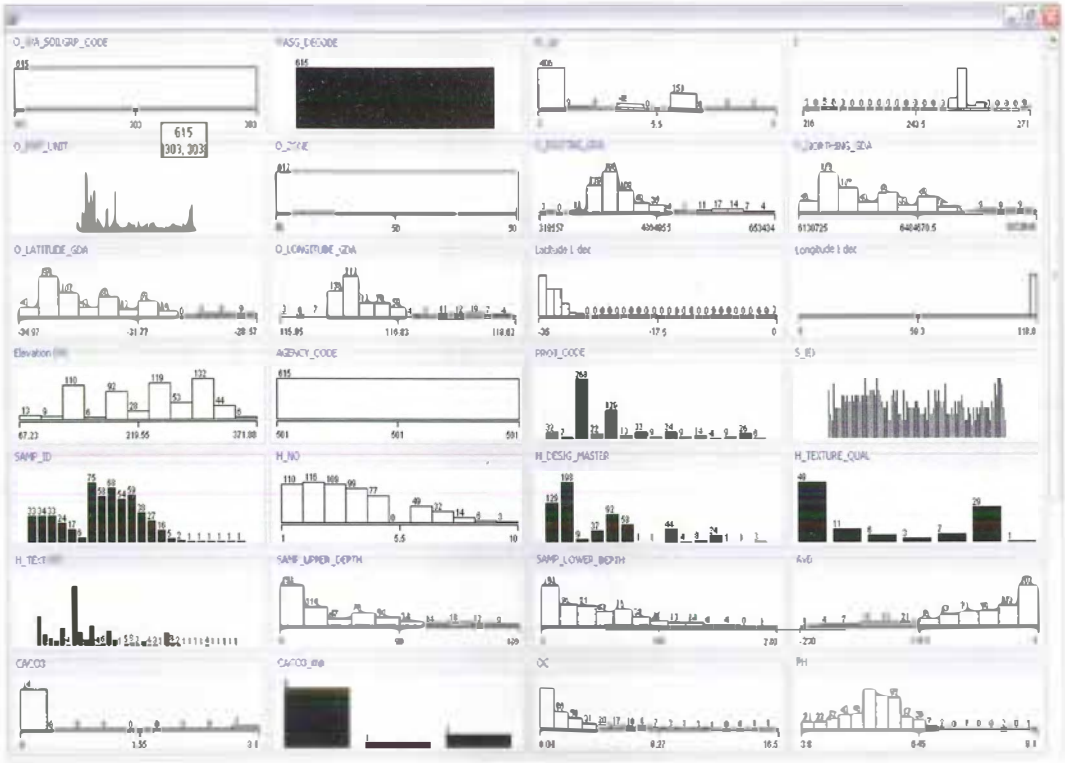
Grey deep sandy duplex (Initial analysis by WEKA)



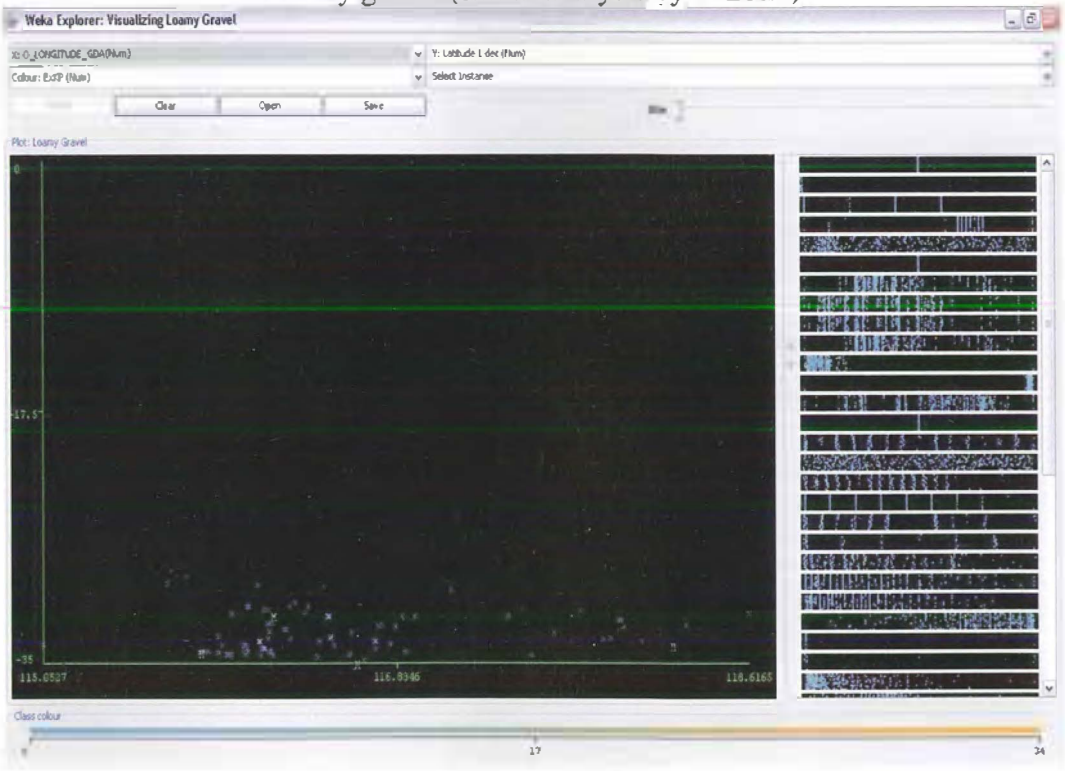
Grey deep sandy duplex (All traits charted against each other)



