University of HUDDERSFIELD

An Effective Approach for Mining Complex Spatial Dataset

Author: Grace Samson

ID: 1251405

Supervisor: Prof. Joan LU

Date: December 2012

School of Computing & Engineering

A Project report submitted in partial fulfilment of the Requirements for the degree of MSc Advanced Computer Science

Abstract

In this research work, we have presented an illustration of spatial ecological predictive modelling. We focused on the unique features that distinguish spatial data mining from classical data mining, and presented major accomplishments of spatial data mining research, especially regarding predictive modelling, spatial outlier detection, spatial co-location rule mining, and spatial clustering.

In a very detailed research based study, we thoroughly investigated methods of mining patterns of a **spatial data set** (which generally describes any kind of data where the location in space of object holds importance) and made predictions based on the outcome of our analyses. We based this research on the analysis of some spatial characteristic of certain objects (that exist in an ecosystem). We began with describing the spatial pattern of events or objects with respect to their attributes, in other words and most specifically, we looked at how to describe the spatial nature/characteristics of entities in an ecological environment with respect to their spatial and non-spatial attributes. Secondly, we were able to predict likelihood of an object with a range of variables **(using spatial analyst tools like – distance, interpolation, overlay, raster creation, reclass, multivariate analysis, maths, surface and conditional tools respectively)** on a sample dataset and then we verified the model performance on the rest of the data. These feats were basically achieved using **data visualization**–which is the visual interpretation of complex relationships in multidimensional data – and **statistical interpretations**.

Results: At the conclusive end of this project, we were able to build a ***prediction/suitability model*** for the prediction of plant species around a river area. The method illustrated in this research work suggests the use of mapping and statistical functions in the prediction of large spatial database. We have tried to achieve this by two (2) major stages – first stage is the ***spatial analysis***, while the second stage is the ***statistical analysis***. The advantages and shortcomings of this approach are discussed in the context of the need for further development of methodology and software

This work is particularly useful to researchers in the field of data mining as it contributes a whole lot of knowledge to different application areas of data mining especially spatial data mining. It can also be useful in teaching and likewise for other study purposes

Dedication

This project work is dedicated to the **almighty GOD** who makes all things possible for us in his own time and my family especially my little wonderful kids, who has been a backbone to my success.

Acknowledgment

I would like to express my sincere gratitude and thanks to the **ALMIGHTY GOD** whom by himself has divinely arranged my coming for this study. You are awesome oh Lord my God.

I would also love to stretch out my heart felt gratitude to the management and staff of **UNIVERSITY OF ABUJA - Nigeria** for their uttermost support both financially and otherwise, I am indeed very grateful and indebted to you.

To the management and staff of **Education Trust Fund (ETF - NIGERIA). It was you who financially made this journey possible and I am eternally grateful.**

I would like to acknowledge the entire **staff** of the school of Computing and Engineering, Dept. of Informatics, for all their support and effort in the course of my master's education at the University of Huddersfield. I specifically want to appreciate my examiner, the ever loving and kind hearted **Dr. David Wilson**, who always has ears for whatever I want to say and always give a positive response, thank you so much.

My most sincere gratitude goes to my supervisor, who began this great journey to a glorious research future for me, if I say I am grateful it would be an understatement, I am very, very grateful and I appreciate all your kind efforts and patience towards the successful completion of this research project.

To every one whom I came across in Huddersfield, our path has been destined to cross, therefore, I never regret meeting any one of you, else your purpose in my life will not be accomplished, it's you that I need to learn the lessons that I have to, thank you so much for being used to help me through.

Finally and most importantly, I do acknowledge you Heaven and Zenith my indescribable stars, you guys are really the stars that give light to my path, I love you so much my awesome kids and thanks for your 12 months patience God bless you. And to you my husband if it was not you, then I would not be here, but because it's you we have seen the end of this unfathomable journey together, thank you for your love and patience and especially for devoting your time to care for the kids, I love you so so.

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter One

INTRODUCTION

The main purpose of this research work is to develop a generalised model for spatial pattern mining, capable of analysing data from a complex spatial system and then producing information that would be useful in various disciplines where spatial data form the basis of general interest. As acknowledged by **Wilson (2002)**, complex spatial systems are defined as those systems described by many variables, with high levels of interdependence between elements, governed by non-linear processes and having significant spatial structures. One would have noticed that the major challenge in trying to build a general complex spatial system model would be; to be able to integrate the elements of these complex systems in a way that is optimally effective in any particular case. Spatial data mining organizes **by location** what is interesting as such, specific features of spatial data that preclude the use of general purpose data mining algorithms are: **rich data types** (e.g., extended spatial objects), **implicit spatial relationships among the variables**, **observations that are not independent** and **spatial autocorrelation** among the features.

As highlighted by **Shekhar et al. (2005),** the explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. This is what motivates our research interest. Although there are some general purpose data mining tools such as Clementine and Enterprise Miner which are designed to analyse large commercial databases, **(Shekhar et al, 2005)**, discovered that general purpose tools for spatial data mining (especially in the case of a complex spatial data) need also to be develop because extracting interesting and useful patterns from spatial data sets is more difficult than the patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. As a result, we seek to develop a predictive model that produces spatial output patterns for spatial data mining. **Krzysztof et al. (1996)** added that mining spatial patterns is particularly interesting because it helps the researcher to discover the existing relationships between spatial and non-spatial data in a large spatial dataset.

1.1 Background

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as, droughts, vegetation diseases, and earthquakes.

We seek to create an explicit spatial model for event prediction using basic spatial data mining algorithms and not any of the general purpose data mining algorithms. In essence we aim to look at modelling (predictive modelling/ knowledge management of complex spatial systems), querying and implementing a complex spatial database (using data structure and algorithms). Critically speaking, the presence of spatial auto-correlation and the fact that continuous data types are always present in spatial data makes it important to create methods, tools and algorithms to mine spatial patters in a complex spatial data set.

The main goal of data mining is all about extracting patterns from an organization's stored or warehoused data. These patterns can be used to gain handful information about some aspects of the organization's operations, and to predict outcomes for future situations as an aid to decision-making.

The basic principles of data mining can be applied to any form of database including**: relational**, **transactional**, **multi-dimensional**, **distributed**, **spatial**, **multi-media**, **data-stream**, **time-series**, **text**, and **web** data respectively.

There are basically three types of complex systems as noted by **Weaver (1948, 1958).** These include the simple, the organised complexity and the disorganised complexity systems respectively. For the purpose and scope of our research work we are going to be considering the organised complex system and then move further into the ***disorganised complex system and complex adaptive systems*** in further research works. Organised complex systems are described by many variables, and all variables have strong interdependencies. *Human beings*, *brains*, *economies*, *cities*, *ecosystems*, *and language all provide examples of organised complex systems. Organised systems are also characterised by the presence of nonlinearities (*consider figure 1)

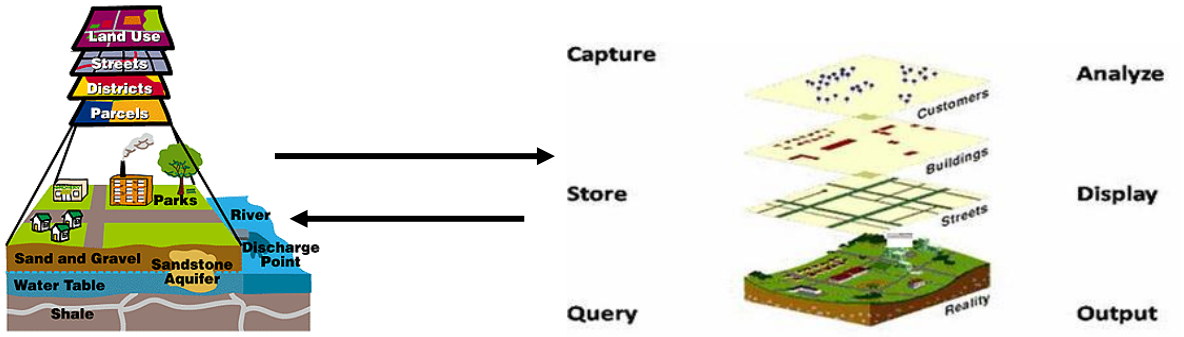


Figure 1: **A diagrammatic view of different spatial data layer and data from such a system is handled for knowledge management in a complex organised system**.

In other to be able to analyse a complex spatial system such as been mentioned above, the major challenge one would have is not having too much data but having too much and too complex a database for the discovery and understanding **structures**, **processes** and **relationships (Leung, 2010).** In addition, Complex spatial dataset forms the basis of many organizational/geographic data (as we have cited earlier) as such; extracting unknown and unexpected information from such spatial dataset of unprecedented large size and high dimensionality requires efficient and effective methods such as the application of knowledge discovery methods like data mining to spatial data. Contemporary data mining methodologies are not suitable for mining spatial data because those methods do not consider **location data** and also they do not support **implicit relationship between objects** which now makes it pertinent to employ appropriate spatial data mining techniques for efficient patterns discovery in a spatial dataset**. Pudi and Krishna (2009**) also noted that in any spatial data, location in space of objects holds importance and also, the bulk of data in the **real world** has some spatial component such as Medical images of the human body, engineering drawing, architectural drawing e.t.c, it therefore becomes pertinent that these spatial data should be dealt with in a specialised manner for pattern discovery.