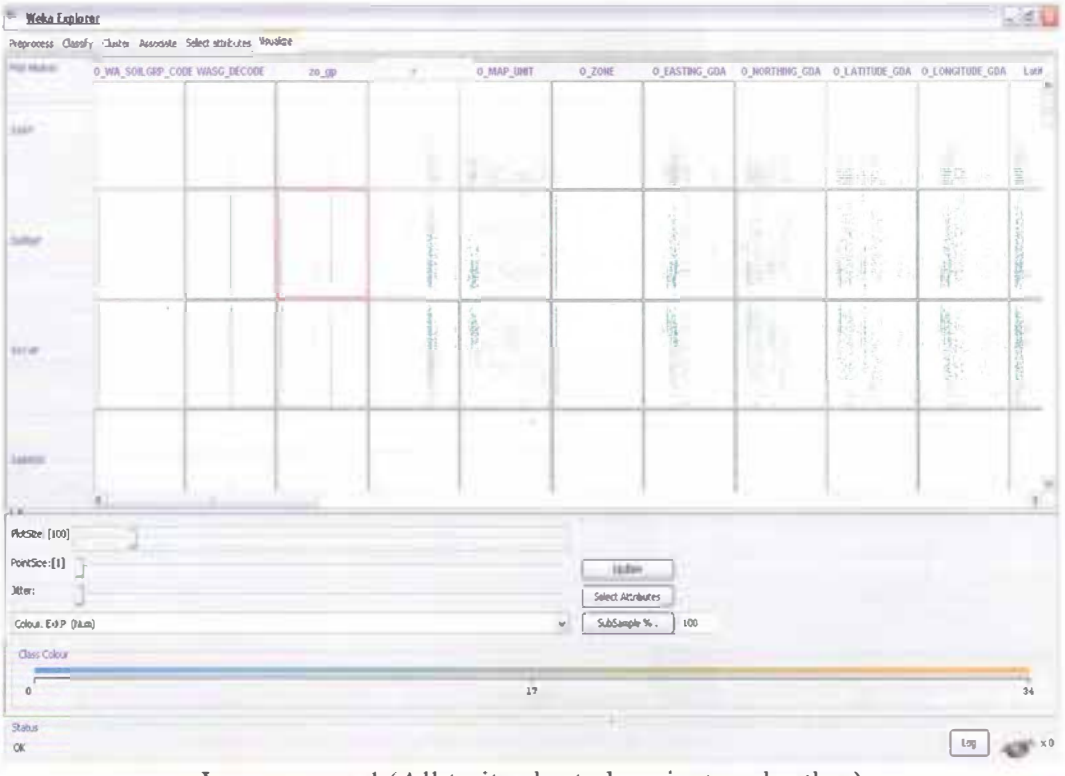
Grey deep sandy duplex



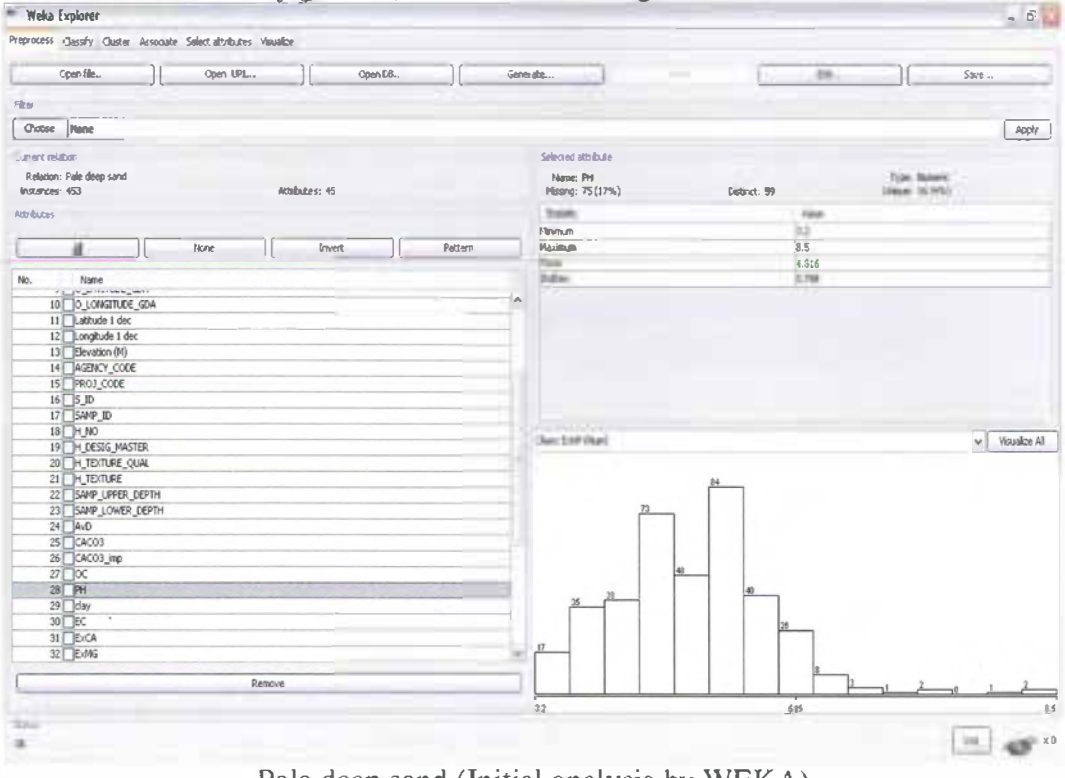
Loamy gravel (Initial analysis by WEKA)



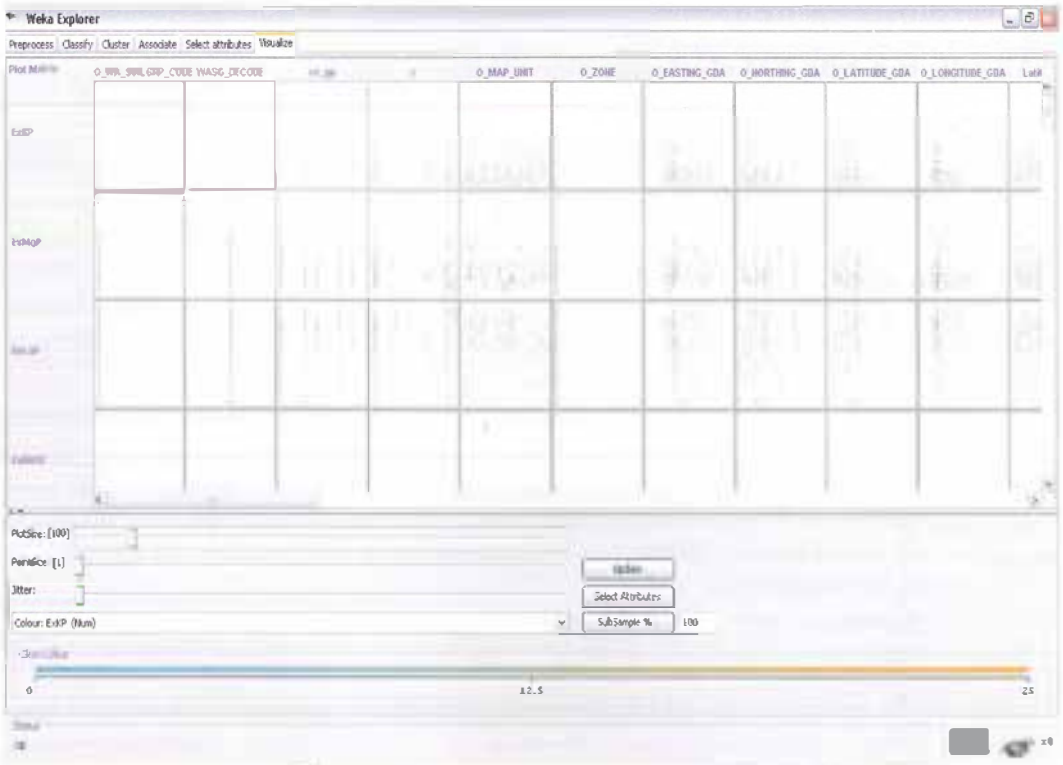
Loamy gravel



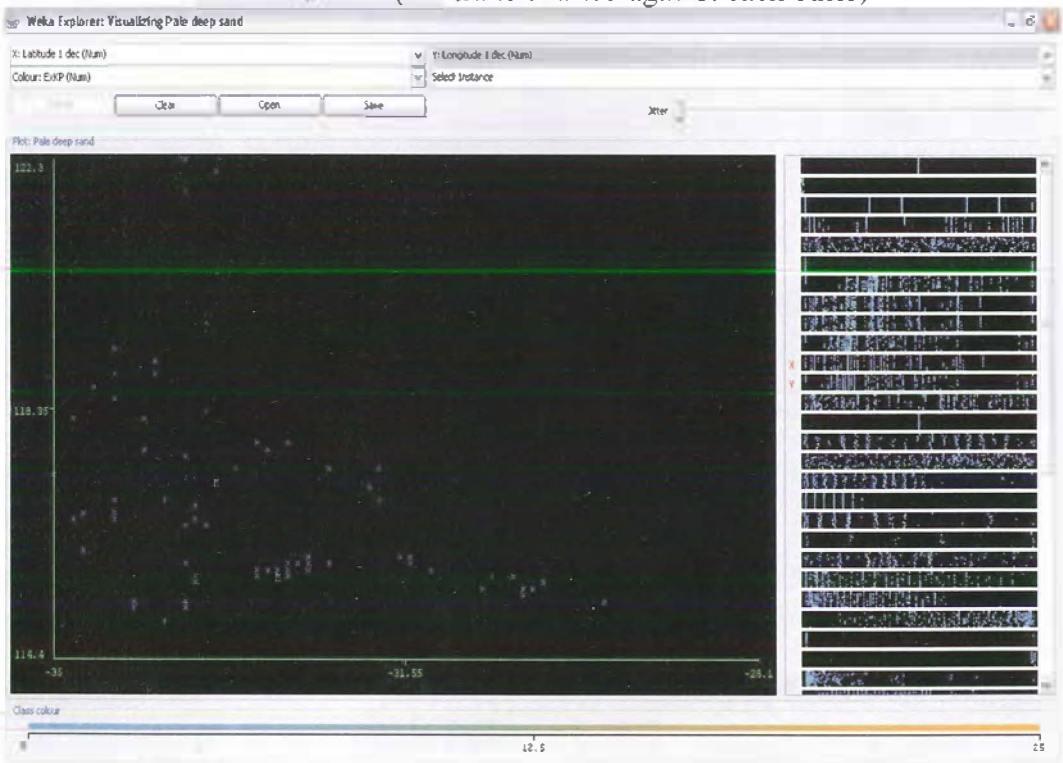
Loamy gravel (All traits charted against each other)



Pale deep sand (Initial analysis by WEKA)



Pale deep sand (All traits charted against each other)



Pale deep sand (Site location in the South Western Agriculture region).

8.4.2 Weka data output

All soils run with EM clustering algorithm.

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: Full dataset
Instances: 2825
Attributes: 28
WASG_DECODE
Row Number
Latitude 1 dec
Longitude 1 dec
Elevation 0 dec
O_LATITUDE_GDA
O_LONGITUDE_GDA
Elevation (M)
CACO3
OC
PH
Clay
EC
ExCA
ExMG
ExK
ExNA
ExCEC
ExSUM
ExESP
ExH
ExMN
ExAL
ExSAT_PC
ExBASE
ExCaP
ExMgP
ExKP
Test mode: evaluate on training data

=== Model and evaluation on training set ===

EM
==

Number of clusters selected by cross validation: 7

Cluster: 0 Prior probability: 0.0703

Attribute: WASG_DECODE
Discrete Estimator. Counts = 1 30.25 102.38 15.74 19.81 1 1 1 1 1 (Total = 174.18)
Attribute: Row Number
Normal Distribution. Mean = 1565.9922 StdDev = 452.2445
Attribute: Latitude 1 dec
Normal Distribution. Mean = -32.8869 StdDev = 0.7736