1.1.1 Data Types

Pudi and Krishna (2009) observed that in trying to discover pattern in real world data, the different models in which real world data is organised and the pattern discovery technique to be applied to this models must be considered. Data types of a spatial set are the major element of a spatial database. A spatial database according to **Güting (1994)** supports spatial predicates (such as equal, disjoint, intersect, touch, overlap, cross, within, contains e.t.c) and offers spatial data

types in its data model and query language, and supports spatial data types in its implementation, providing at least spatial indexing and spatial join methods. Spatial data types, e.g. POINT, LINE, REGION, provide a fundamental concept for modelling the structure of geometric entities in **space** (The space of interest can be, for example, the two-dimensional abstraction of the surface of the earth – that is, geographic space, the most prominent example –, a man-made space like the layout of a VLSI design, a volume containing a model of the human brain, or another 3d-space representing the arrangement of chains of protein molecules) as well as their relationships. The types used may of course, depend on a class of applications to be supported. Without spatial data types a system does not offer adequate support in modelling **(Güting 1994).** In his view, **Schneider** in **Schneider (2002)** added that spatial data types provide a fundamental abstraction for modelling the geometric structure of objects in space; their relationships, properties and operations. He also added that there are more complex types such as partitions and graphs (networks). To a large degree he said, their definition is always responsible for successful design of spatial data models and the performance of spatial database systems and exercises a great influence on the expressive power of spatial query languages. **Sherkar et al. (2005)** focused on the unique features that distinguish spatial data mining from classical data mining and was able to classify them into the following four categories: **data input, statistical foundation, output patterns**, and **computational process.** Consideration of spatial predicates becomes very imminent when querying spatial databases; this is because spatial databases are sometimes faced with the presence of constraint (such as topological constraint) that makes current classical database solutions in appropriate **(Clementini et al, 1994).** Table 1 below gives a brief description of spatial data types.

Table 1: spatial data types, a brief description

Types	Examples	Description
Polygon/Areas	Locations of cities, object, factories, entities of study e.t.c	Specific locations that does not consider extent in any bearing
Lines	Roads (networks), streets, distance between any two points around an object, rivers e.t.c	Has a sequence of x, y coordinates, showing the distinct starting and ending points.
Points	Land-cover, areas covering administrative boundary around a study zone e.t.c	Linked lines bounded around an area (for xample you can measure perimeter e.t.c)

Spatial can also be represented as continuous surfaces (e.g. elevation, temperature, precipitation, pollution, noise e.tc) using the grid or raster data Model in which a mesh of square cells is laid over the landscape and the value of the variable defined for each cell.

1.1.2 Spatial Data – Features

Data Input

The data inputs of spatial Data Mining are more complex than the inputs of classical Data Mining because they include extended objects such as **points, lines**, and **polygons**. The data inputs of spatial Data Mining have two distinct types of attributes: non-spatial attribute and spatial attribute

Statistical Foundation

According to **Chang (2004),** spatial statistics arises when the data to be analysed are points in some Euclidean space, where the distance are usually represented by $\mathbf{R}^n$ in an n-dimension space. Also data that could be measured on some surfaces which consist of locations of points could also give rise to statistical analysis. Statistical analysis of spatial data can include finding of the likely variables present in a spatial data so as to be to create the parameters for an intended

model. Possible features like co-located objects, spatial outliers, spatial relationships and even spatial trends can be discovered using spatial statistics.

Output Patterns

According to **Shekhar et al. (2003**) some of the basic output patterns of a spatial process or spatial analysis based on literature, come in the forms listed below;

- **Predictive models** which basically arise form spatial classification
- **Spatial outliers** derived from spatial outlier detection
- **Spatial co-location/association** rules derived from colocation mining of spatial datasets and
- **Spatial clustering**

Computational process

Adhikary (1996) acknowledged that there are two major types of spatial operation that could be carried on a spatial dataset; **spatial join** and **map overlay**. Spatial join may be achieved using the R-Tree algorithm as suggested by **Brinkhoff et al. (1993)**, but map overlay involves the combination of the features and attributes of two or more data layers on a spatial map (data frames) layout produce a desired output. Other operations as identified by **(Güting 1994)** can be grouped into four (4) major classes depending on the nature of data input. These algebraic operations groups include spatial operations;

a. On a set of objects

Examples: **sum, closest**

b. Running numbers

Examples: **distance, perimeter, area**

c. Returning atomic spatial data types

Examples: **intersection, plus, minus, contour**

d. (Predicates) expressing relationships

Examples: **inside, intersect, meets, adjacent, encloses**

1.2 Knowledge/Pattern Discovery Task

The main purpose of spatial data mining is to search for interesting, valuable, and unexpected spatial patterns; which can be useful in so many application domains. Most often than not the pattern discovered always provide a new understanding of the real world, but it is very obvious here that this search must be a non-trivial one and should be as automated as possible with a large search space of **plausible hypothesis.**

Some of the pattern that one can discover in mining a spatial dataset would involve but not limited to those shown in table 2:

Table 2: knowledge discovery task in spatial data mining

Pattern	Description	Example
Location Prediction Predict	Trying to identify where a phenomenon will occur.	➢ predicting location of protein sub cellular (**Chou and Shen 2007)** ➢ Predicting location of a mobile cellular networks user (**Anagnostopoulos et. Al 2012)**
Spatial Interactions	The researcher is trying to find out which subsets of spatial phenomena interact?	➢ Application of spatial information to mobile computing **(Fröhlich et al, 2007)** ➢ Applying **spatial interactions** to the analysis of crime incidents **(Kakamu, 2008)**
Hot spot	Finding which locations are unusual or share commonalities through spatial clustering	➢ Detecting **spatial hot spots** in landscape ecology **(Nelson and Boots, 2008)** ➢ **Spatial** Organization of DNA in the Nucleus May Determine Positions of Recombination **Hot Spots (Razin and Laroaia, 2005)** ➢ Applying clustering techniques to crime hot-spot analysis **(Estivill-Castro and Lee, 2002)** ➢ Other application areas include earthquake analysis, vehicle crashes, agricultural situations …..

Spatial outliers' detection	Trying to identify abnormal patterns (outliers) from large data sets	➢ **Detecting Outliers** in Gamma Distribution **(Nooghabi et al. 2010)** ➢ Bearing Based Selection in Mobile Spatial Interaction (**Strachan and Murray-Smith, 2009**)

1.3 Motivation/Justification of study

The major motivating factor behind the modelling or mining of a spatial data lies in the differences that exit between spatial and non-spatial data. These differences as depicted in figure 2 below, explains the typical nature of spatial and non-spatial data as they affect their computation process, query language, and mining techniques.

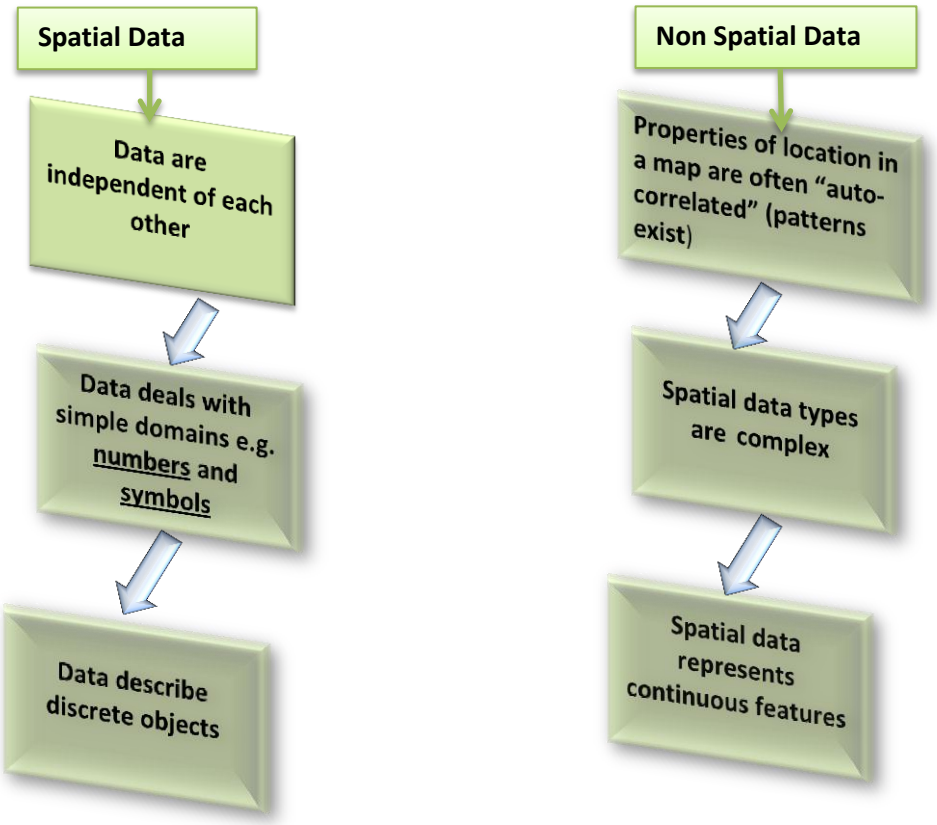


Figure 2: **difference between spatial and non-spatial data**

Some other reasons why classical data mining methods cannot be used for mining spatial data are listed below:

- ❖ The existence of spatial autocorrelation
- ❖ The fact that space is continuous
- ❖ Complex spatial data types
- ❖ Seeking regional knowledge
- ❖ Large dataset and many possible patterns
- ❖ Importance of maps as summaries

1.4 Research Objectives

Because complex spatial systems are those with significant spatial structures, we shall concern ourselves majorly with ***three main tasks***: these forms our major aims and objectives

We are interested in investigating, examining and analysing the different range of disciplines where complex spatial data has a significant function and then we shall link some of these functions to the development of complex spatial dataset. We would also investigate existing models for representing these kinds of systems and then try to develop a better, general, and analytical/predictive model for such complex systems, by implementing and querying a given complex spatial database. We would then take an example from one of the various existing models (e.g. ***ecosystems analysis***, ***accident analysis system***, ***transport control system*** e.t.c) which will form an important antecedent of the programme of building general models of complex systems within the field of complex spatial data analysis system.

Finally we shall make a conclusion based on what we have discovered from the proposed extensive study of complex spatial systems. However, the six (6) basic areas of interest of spatial data mining as listed below, would be a major term of reference to what we intend to achieve (and our major task would be to develop an algorithm for some of these tasks and then identify their application areas).

1. Predictive modelling/ Knowledge Management (for event prediction)
2. Spatial outlier detection
3. Spatial co-location rule/patterns mining
4. Spatial clustering.

5. Spatial trend

6. Spatial classification

1.4.1 What needs to be represented?

The main application driving research in spatial database systems are GIS. Hence we consider some modelling needs in this area which are typical also for other applications. Examples are given for two dimensional **space (length and breadth)**, but almost everywhere, extension to the three - or more-dimensional case is possible. There are two important alternative views of what needs to be represented:

(i) **Objects in space**: We are interested in distinct entities arranged in space each of which has its own geometric description.

(ii) **Space**: We wish to describe space itself, that is to say something about every point in space.

Point (i) allows us to model, for example, ***cities***, ***forests***, or ***rivers***, while (ii) is of thematic maps describing e.g. ***land use/cover*** or the partition of a ***country into districts***. Since ***raster images*** say something about every point in space, they are also closely related to the second point. We can reconcile both views to some extent by offering concepts for modelling in point (i) for instance**,** ***single objects***, and point (ii) can help us consider ***spatially related collections of objects***.

1.5 Research Methodology

Research methodology describes the objectives of your study by determining the type of research which is (***descriptive***, ***co-relational*** and ***experimental***) and then establishes the type of research design you need to adopt to achieve them. Because we are interested in finding out if an increase or decrease (or any form of change) in physical phenomena (***e.g climatic structure***) has an impact on the existence of objects in any given geographical space (***e.g ecosystem - especially on plant and animal species***), we have adopted the ***experimental method*** of research to achieve our objectives as you can see in figure 3 below.

The **experimental method** also known as the **cause** and **effect** method or the ***empirical*** research method is a data-based research method. Conclusions made at the end of the research are always capable of being verified with experiments or observations. It always involves two types of variable **dependent** and **independent variables** where the effect of substituting the value of one or more of the independent variable affects the outcome of the dependent variable based on a deliberate manipulation of one of them in other to learn its effects.

For example given an equation of the form **2x + v = y……..**

Where:

Y is the dependent variable

X is the independent variable

and V could be any constant of another independent variable

For every value of x and v, y takes a new value. This kind of research is appropriate when all the researcher seeks to establish a **proof** that an independent variable always affects the dependent variable (as we saw in the equation above). In other words, there is always a necessity to start with the reality in the first place, starting from the foundation, and then actively go about doing certain things to stimulate the production of desired information. All this can be achieved by the presence of a working hypothesis which will state the possible result.

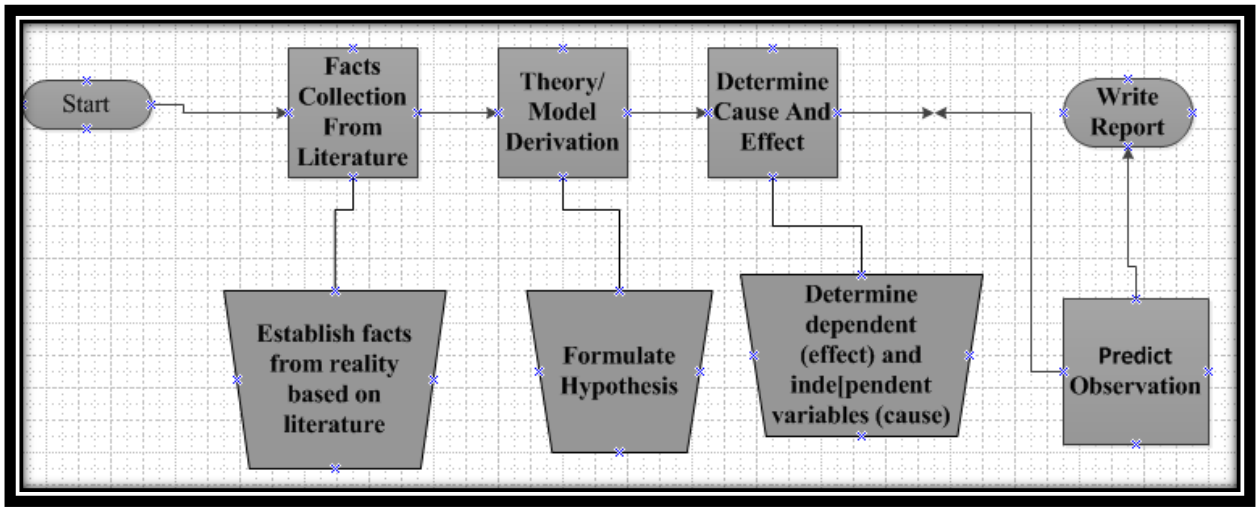


Figure 3: **our research methodology**

1.6 Scope and limitation of the Study

The scope of this study is limited to the peripheral computing of the complexity of spatial data mining; this is to give the common reader a clear understanding of the issues that surrounds the subject and field of spatial data mining. However, the study also entails a wide coverage of methods, techniques, theories, complex interactions among elements of a complex spatial system and examples of spatial data and spatial data mining. Thus, it will be useful both to the professional researcher and also for the amateur. The basic limitation of this study is the fact that the field of spatial data mining is very large and complex and as such time constraints may also limit the expected output of this research project (though we would give in all our best to achieve an effective system). So basically this project may not be able to look critically at the individual algorithms and techniques of spatial data mining in details but will mention each and every one of them with explanations, examples and typical applications.

1.7 Organization of work

1. **Chapter one** gives a detailed introduction of the project
2. **Chapter two** starts by conducting an intensive literature review on Data mining, Spatial data mining, Complex systems, Spatial patterns mining, Complex spatial system, Methods for modelling and querying a spatial database, Predictive modelling/ Knowledge Management and Creating mathematical models for computer simulation based on spatial data and representing spatial data using graphical features.
3. **Chapter three** undertakes a thorough analysis of existing models and algorithms for predictive modelling/ knowledge management of CSS.
4. **Chapter four** tries to work out a new and typical application model for modelling a complex spatial system and Design a system for visualization of event prediction result using Arcgis 10.1 software
5. **Chapter five** explains the analysis and design simulation system for complex spatial system prediction
6. **Chapter six** gives detailed generalisation and interpretation of the prediction system
7. **Chapter seven** explains all the design methods used
8. **Chapter eight** discusses relevant professional and ethical issues in system development
9. **Chapter nine** is an evaluation of the research
10. **Chapter ten** is the **Conclusion and future works.**

1.8 Choice of programming language

We have employed the functionalities of the **Arcgis 10.1** software in this project for spatial analysis. The software is a ***geographical information*** (GIS) based system that can help us to analyse spatial data in form of maps. In a range of application, spatial data consist of geographic information which can be mapped and analysed; as such GIS(s) help us extract geographic information from a spatial data set, represent such information using map and then analyse the information that has been mapped to produce a required output. This output could be the final information that we sought (in which case it is documented and shared) or it could be an input for further spatial analyses. **Arcgis 10.1** as a geographical information (GIS), has the functionalities that allows it to be able to manage geographic information system or a spatial database (which is always difficult using classical database management systems).