Attribute: Longitude 1 dec
Normal Distribution. Mean = 115.8674 StdDev = 0.3188
Attribute: Elevation 0 dec
Normal Distribution. Mean = 27.9008 StdDev = 21.8563
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -32.892 StdDev = 0.7686
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 115.8717 StdDev = 0.3188
Attribute: Elevation (M)
Normal Distribution. Mean = 27.9383 StdDev = 21.8729
Attribute: CACO3
Normal Distribution. Mean = 4.6424 StdDev = 0.3757
Attribute: OC
Normal Distribution. Mean = 1.4005 StdDev = 0.9329
Attribute: PH
Normal Distribution. Mean = 4.42 StdDev = 0.6196
Attribute: clay
Normal Distribution. Mean = 5.4024 StdDev = 4.627
Attribute: EC
Normal Distribution. Mean = 6.1757 StdDev = 8.182
Attribute: ExCA
Normal Distribution. Mean = 1.1611 StdDev = 1.5895
Attribute: ExMG
Normal Distribution. Mean = 0.5098 StdDev = 0.6851
Attribute: ExK
Normal Distribution. Mean = 0.0641 StdDev = 0.063
Attribute: ExNA
Normal Distribution. Mean = 0.1682 StdDev = 0.2214
Attribute: ExCEC
Normal Distribution. Mean = 10.9053 StdDev = 0.0383
Attribute: ExSUM
Normal Distribution. Mean = 1.9031 StdDev = 2.1627
Attribute: ExESP