Chapter Two

Literature review: Theoretical Framework for Mining Complex Spatial Dataset

2.1 INTRODUCTION

Data mining (DM) deals with extracting interesting knowledge from real-world, large and complex data sets; it is the core step of a broader process, called knowledge discovery in databases (KDD)- (see figure 4 below). In addition to the DM process, which actually extracts knowledge from data, KDD process includes several pre-processing (***data preparation***) and post-processing ***(knowledge refinement***) phases **(Ghosh and Freitas, 2003).** From the work of **Leung (2010),** we discovered that the problem of extracting knowledge is not the issue of not having enough data, but having too much and too complex a database for the discovery and understanding of **structures**, **processes**, and **relationships** this like we may expect has left useful knowledge often hidden in the sea of data that awaits discovery.

Data mining bridges many technical areas, including ***databases***, ***statistics, machine learning***, and ***human-computer interaction***. The set of data mining processes used to extract and verify patterns in data is the core of the knowledge discovery process **(Hammawa and Sampson, 2011).** In their own view, **Smyth et al. (2001)** stated that data mining can be characterised as a secondary analysis tool which seeks to find unexpected and unforeseen information that could be hidden in a given data set. This means that for a large number of times, data miners and knowledge seekers are not typically involved directly with data collection process. A generic problem in data mining is to find relationship between variables; that is to say, is an action performed to a given data set, say **z** likely to affect the action performed to another data set **x**. The data set involved in a data mining process is known as a database which is a large record of data pertaining to a given discipline or field. According to **Frawley et al. (1992),** data mining or knowledge discovery in databases refers to the discovery of interesting, implicit, and previously unknown knowledge from a large database. **Dasarathy (2003)** noted that knowledge discovery is clearly one of the many potential objectives of information fusion process and then he described data mining as a means of accomplishing the objective of knowledge discovery. Contrary to

these conventional views, **Imielinski and Mannila (1996)** objected that "there is no such thing as discovery; it is all in the power of the **query language".** [1]

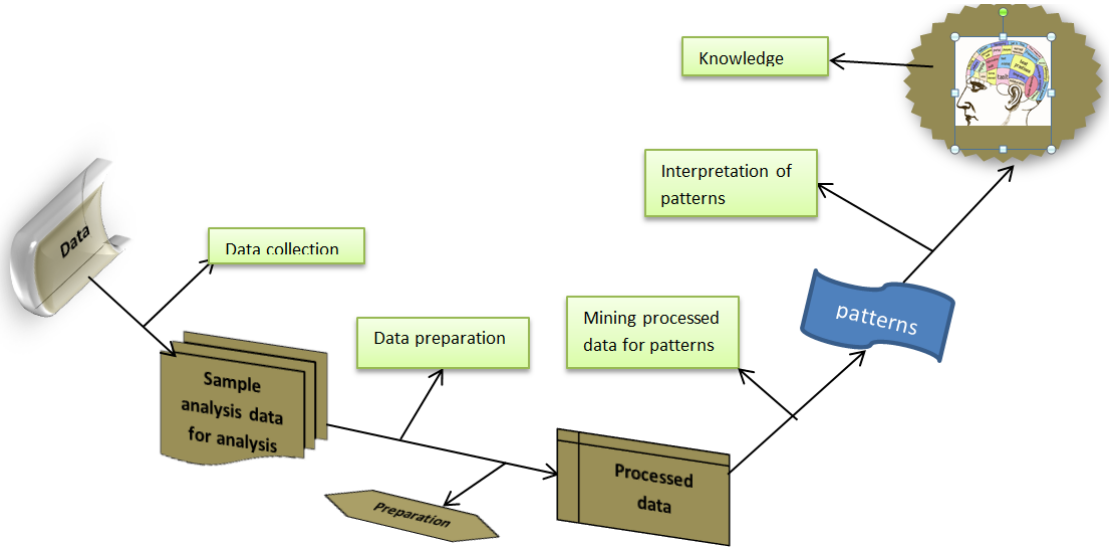


Figure 4: **The process of knowledge discovery in a database**

Another important contribution to the research of databases and data mining processes is found in **Fayyad et al. (996**); here they stated that Data mining can be viewed as the automated application of algorithms to detect patterns in data.[2] **Bhandari et al. (1997)** supported this idea by adding that the process of interpreting patterns represents knowledge discovery, and traditionally requires activity on the part of a domain expert. **Cios (2000)** stated that data mining is inherently associated with databases and as such data mining and knowledge discovery are tools that can help in dealing with the problem of the acute and widening gap between data collection and data comprehension; thus according to him, data mining methods are algorithms that are used on databases, after initial data preparation for model building or for finding patterns in a data set.[3] Some of these data mining algorithms would be examined in section 2.2.

[1] We could deduce from this proposition that an effective data mining process would involve a good database querying performance.

[2] We believe this is the basic framework for all branches of data mining
[3] This is the procedure we adopted in this project that has helped us to reach a conclusive end

2.2 Data Mining methods and algorithms

Data mining algorithms according to **Cios (2000)** are basically the building blocks of any database both for finding patterns and for building models. **In Smyth et al (2001)**, data mining algorithms could be seen as some well-defined procedures that take data as input and produce output in the form of models and patterns. **Wu et al. (2008)** investigated on the top ten algorithms that are among the most influential data mining algorithms in the data mining research community. Amongst these algorithms are ***AdaBoost, C4.5, k-Means, EM, PageRank, kNN, SVM***, ***Naive Bayes, A priori, and CART***. These algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all amongst the most important topics in data mining research and development.

C4.5 algorithm creates a decision tree (that can then be tested against unseen labeled test data to quantify how well it generalizes) based on a set of labeled input data, it is robust in the presence of noise, it construct classifiers by taking as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

Some other algorithms have been developed over the years as a product of intensive and continuous research in this all useful discipline of data mining. Based on the model-induction mode of data mining (which is a way of deducing a closed-form explanation based solely on observations, in other to infer models from data), **Babovic (2000)** developed a Special kind of evolutionary algorithm called genetic programming. Evolutionary algorithms according to him are engines simulating grossly simplified processes occurring in nature and implemented in artificial media—such as a computer. Explaining this approach of data mining, Babovic in **Babovic (2000)** gave an elucidation that in this process as the genetic programming, the evolutionary force is directed toward the creation of models that take a symbolic form and evolving entities are presented with a collection of data, and the evolutionary process is expected to result in a closed-form symbolic expression describing the data. Symbiotic Bid-Based Genetic Programming (SBB) has been described in **Doucette et al. (2012)** as an algorithm that employs cooperative and competitive co-evolution for discovering knowledge from large databases with many attributes. This method opposes to the filter or wrapper methodologies address both tasks simultaneously

2.3 Data Mining Tasks and Techniques

Data mining according to **Guo and Mennis (2009)** encompasses various tasks (which include – ***classification*** (supervised classification), ***association rule mining***, clustering (unsupervised classification) and multivariate geo-visualization.) and for each task a number of different methods are often available, which be ***computational***, ***statistical***, ***visual***, or some combination of them.[4]

The aim of data mining and knowledge discovery is to provide tools to facilitate the conversion of data into a number of forms, such as equations (or models) that provide a better understanding of the process generating or producing these data (Babovic, 2000).[5] According to **Raza (2012),** the two "high-level" primary goals of data mining, in practice, are ***prediction*** and ***description***.[6] However, the basic task of mining data for pattern discovery includes classification and clustering and the major difference between them is based on their specific requirements regarding the structure of their input data **(Pudi and Krishna, 2009)**. These techniques help to analyse the observations made from physical systems in other to search for the information that they encode, they are basically categorised into two namely: ***numerical*** and ***knowledge based*** techniques. The numerical technique is further classified into three i.e ***statistical*** (where all analyses are treated as hypothesis tests or exercises in parameter estimation as stated by **Hochachka et al. (2007)**), **heuristic** (i.e the process of extracting patterns from data sets which are then used to gain insight into relational aspects of the phenomena being studied and to predict outcomes to aid decision making according to **Flentje et al. (2007)**) and **deterministic** respectively. The knowledge based techniques has to do majorly with data mining approaches whereby key data sets are assessed to establish inter-relationships with the primary training set.[7] We shall look at the various DM (Data Mining) tasks and the suitable tools/technique used for this task below. Generally, in trying to mine information from a given data store, four types of interactions is basically being aimed at:[8]

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.

[4] We are going to be applying these three methods in carrying out the data mining tasks we have chosen in this project.
[5] This has given us the insight to what we are trying to achieve in this project
[6] This are the goals we are actually set to achieve
[7] We shall apply both techniques in these research work (i.e the numeric and knowledge based technique)
[8] We have already explored these in section 1.2 above

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. An example of associative mining could be trying to relate the sale of a particular good to be determined by the sale of another non similar good when there seems to be a kind of relationship between both.

Sequential patterns: Data is mined to anticipate behaviour 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.

Based on these interactions as mentioned above the following techniques for mining data in a large database has been established:

Classification -: Classification according **Agrawal et al. (1993)** is about grouping data items into classes (***categories***) according to their properties (***attribute values***). It requires a ***training set*** to train (or configure) the classification model, a ***validation set*** to validate (or optimize) the configuration, and a ***test set*** to evaluate the performance of the trained model. From **Nisbet et al. (2009)**, classification involves the implementation of most of the **tree based** data mining algorithms as a way of making decisions based on the solution to a previous problem of the same nature. It involves structuring a decision tree as a sequence of simple questions. Whereby the answers to the first given question determines the next question that is posed, if there is any. The result is a network of questions that forms a tree-like structure. The "ends" of the tree are terminal "leaf" nodes, beyond which there are no more questions. Classification otherwise known as supervised learning is used **Predictive Modelling that is being able to use observations to learn to predict.**

In ***classification***, a collection of records (training set) is made whereby each record contains a set of attributes, and one of the attributes is the class. The main task here is to find a model for class attribute as a function of the values of other attributes. It is always important to note that in classification, a test set is used to determine the accuracy of the model.

The given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. **Witten, I.H. (2008)** describes classification as a way of using a set of classified examples (***known as instances***) to produce a method of classifying new examples by using a set of attributes (***i.e a fixed set of features***) of the instance class. The major characteristics of creating classes (which could be discrete or continuous) ***are to be able to arrive***

at a model which describes how the decision of class group was made. Classification according to **Wu et al. (2008)** accurately predicts the class to which a new case belongs. In other words classification as a data mining (machine learning) technique used to predict group membership for data instances i.e it is a useful resource when prediction and forecasting future events/trends is of paramount importance. Some of the data mining techniques used for classification include:

Decision Tree based Methods

Rule-based Methods

Memory based reasoning

Neural Networks

Naïve Bayes and Bayesian Belief Networks

Support Vector Machines

CART

CHAID

Classification by back propagation

Clustering: - In data mining ***clustering*** (generally known as unsupervised classification) simply means the logical detecting and grouping of a set of similar subgroups among a large collection of cases and to assign those observations to the clusters **(Wu et al., 2008**). More practically, it will involve finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. For example, we group a set of ***related documents for browsing***, ***group genes*** and ***proteins*** that have similar functionality, or group stocks with similar price fluctuations. **Brendan and Delbert (2007)** described clustering as a data mining technology that divides data objects into more than one class or classes. The major characteristics of the clustering techniques according to **Ren and Yin (2010)** includes the fact that it takes as input a sample matrix,[9] which is to think a sample to be a point in the characteristic variable space. The output of the clustering algorithm is usually a cluster genealogy diagram to reflect all the classification. In clustering analysis, the Partitioning clustering methods organises a data item is assigned to the ‘‘closest” cluster based on a proximity or dissimilarity measure while the hierarchical clustering, on the other hand, organizes

[9] We shall be doing in the modelling section

data items into a hierarchy with a sequence of nested partitions or groupings **(Jain and Dubes, 1988**). However, clustering as a data mining ***technology is used when we need to find the number of clusters as well as the members of each***. The clusters are assigned a sequential number to identify them in results reports. Cases within a group should be much more similar to each other than to cases in other clusters. Data mining techniques used for clustering purposes include:

- *k*-Means clustering
- *EM (*Expectation-Maximization*)cluster analysis*

***Associations*: -** Data mining like we mentioned earlier seeks to find all forms of pattern which could be hidden in a database in other words data mining is a process to extract the implicit, not known in advance and potentially useful information and knowledge from a large number of incomplete, noisy, vague and random practical application data. It is a reliance on the application, and thus different applications may require different data mining techniques. Mining associations according to **Agrawal et al. (1993)** is intended to discover regularities between items in large transaction databases by finding all rules from transaction data satisfying the minimum support and the minimum confidence constraints. Association rule mining is one of most popular data analysis methods that can discover associations within data. It is used for creating **Link Analysis** that is, presenting links between individuals rather than characterising whole. An association rule is an expression denoted by (Association Rule – X → Y;

$X \subseteq I, Y \subseteq I$ and $X \bigcap Y = \phi$ as **stated by Agrawal and srikant (2003)**)**X ⇒ Y**, where **X** and **Y** are sets of items typically in association rule mining, support and confidence are used to measure the significance and certainty of a rule **(Peng, 2010)**. Association mining in data mining activities has been very imperative in mining the significant association rules between items in a trade database, which have reflected the behaviour mode of the customers. An association rule can also be applied to ***website's structure optimization, storage planning, network accident analysis, designing of business catalogue***, ***add to sales,*** etc. The rule is given as the following statement;

"*Let* ***j = j1, j2… jm (where*** *m* ***may range from 0 to*** *∞) be a set of items. Given a database* ***D*** *of t****ran****sactions, where each transaction* ***t*** *is represented as a set of items, with* ***t[j] = 1*** *if* ***t*** *bought the item* ***ji*** *and* ***t[j] = 0*** *otherwise. Let* ***x*** *be a set of some items in* ***j*** (for instance if ***j*** represents

bags ***x*** may be used to represent sizes of the bags or other bag attributes). *We say that a transaction **t** satisfies **x** if for all items **ji** in **x,** $t[j_x] = 1$ if an item x is bought and $t[j_x] = 0$ otherwise"* **(Furong et al, 2010)**.

Having mentioned earlier that *the goal of association rules mining is to detect and analyse relationships or associations between specific values of categorical* variables in large data sets, we wish to add that the technique can be used to analyse simple categorical variables, dichotomous variables, and/or multiple target variables. When **Agrawal et al. (1993)** first proposed the mining of association rule in a transaction database; they presented a case study in this form:

"*Suppose you are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. Present an efficient algorithm that generates all significant association rules between items in the database. The algorithm should incorporate buffer management and novel estimation and pruning techniques. Also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm"* **Agrawal et al. (1993)**

Solution idea for the above problem from **Agrawal and srikant (2003):**

"Association Rule – X → Y; $X \subseteq I, Y \subseteq I$ and $X \bigcap Y = \phi$

- Say ABCD and AB are large item-sets
- Compute conf = support(ABCD) / support(AB)
- If conf >= minconf

AB → CD holds.

Important note for confidence and support:

- Association rule X→Y has confidence c,
- c% of transactions in D that contain X also contain Y.
- Association rule X→Y has support s,
- s% of transactions in D contain X and Y" (**Agrawal and srikant 2003)**.
-

Existing algorithms used for mining association rules include:

- Fitting of function e.t.c
- AprioriTid Algorithm
- Knowledge Discovery
- Induction of Classification Rules
- Discovery of causal rules
- Apriori

2.3.1 Applications of data mining

Data mining applications have proved highly effective in addressing many important areas of human activities and endeavours; we expect to see the continued construction and deployment of KDD applications for crucial decision support systems. Exemplary applications employing data mining analytical techniques require the KDD technical community to keep improving the underlying techniques for model building and model understanding **(Chidanand et al., 2000)**. For instance in **Chidanand et al. (2002)**; **Soares et al (2008**); **Kohavi and Provost (2001)**, data mining techniques has been applied to business management especially in the area of electronic commerce and e-business transaction generally. **Raza in Raza (2012)** explored the application of data mining in ***bioinformatics*** and the application of data mining techniques in ***pharmacovigilance*** was examined in **Thabane and Holbrook (2004)**. A particular active area of research in ***human health***, ***psychology*** and ***well-being*** is the application and development of data mining techniques to solve ***real-world human health*** related problems for example in **Bhramaramba et al (2011)** a data set taken from protein data pertaining to diabetes mellitus obtained from a genomic database was mined in search of useful patterns and information using data mining techniques on diabetes related proteins. Using data mining techniques for evaluating the psychological performance of human mental health has been examined in **Hengqing and Li (2008)** by using a generalized linear regression model (an Evaluation Model in which the dependent variable and regression coefficients all are unknown) with convex constraint. **Li et al (2010)** adopted an improved **frequential pattern** algorithm of data mining to increase the mining speed of intrusion detection systems which are used to identify any activities of damage to the computer system ***security***, ***integrity*** and ***confidentiality***. In **Huang et al (2009)** data mining was used for automatic frog identification by using DM techniques in identifying frog calls (frog calls are sounds that can be seen as an organized sequence of brief sounds from a

species-specific vocabulary) during an online consultation. It is well known that the number and variety of application areas of data mining is growing drastically as such had made it impossible to exhaust the various area in which it is applied.

Nonetheless other major areas where the advantageous use of data mining techniques applied includes ***human resources management*** and ***control***, ***engineering***, ***pharmaceutics***, ***health, government***, ***medicine***, ***manufacturing, design***, ***telecommunication***, ***education e.t.c***

However, despite these specific application domains where data mining approaches seem ideally suited for, the extensive knowledge discovery capabilities of data mining techniques have also been evident and very pervasive in several other general data analysis activity. Analysing and mining data models such as listed below are some of the implications of the usefulness of data mining intelligence applied in knowledge discovery being functional in exploring the possibility of hidden knowledge that resides in these data:

- Relational data
- Transactional data
- Multi-dimensional data
- Distributed data
- Spatial data
- Multi-media data
- Time-series data
- Text data and
- Data streams and web data

2.4 Knowledge discovery process in data mining

According to **Fayyad et al (1996)**, The general task of discovery knowledge from a database involves the process of retrieving the data from a large data warehouse (or some other source); selecting the appropriate subset with which to work; deciding on the appropriate sampling strategy; selecting target data; dimensionality reduction; cleaning; data mining, model selection (or combination), evaluation, and interpretation; and finally, the consolidation and putting into practical use of the extracted "knowledge."