Normal Distribution. Mean = 14.9238 StdDev = 14.1601
Attribute: ExH
Normal Distribution. Mean = 4.3294 StdDev = 0.3726
Attribute: ExMN
Normal Distribution. Mean = 0.0251 StdDev = 0.0162
Attribute: ExAL
Normal Distribution. Mean = 0.225 StdDev = 0.3166
Attribute: ExSAT_PC
Normal Distribution. Mean = 84.5868 StdDev = 0.3536
Attribute: ExBASE
Normal Distribution. Mean = 61.2089 StdDev = 86.5557
Attribute: ExCaP
Normal Distribution. Mean = 49.4834 StdDev = 23.8314
Attribute: ExMgP
Normal Distribution. Mean = 29.5674 StdDev = 16.4542
Attribute: ExKP
Normal Distribution. Mean = 6.0507 StdDev = 5.3628

Cluster: 1 Prior probability: 0.1061

Attribute: WASG_DECODE
Discrete Estimator. Counts = 1 1.03 72.41 172.2 1.06 1 1 5.14 7.14 1 (Total = 262.98)
Attribute: Row Number
Normal Distribution. Mean = 2190.0589 StdDev = 195.6033
Attribute: Latitude 1 dec
Normal Distribution. Mean = -30.6898 StdDev = 2.1722
Attribute: Longitude 1 dec
Normal Distribution. Mean = 116.1098 StdDev = 1.5522
Attribute: Elevation 0 dec
Normal Distribution. Mean = 218.0636 StdDev = 87.7154
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -30.6873 StdDev = 2.1751
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 116.1102 StdDev = 1.549
Attribute: Elevation (M)
Normal Distribution. Mean = 218.0736 StdDev = 87.7281
Attribute: CACO3
Normal Distribution. Mean = 4.4152 StdDev = 1.0216
Attribute: OC
Normal Distribution. Mean = 0.2866 StdDev = 0.3413
Attribute: PH
Normal Distribution. Mean = 5.1637 StdDev = 0.7109
Attribute: clay
Normal Distribution. Mean = 6.4155 StdDev = 5.013
Attribute: EC
Normal Distribution. Mean = 2.3069 StdDev = 2.482

Attribute: ExCA
Normal Distribution. Mean = 0.4027 StdDev = 0.3473
Attribute: ExMG
Normal Distribution. Mean = 0.2102 StdDev = 0.2107
Attribute: ExK
Normal Distribution. Mean = 0.0461 StdDev = 0.0466
Attribute: ExNA
Normal Distribution. Mean = 0.0632 StdDev = 0.1049
Attribute: ExCEC
Normal Distribution. Mean = 8.4567 StdDev = 4.2253
Attribute: ExSUM
Normal Distribution. Mean = 0.7223 StdDev = 0.526
Attribute: ExESP
Normal Distribution. Mean = 10.6726 StdDev = 12.9666
Attribute: ExH
Normal Distribution. Mean = 4.3165 StdDev = 0.4118
Attribute: ExMN
Normal Distribution. Mean = 0.0147 StdDev = 0.0065
Attribute: ExAL
Normal Distribution. Mean = 0.132 StdDev = 0.1604
Attribute: ExSAT_PC
Normal Distribution. Mean = 78.6698 StdDev = 16.1026
Attribute: ExBASE
Normal Distribution. Mean = 17.711 StdDev = 16.9294
Attribute: ExCaP
Normal Distribution. Mean = 54.0298 StdDev = 19.6102
Attribute: ExMgP
Normal Distribution. Mean = 27.8871 StdDev = 14.1181
Attribute: ExKP
Normal Distribution. Mean = 7.4848 StdDev = 5.4117

Cluster: 2 Prior probability: 0.2389

Attribute: WASG_DECODE
Discrete Estimator. Counts = 141.35 132.29 113.86 71.19 160.54 24.04 24 12 25.26 4 (Total = 708.54)
Attribute: Row Number
Normal Distribution. Mean = 1413.1403 StdDev = 694.2764
Attribute: Latitude 1 dec
Normal Distribution. Mean = -33.0378 StdDev = 1.4416
Attribute: Longitude 1 dec
Normal Distribution. Mean = 116.8079 StdDev = 1.0147
Attribute: Elevation 0 dec
Normal Distribution. Mean = 194.4666 StdDev = 111.2184
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -33.044 StdDev = 1.4373

Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 116.808 StdDev = 1.014
Attribute: Elevation (M)
Normal Distribution. Mean = 194.4564 StdDev = 111.2061
Attribute: CACO3
Normal Distribution. Mean = 4.6731 StdDev = 4.1143
Attribute: OC
Normal Distribution. Mean = 1.0274 StdDev = 0.4033
Attribute: PH
Normal Distribution. Mean = 5.4766 StdDev = 0.5903
Attribute: clay
Normal Distribution. Mean = 21.9738 StdDev = 19.5629
Attribute: EC
Normal Distribution. Mean = 21.2298 StdDev = 27.2513
Attribute: ExCA
Normal Distribution. Mean = 2.3968 StdDev = 0.5776
Attribute: ExMG
Normal Distribution. Mean = 2.2803 StdDev = 0.4203
Attribute: ExK
Normal Distribution. Mean = 0.2973 StdDev = 0.074
Attribute: ExNA
Normal Distribution. Mean = 1.0364 StdDev = 0.2246
Attribute: ExCEC
Normal Distribution. Mean = 10.8948 StdDev = 0.2989
Attribute: ExSUM
Normal Distribution. Mean = 6.0072 StdDev = 0.7892
Attribute: ExESP
Normal Distribution. Mean = 12.7097 StdDev = 3.0212
Attribute: ExH
Normal Distribution. Mean = 4.3756 StdDev = 0
Attribute: ExMN
Normal Distribution. Mean = 0.0306 StdDev = 0.1055
Attribute: ExAL
Normal Distribution. Mean = 0.1754 StdDev = 0.0332
Attribute: ExSAT_PC
Normal Distribution. Mean = 84.6159 StdDev = 0.5825
Attribute: ExBASE
Normal Distribution. Mean = 63.426 StdDev = 15.0703
Attribute: ExCaP
Normal Distribution. Mean = 46.0131 StdDev = 6.6722
Attribute: ExMgP
Normal Distribution. Mean = 36.1884 StdDev = 5.7596
Attribute: ExKP
Normal Distribution. Mean = 5.0828 StdDev = 1.1085

Cluster: 3 Prior probability: 0.1293

Attribute: WASG_DECODE
Discrete Estimator. Counts = 23.64 123.6 126.29 57.49 47.5 1 1.02 6.39 38.39 19.76 (Total = 445.08)
Attribute: Row Number
Normal Distribution. Mean = 1551.5659 StdDev = 577.3315
Attribute: Latitude 1 dec
Normal Distribution. Mean = -33.0444 StdDev = 1.4075
Attribute: Longitude 1 dec
Normal Distribution. Mean = 117.6117 StdDev = 1.6092
Attribute: Elevation 0 dec
Normal Distribution. Mean = 238.2696 StdDev = 90.8335
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -33.047 StdDev = 1.408
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 117.6099 StdDev = 1.6065
Attribute: Elevation (M)
Normal Distribution. Mean = 238.2584 StdDev = 90.8568
Attribute: CACO3
Normal Distribution. Mean = 3.497 StdDev = 1.9958
Attribute: OC
Normal Distribution. Mean = 0.817 StdDev = 0.7258
Attribute: PH
Normal Distribution. Mean = 5.0108 StdDev = 0.5998
Attribute: clay
Normal Distribution. Mean = 3.3867 StdDev = 3.0851
Attribute: EC
Normal Distribution. Mean = 4.0642 StdDev = 4.6493
Attribute: ExCA
Normal Distribution. Mean = 1.3359 StdDev = 1.1836
Attribute: ExMG
Normal Distribution. Mean = 0.4726 StdDev = 0.4281
Attribute: ExK
Normal Distribution. Mean = 0.0788 StdDev = 0.0896
Attribute: ExNA
Normal Distribution. Mean = 0.105 StdDev = 0.1136
Attribute: ExCEC
Normal Distribution. Mean = 8.6886 StdDev = 3.7777
Attribute: ExSUM
Normal Distribution. Mean = 1.9924 StdDev = 1.5711
Attribute: ExESP
Normal Distribution. Mean = 6.3955 StdDev = 5.5722
Attribute: ExH
Normal Distribution. Mean = 4.3756 StdDev = 0.81
Attribute: ExMN
Normal Distribution. Mean = 0.0176 StdDev = 0.0078

Attribute: ExAL
Normal Distribution. Mean = 0.136 StdDev = 0.0913
Attribute: ExSAT_PC
Normal Distribution. Mean = 83.179 StdDev = 10.6708
Attribute: ExBASE
Normal Distribution. Mean = 115.0361 StdDev = 176.103
Attribute: ExCaP
Normal Distribution. Mean = 64.6085 StdDev = 16.2468
Attribute: ExMgP
Normal Distribution. Mean = 24.3946 StdDev = 13.7052
Attribute: ExKP
Normal Distribution. Mean = 4.5922 StdDev = 3.6083