2.4.1 Spatial Data mining

Attention to location, spatial interaction, spatial structure and spatial processes lies at the heart of activities in several disciplines today and as such demands the urgent development of tools capable of analysing and managing such data which typically can only be represented by means of **geometric features**, for instance, consider the examples of spatial data described by **Perry et al. (2002)** as given below;

1) Percentage cover of woody plants along a line division;

2) Land cover from some rangeland types within a specified area of a coastal region; these include some special cases of spatial data. Finding implicit regularities, rules or patterns hidden in spatial databases is an important task, e.g. for geo-marketing, traffic control or environmental studies **Esther et al. (2001)**. The ultimate goal of spatial data mining is to integrate and further extend methods of traditional data mining in various fields for the analysis and management of large and complex spatial data. The underlying concept is based on the fact that spatial data types (e.g ***points, lines, polygons and regions***) are not supported by the ***conventional database management system***.[10] Studying spatial data management helps us to discover the relationship between spatial and non-spatial data and to be able to build and query a spatial knowledgebase. Geospatial data is the data or information that identifies the geographic location of features and boundaries on earth (such as natural or constructed features), oceans e.t.c. spatial data are usually stored as ***co-ordinates*** and ***topology*** that **can be mapped**. ***They are often accessed, manipulated and analysed through geographic information system***.[11] Spatial data mining and geographic knowledge discovery has emerged as an active research area focusing on the d***evelopment of theory, methodology***, and ***practice*** for the extraction of useful information and knowledge from ***massive and complex spatial databases***, Therefore, there is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity **(Mennis and Guo 2009)**.[12] According to **Gunther and Buchmann (1990**) geographic information systems contain high level spatial operators that are uncommon in conventional database management system (DMS). This has led to an increased development of research issues that focus on ***technologies***, ***techniques*** and ***trends*** that identifies properties that a spatial data model, dedicated to support spatial data for cartography, topography, cadastral and relevant applications, should satisfy.

[10] This has been highlighted in section 1.3 figures 2.
[11] Which is why we have used the Arcgis 10.1 geographic information system software for spatial analyses
[12] Which is the main reason we are working on this project

These properties concern the ***data types***, data structures and spatial operations of the model **(Papadopoulos et al. 2004)**. In their work, **Koperski and Han (1995)** asserted that the for every spatial data object, the attribute data are referenced to a specific location; which means that they are highly ***dependent on location*** and also influenced by neighbouring object (which has given rise to the mining of collocation pattern between spatial objects).

2.4.2 Spatial database

The term spatial database system is associated with a view of a database as containing sets of objects in space rather than images or pictures of a space. The basic element for querying a spatial database is to connect the operations of a spatial algebra (including predicates to express spatial relationships) to the facilities of a DBMS query language. Existing DBMS do not support complex spatial relations that exist between spatial objects thus to achieve this, the functionalities of the DBMS should be extended to incorporate the facilities of these complex spatial relations into their query language by providing for the DBMS a model of how to process and optimize queries over spatial relations (**Clementini et al., 1994)**. Wide application of remote sensing technology and automatic data collection tools has made it possible for tremendous amount of spatial and non-spatial data to be collected and stored in large spatial databases. Spatial database management refers to the extraction of implicit knowledge, spatial relations or other patterns not explicitly stored in spatial databases. Traditional data organisation and retrieval tools can only handle the storage and retrieval of explicitly stored data (**Koperski and Han, 1995**).

The interest of managing a spatial database derives from the need to deal with geometric, geographic or spatial data (i.e data related to space). One remarkable feature of a spatial database is based on the fact that the management of geographic data is split into two distinct types of processing, ***one for the spatial data and another for the attributes of conventional data and their association with spatial data*** **Papadopoulos et al (2004**). Some of the properties that should be considered in a spatial database would include the **data types** the **data structures used**, the operations supported by the data model for the management of cartography, topography, cadastral and relevant applications. Spatial database management system works with an underlying traditional database management system which supports:

- Spatial data models
- Spatial abstract data types and a query language from which these data types are callable
- Spatial indexing, efficient algorithm for processing spatial operations/join and domain specific rules for query optimization.

In general spatial database systems offer the fundamental database technology for geographic information systems and other applications and querying this database is to connect the operations of a spatial algebra (including predicates to express spatial relationships) to the facilities of a DBMS query language **(Güting 1994).**

2.4.3 Spatial data representation

Basically, geographical data can be described in two categories; ***spatial data*** and ***attribute data***. Spatial data describes the location of the object of concern while attribute data tries to specify characteristics at that location (e.g how much, when e.t.c). However representing these data in the form that the computer would understand requires grouping the data into layers according to the individual components with similar features (example layer could be waterlines, elevation, temperature, topography e.t.c).[13] Nonetheless, the data properties of each layer (such as scale, projection, accuracy, and resolution) needs to be set by selecting appropriate properties for each of these layers. In general, two distinct data structures are considered when representing spatial data digitally; ***(i) raster data structure (ii) vector data structure.***

Raster data structure: - Raster data structure according to **Gregory et al. (2009)**, is similar to placing a regular grid over a study region and representing the geographical feature found in each grid cell numerically: for example, **1** for loamy, **2** for clay and so on (in the study of land use/cover or the study of soil types over a region as shown in figure 5 below). ***Rasters are associated with remote sensing, image processing and dynamic modelling, and are easily manipulated using map algebra (e.g. multiplying geographically corresponding cell values in two or more datasets***) and neighbourhood functions (e.g. returning the sum of values in a 3 by 3 cell window). Rasters are simple but often voluminous. Patterns in the data are therefore compressed using run length encoding, quadtrees or wavelets. Raster data represents geographic data by discretizing it equally spaced and quantizing each raster cell. A raster cell is usually a square, but could theoretically be another regular polygon that is able to fully cover an image

[13] This has been clarified in chapter 4 of this project work

area without leaving holes in the covered region, e.g. a triangle, hexagon or rectangle **(Neuman et al. 2010)**. A raster consists of a matrix of cells (or pixels) organized into rows and columns (or a grid) where each cell contains a value representing information, such as temperature. Data stored in a raster format represents real-world phenomena, such as; **Thematic** data (also known as discrete), representing features such as land-use or soils data and **Continuous** data, representing phenomena such as temperature, elevation or spectral data such as satellite images and aerial photographs **ESRI (2010**).[14] Raster data structures are the pixels of an object in a raster representation. The main reason for storing spatial data as a raster data is that:

1. Raster data structure is a simple data structure—A matrix of cells with values representing a coordinate and sometimes linked to an attribute table
2. The raster data model is a powerful format for advanced spatial and statistical analysis
3. Raster data has the ability to represent continuous surfaces and to perform surface analysis
4. It has the ability to uniformly store points, lines, polygons, and surfaces and also
5. Raster data can perform fast overlays with complex datasets.

Application of raster data structure: Raster data structure can be used for;

- Modelling Elevation (DEM)
- Land-cover Analysis
- Modelling Terrain
- Hydrologic modelling and Analysis
- General GIS surface modelling and analysis of continuous surfaces.[15]

[14] Retrieved from http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=What_is_raster_data%3F
[15] This is exactly what we will be doing in the analysis phase.

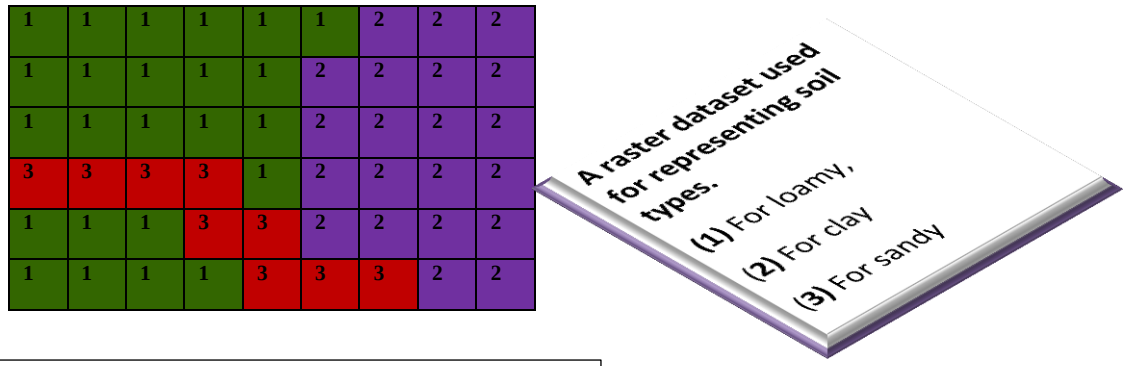


Figure 5: **example raster data representation**

Vector data structure:-

Vector data structure represents geographic objects with the basic elements ***points***, ***lines*** and ***areas***, also called polygons. From the description given by **Gregory et al. (2009),** vector data is based on recording point locations (zero dimensions) using ***x and y coordinates, stored within two columns of a database.*** By assigning each feature a unique ID, a relational database can be used to link location to an attribute table describing what is found there. Line segments (one dimensional) have two points: a start and end node. Polylines are connected line segments; for polygons (two-dimensional) the start and end node is the same. We can also use the vector data structure to encode topological information. Every element in a vector model is described mathematically and bases on points that are defined by Cartesian coordinates **(Neuman et al. 2010**). Vector objects are discrete but sometimes represent continuous fields; for example, as contours. **Esther et al. (2001)** viewed a vector data structure as a data structure used for representing a polygon (area) by its edges or by the points contained in its interior. The most important characteristics of representing data as vectors is that the vector data model can be used to render geographic features with great precision (although this may increase the complexity in data structure which translates to slow processing speed). The reasons for storing spatial data as vector are as follows:

- Small amount of data
- Easy to update
- Logical *data structure*
- Attributes are combined with objects
- Preserves quality after interactivity (e.g. scaling)
- More sophisticated in spatial analysis

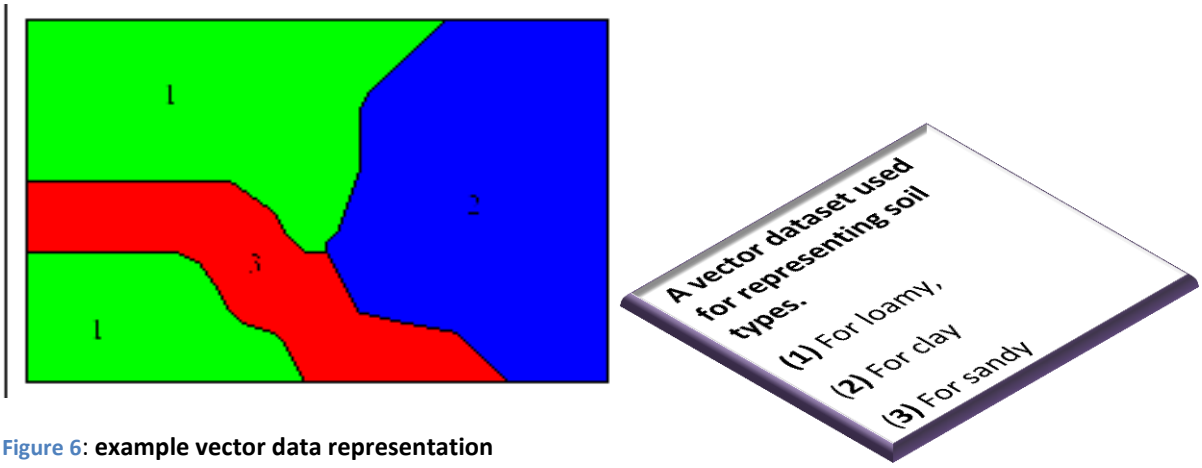


Figure 6: **example vector data representation**

2.5 Knowledge discovery task in spatial data mining

The essence of data mining is to demonstrate the possible contribution of general KDD methods that are not specifically designed for spatially referenced data. Knowledge discovery in a ***spatial database*** involves finding ***implicit regularities***, ***rules*** or p***atterns*** hidden in spatial databases. These are grouped under several basic categories in terms of the kind of knowledge to be discovered. Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, which could be ***computational***, ***statistical***, ***visual***, or some combination of them.[16] Some common spatial data mining task includes:

- ***Spatial classification/prediction***
- ***Spatial association rule mining***
- ***Spatial cluster analysis***
- ***Geo-visualization e.t.c***

[16] We have mentioned this earlier in section 2.3, and have also established that we are going to be carrying out all these tasks.

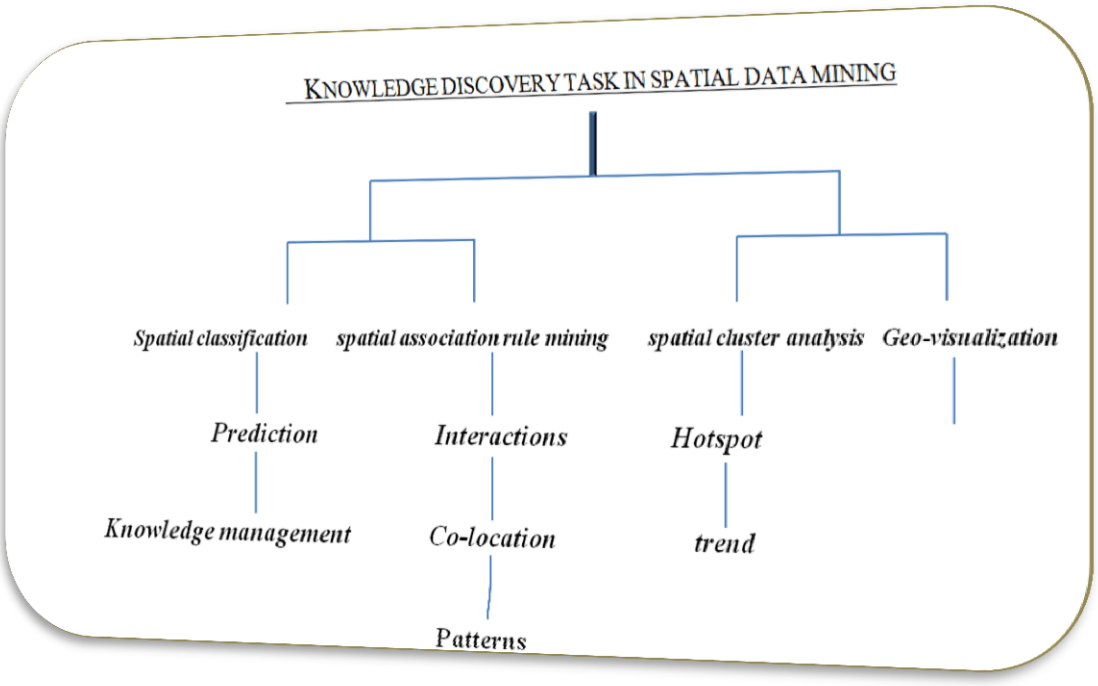


Figure 7: hierarchical view of spatial data mining knowledge discovery: this shows the types of patterns that could be discovered from each different kind of task

2.5.1Spatial classification

Sumathi et al (2001) described spatial classification as predictive spatial data mining, because it involves the initial task of creating a model according to which the whole dataset is analysed. Spatial classification methods extend the general-purpose classification methods to consider not only attributes of the object to be classified but also the attributes of neighbouring objects and their spatial relations **(Guo and Mennis, 2009; Ester et al, 1997)**.[17] For example **Andrienko and Andrienko (1999)** considered applying techniques of knowledge discovery in databases (KDD) to spatially referenced data by combining such techniques with various methods of interactive classification of spatial objects supported by map displays ***(using Descartes for visualization and Kepler knowledge discovery process)***. In that study, they presented an interactive visual (map presentation) method of preparing data for mining and interpreting the result of the C4.5 KDD classification learning algorithm when applied to spatially referenced data. In essence they were able to achieve a synergy of two approaches to exploration of spatial data, visual analysis with the use of interactive cartographic displays and KDD methods. In their

[17] This is the major distinguishing facts about spatial data

work **Wu and Sharma (2012)** examined the ***role spatial contiguity in housing submarket classification***, in this study, they obtained a spatially integrated housing market segments by applying a spatially constrained data-driven submarket classification methodology, the outcome of the study is meant to improve the decision making ability of current and future homeowners on their residential choices

2.5.2 Spatial clustering

Spatial clustering algorithms according to **Sumathi et al (2001)** can be separated into four general categories: partitioning method, hierarchical method, density-based method and grid-based method.

Partitioning Method: - partitioning algorithm organizes the objects into clusters such that the total deviation of each object from its cluster centre is minimized.

Hierarchical Method: - Hierarchical method hierarchically decomposes the dataset by splitting or merging all clusters until a stopping criterion is met.

Density-Based Method: - The method regards clusters as dense regions of objects that are separated by regions of low density (representing noise). In contrast to partitioning methods, clusters of arbitrary shapes can be discovered. Density-based methods can be used to filter out noise and outliers.

Grid-Based Method: - Grid-based clustering algorithms first quantize the clustering space into a finite number of cells and then perform the required operations on the grid structure. Cells that contain more than a certain number of points are treated as dense.