Cluster: 4 Prior probability: 0.1515

Attribute: WASG_DECODE
Discrete Estimator. Counts = 388.3 59.25 21.12 1.65 28.49 1.91 1 2.86 2.06 1 (Total = 507.63)
Attribute: Row Number
Normal Distribution. Mean = 477.8362 StdDev = 483.252
Attribute: Latitude 1 dec
Normal Distribution. Mean = -33.1554 StdDev = 1.3364
Attribute: Longitude 1 dec
Normal Distribution. Mean = 116.2295 StdDev = 0.4123
Attribute: Elevation 0 dec
Normal Distribution. Mean = 207.469 StdDev = 83.843
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -33.1598 StdDev = 1.3368
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 116.2228 StdDev = 0.4183
Attribute: Elevation (M)
Normal Distribution. Mean = 207.4728 StdDev = 83.8767
Attribute: CACO3
Normal Distribution. Mean = 4.6731 StdDev = 4.1143
Attribute: OC
Normal Distribution. Mean = 2.0281 StdDev = 2.34
Attribute: PH
Normal Distribution. Mean = 5.3068 StdDev = 0.6498
Attribute: clay
Normal Distribution. Mean = 19.6771 StdDev = 15.1592
Attribute: EC
Normal Distribution. Mean = 4.0851 StdDev = 4.4176
Attribute: ExCA
Normal Distribution. Mean = 2.8405 StdDev = 3.5629
Attribute: ExMG
Normal Distribution. Mean = 1.416 StdDev = 1.6533

Attribute: ExK
Normal Distribution. Mean = 0.1673 StdDev = 0.191
Attribute: ExNA
Normal Distribution. Mean = 0.2026 StdDev = 0.3294
Attribute: ExCEC
Normal Distribution. Mean = 10.9055 StdDev = 0.202
Attribute: ExSUM
Normal Distribution. Mean = 4.6257 StdDev = 4.6466
Attribute: ExESP
Normal Distribution. Mean = 5.3051 StdDev = 5.1812
Attribute: ExH
Normal Distribution. Mean = 4.4684 StdDev = 1.861
Attribute: ExMN
Normal Distribution. Mean = 0.0207 StdDev = 0.0117
Attribute: ExAL
Normal Distribution. Mean = 0.2226 StdDev = 0.3212
Attribute: ExSAT_PC
Normal Distribution. Mean = 84.5689 StdDev = 1.1706
Attribute: ExBASE
Normal Distribution. Mean = 83.9302 StdDev = 254.4062
Attribute: ExCaP
Normal Distribution. Mean = 56.3048 StdDev = 23.2493
Attribute: ExMgP
Normal Distribution. Mean = 34.015 StdDev = 21.8793
Attribute: ExKP
Normal Distribution. Mean = 4.3883 StdDev = 4.3627

Cluster: 5 Prior probability: 0.1382

Attribute: WASG_DECODE
Discrete Estimator. Counts = 72.92 116.43 18.06 7.48 48.77 1.04 1.01 19.58 3.15 2.86 (Total = 291.29)
Attribute: Row Number
Normal Distribution. Mean = 1106.8823 StdDev = 602.9314
Attribute: Latitude 1 dec
Normal Distribution. Mean = -33.238 StdDev = 1.1143
Attribute: Longitude 1 dec
Normal Distribution. Mean = 117.0566 StdDev = 0.8534
Attribute: Elevation 0 dec
Normal Distribution. Mean = 240.1944 StdDev = 98.6727
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -33.2354 StdDev = 1.1172
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 117.0539 StdDev = 0.85
Attribute: Elevation (M)
Normal Distribution. Mean = 240.2216 StdDev = 98.683

Attribute: CACO3
Normal Distribution. Mean = 4.2056 StdDev = 1.386
Attribute: OC
Normal Distribution. Mean = 0.7523 StdDev = 0.7708
Attribute: PH
Normal Distribution. Mean = 5.4114 StdDev = 0.7685
Attribute: clay
Normal Distribution. Mean = 28.9664 StdDev = 16.4267
Attribute: EC
Normal Distribution. Mean = 11.5177 StdDev = 12.5901
Attribute: ExCA
Normal Distribution. Mean = 0.8922 StdDev = 0.8803
Attribute: ExMG
Normal Distribution. Mean = 2.4375 StdDev = 1.7555
Attribute: ExK
Normal Distribution. Mean = 0.1054 StdDev = 0.1092
Attribute: ExNA
Normal Distribution. Mean = 0.7061 StdDev = 0.781
Attribute: ExCEC
Normal Distribution. Mean = 7.708 StdDev = 3.5656
Attribute: ExSUM
Normal Distribution. Mean = 4.1412 StdDev = 2.6732
Attribute: ExESP
Normal Distribution. Mean = 16.2854 StdDev = 12.5127
Attribute: ExH
Normal Distribution. Mean = 4.3497 StdDev = 0.233
Attribute: ExMN
Normal Distribution. Mean = 0.02 StdDev = 0.0053
Attribute: ExAL
Normal Distribution. Mean = 0.1589 StdDev = 0.0649
Attribute: ExSAT_PC
Normal Distribution. Mean = 77.7705 StdDev = 19.8733
Attribute: ExBASE
Normal Distribution. Mean = 19.2 StdDev = 18.7999
Attribute: ExCaP
Normal Distribution. Mean = 24.0435 StdDev = 17.8521
Attribute: ExMgP
Normal Distribution. Mean = 56.031 StdDev = 18.3099
Attribute: ExKP
Normal Distribution. Mean = 3.5729 StdDev = 4.3042

Cluster: 6 Prior probability: 0.1657

Attribute: WASG_DECODE
Discrete Estimator. Counts = 20.8 76.14 5.88 7.25 58.82 66.01 203.97 11.03 1 54.39 (Total = 505.29)

Attribute: Row Number
Normal Distribution. Mean = 1971.6654 StdDev = 835.3439
Attribute: Latitude 1 dec
Normal Distribution. Mean = -32.8494 StdDev = 1.3774
Attribute: Longitude 1 dec
Normal Distribution. Mean = 118.4897 StdDev = 1.7776
Attribute: Elevation 0 dec
Normal Distribution. Mean = 237.5186 StdDev = 94.1927
Attribute: O_LATITUDE_GDA
Normal Distribution. Mean = -32.8479 StdDev = 1.3827
Attribute: O_LONGITUDE_GDA
Normal Distribution. Mean = 118.4861 StdDev = 1.7752
Attribute: Elevation (M)
Normal Distribution. Mean = 237.4492 StdDev = 94.2222
Attribute: CACO3
Normal Distribution. Mean = 6.1137 StdDev = 9.3753
Attribute: OC
Normal Distribution. Mean = 0.8068 StdDev = 1.1615
Attribute: PH
Normal Distribution. Mean = 7.2326 StdDev = 1.4072
Attribute: clay
Normal Distribution. Mean = 34.8203 StdDev = 14.6525
Attribute: EC
Normal Distribution. Mean = 84.6138 StdDev = 109.4385
Attribute: ExCA
Normal Distribution. Mean = 5.3452 StdDev = 5.8022
Attribute: ExMG
Normal Distribution. Mean = 6.1859 StdDev = 3.5874
Attribute: ExK
Normal Distribution. Mean = 0.9941 StdDev = 0.9241
Attribute: ExNA
Normal Distribution. Mean = 3.912 StdDev = 3.7489
Attribute: ExCEC
Normal Distribution. Mean = 15.9383 StdDev = 7.1238
Attribute: ExSUM
Normal Distribution. Mean = 16.4309 StdDev = 8.5632
Attribute: ExESP
Normal Distribution. Mean = 24.0906 StdDev = 16.9232
Attribute: ExH
Normal Distribution. Mean = 4.3427 StdDev = 0.2914
Attribute: ExMN
Normal Distribution. Mean = 0.0235 StdDev = 0.0202
Attribute: ExAL
Normal Distribution. Mean = 0.2033 StdDev = 0.1855
Attribute: ExSAT_PC
Normal Distribution. Mean = 92.7342 StdDev = 7.9436

Attribute: ExBASE
Normal Distribution. Mean = 66.4449 StdDev = 66.5438
Attribute: ExCaP
Normal Distribution. Mean = 31.0246 StdDev = 23.083
Attribute: ExMgP
Normal Distribution. Mean = 38.575 StdDev = 13.1521
Attribute: ExKP
Normal Distribution. Mean = 6.2839 StdDev = 5.3136
Clustered Instances

0	200 (7%)
1	303 (11%)
2	675 (24%)
3	356 (13%)
4	425 (15%)
5	402 (14%)
6	464 (16%)

Log likelihood: -62.91778

All soils run with FarthestFirst clustering algorithm

=== Run information ===

Scheme: weka.clusterers.FarthestFirst -N 2 -S 1
Relation: Full dataset
Instances: 2825
Attributes: 28
WASG_DECODE
Row Number
Latitude 1 dec
Longitude 1 dec
Elevation 0 dec
O_LATITUDE_GDA
O_LONGITUDE_GDA
Elevation (M)
CACO3
OC
PH
clay
EC
ExCA
ExMG
ExK
ExNA
ExCEC
ExSUM
ExESP
ExH
ExMN
ExAL
ExSAT_PC
ExBASE
ExCaP
ExMgP
ExKP
Test mode: evaluate on training data

=== Model and evaluation on training set ===

FarthestFirst
==========

Cluster centroids:

Cluster 0
Pale_deep_sand 1838.0 -33.0 115.8 20.0 -33.043008 115.798875 20.0 4.673083700440528 1.0605508565310449 3.6 4.0 2.0 0.72 0.11 0.05 0.35 10.906117021276598 1.23 28.0 4.375641025641024 0.01 0.11 84.59387483355526 30.0 59.0 9.0 4.0
Cluster 1
Calcareous_loamy_earth 2603.0 -33.3 117.9 343.0 -33.324642 117.940119 342.83 10.0 0.14 8.9 49.8 131.0 3.2 14.84 1.11 26.6 43.0 45.75 58.0 4.375641025641024 0.022810383747178475 0.1798803827751198 100.0 91.0 7.0 32.0 2.0

Clustered Instances

0 2561 (91%)
1 264 (9%)