Also, **Fangju and Sun (2002)** described spatial clustering as the grouping together of similar object so that they can be stored together based on the grouping and then referenced together as similar object, the main highlight of this study is to reveal spatial buffering as a means of carrying out spatial clustering in other to overcome the complex data structure of spatial objects. According to **Deng et al. (2012),** there are two (2) types of clustering methods, i.e spatial clustering based on spatial attributes of points only (the geometric coordinates) and spatial clustering that considers mutually, spatial and non-spatial attribute of points. In their own description, they stated that spatial clustering is a technique designed for the classification of a spatial database into several clusters whereby points in the same cluster are similar while points in different clusters are not similar to each other. This classification is done without any previous knowledge (example, probability distribution and the number of clusters). In addition,

Thirumurugan and Suresh (2008) identified an advantage for the use of the clustering method over other method of spatial data mining and they pointed out that so much like the unsupervised learning, the clustering method does not require any prior knowledge in finding interesting structures or clusters.

2.5.3 Spatial association rule mining

Chen et al. (2012) described spatial association rule mining as the discovery of interesting meaningful rule from a spatial database without considering the presence of autocorrelation among the spatial data involved. In **Agrawal et al. (1993)**, association rule mining was described as a tool for computing the statistical significance of any discovered relationship proximity relationship between spatial entities. Spatial association rule **Bembenik and Rybiński (2009)** can be used to discover interesting, useful and hidden patterns in any given spatial database. In a detailed explanation, **Ding et al. (2011)** characterized spatial association rule mining/scoping as consisting of three steps which they listed as ***(i) Discovery (ii) Rule Mining (iii) and scoping***. Discovery according to them has to do with identifying interesting associations rule among the region of study, rule mining depicts mining association rules between the patterns discovered and scoping simple entails determining the scope of the association rule in any given region. Identifying spatial association rule mining as the most important key task of spatial data mining, **Fang et al. (2008)** pointed out that there are specifically two types of spatial association rule that currently needs to be solved (i) lengthways and (ii) transverse spatial association; both of which they said must be computed in any spatial data mining process in other to avoid debasing the efficiency of the system of study.

2.6 Application of spatial data mining

Spatial data mining has been applied to various fields of discipline and human-based activities in general. **Sumathi et al. (2001)** presented two (2) basic application areas of spatial data mining; these include (i) Trend Detections in GIS and (ii) Characterization of Interesting Regions. In **Franklin (1995)** spatial data mining was applied to predictive vegetation mapping; which focused on the development of a remote sensing-based vegetation mapping; a method that was used to illustrate and model the relationship between vegetation and its dependency on ecological niche. Franklin also considered the prediction of plant species distribution, vegetation

pattern over a given region and their dependencies on some environmental constraint e.g precipitation, rainfall, climate, soil e.t.c. Spatial data mining was applied in ecological analysis and in the generally discovery of spatial patterns by **Legendre and Fortin (1989).** In that work, they were able to demonstrate that lots off basic statistical methods used in ecological analyses are compromised by the presence of autocorrelation thus, they presented better ways of performing statistical test irrespective of the spatial contiguity constraint. **Brown (1994)** applied spatial data mining techniques in predicting vegetation types around a tree line. In that study, he was able to present tools and techniques for predicting land-scape vegetation patterns and testing hypothesis about spatial controls on such patterns. Other applications of spatial data mining also exist in other fields that do not have to do with ecosystem study or environmental study for instance **Chen and Chen (2010)** applied spatial data mining in the mining information about the heterogeneity of foreclosed mortgages. They were able to apply spatial data mining techniques to determine the heterogeneity of the portfolio across region in other to make an accurate assessment of the credit risk associated with each of the loan portfolio. **Gaixiao et al. (2010)** applied spatial data mining techniques to marine geographic information system and the output of their work is a new direction for the survey of hydrographic research area. **Pérez-Ortega et al. (2010),** applied SDM techniques in a population – based study of cancer data warehouse. They proposed a spatial clustering algorithm that can generate patterns of stomach cancer this was used as a means of applying data mining to the study of epidemiology. **SDM** has also been applied to image analysis as we can see in the study of **Lee et al. (2007).** They proposed a novel spatial data mining algorithm that can mine the spatial association rules from an image database. In the algorithm called 9DLT-Miner, the image itself is described by the 9DLT representation. **Fang et al. (2008)** applied the proficient power of data miming to the extraction of spatial association among correlation between spatial data and location. This extraction provides potential and useful information for a mobile intelligent client in the field of mobile computing.

Conclusively, we would state that every other discipline that depends on complex decision making (especially when the decision is based on some spatial properties) has benefitted from the tools presented by spatial data mining research.

2.7 Challenges of spatial data mining

In section 1.3, we have highlighted the major challenges faced by a spatial data miner (which we depicted using a diagram) which includes the facts that space is continuous and so on. Other challenges that spatial data mining can contend with has to do generally with modelling spatial data which in most cases has to deal with not only the geographically aspect of the data to be analysed but also the induced complexity caused by change in pattern and time of the spatial data we are considering. **Bailey-Kellogg et al. (2006)** noted that; and they quote:

"There is a complex interplay between 'spatial' in the geographical sense and 'spatial' according to distance in a social network – propagation in one context appears as a discontinuous jump in the other" (**Bailey-Kellogg et al., 2006)**.

Also **Shekhar et al. (2002)** pointed out that spatial context such as ***autocorrelation*** is the key challenge in spatial data mining especially in the area of spatial classification. And then we saw the most obvious challenge of spatial data mining (which is a general problem in field on data mining) in **Wang (2003)** as missing data. Wang acknowledged that since data mining process deals greatly with the development of association rule, patter recognition, classification, estimation and prediction, it will be very pertinent to have serious concern on the accuracy of the database to be modelled and on the sample data chosen for building a training set, in other words, the issue of ***missing data*** must be addressed since ignoring this problem can lead to a partial judgement of the models being evaluated and then finally lead to inaccurate data mining conclusions.

2.8 spatial data mining versus traditional data mining

The ***complexity*** of ***spatial data*** and ***intrinsic spatial relationships*** limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets. **Waller and Gotway (2004)** specified spatial information as being comprised of data that can be located or considered in ***two***, ***three*** or ***more dimensions.*** **Waller and Gotway** also added that the major difference between a spatial and a relational database is in the mode of operations performed on their data. A spatial database not only queries data based on their attributes alone, but also has the ability to query data elements with respect to their locations. According to **(Bolstad, 2002)**, Non-spatial attributes are used to characterize non-

spatial features of objects, such as ***name***, ***population,*** and ***unemployment rate for a city***. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used to define the ***spatial location*** and ***extent of spatial objects*** (see table 3). The spatial attributes of a spatial object most often include information related to spatial locations, e.g., ***longitude, latitude*** and ***elevation, shape***, ***area*** e.t.c. Relationships among spatial objects are **often implicit,** such as ***overlap***, ***intersect***, ***behind*** …. This is quite unlike that of non-spatial objects that are explicit in data inputs according to **Agrawal and Srikant, (1994); Jain and Dubes, (1988)**. One feasible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques - although the materialization may result in loss of information.

Spatial and non-Spatial Data features

Spatial data	*Non-spatial*
Multidimensional	One dimensional
Auto-correlated	Independent

Spatial and non-Spatial Data Processing

Spatial data	*Non-spatial*
Nearby	Sorting
Nearest neighbour	

Spatial and non-Spatial Data Characteristics

Spatial data	*Non-spatial*
Location	Features
Shape	Age
Size	Income
Orientation	

A simple illustration to differentiate between spatial and non-spatial data could be given in the example below consider two cases of similar objects for instance climate and climate change, how do we classify them into spatial and non-spatial; we would observe that climate has

characteristics which does not have anything to do with the location (like climate type, name…) while climate change is dependent on the location of consideration.

Table 3: characteristics of spatial and non-spatial datasets

Spatial data mining versus traditional data	Non spatial data mining	Spatial data mining
Attributes	**Example:** Name, age, height…,	**Example**: spatial location (e.g longitude, latitude and elevation, shape), extent of spatial objects
Relation	**Example: join, relate ….**	**Example:** Overlap, intersect, behind, near, distance e.t.c
Data types	**Example: attributes,**	**Example:** Points, areas or polygon, and lines
Operations	**Example: insert, delete, update e.t.c**	**Example:** Some of the basic operations in mining a spatial database include ; **spatial query**, **layering/overlaying**, **buffering**

Chapter Three

Application of spatial data mining and spatial analysis to the study of ecological behaviour:

INTRODUCTION

The nature of living things and their environments is based on the complex spatial relationships between both entities as such, patterns that are generated from this complexity can only be easily handled by projecting the extracted information into a geographical map which is superimposed to migration patterns or correlated to environmental factors; thus incorporating these environmental, spatial and complex data into models of geographic framework requires a geographic information system **(Sloan et al., 2009)**. The Geographic Information Systems (GIS) is used to integrate these multiple layers of information as a set of powerful hardware and software for inputting, managing, displaying and analysing geographically referenced information **Urbach and Moore (2011).** The technique or method that is applied to such analysis or data integration is referred to as spatial data mining.

Spatial data mining is the quantitative study of phenomena that is located in space. This means that there is an explicit consideration of the location and spatial arrangement of the object to be analysed **(Gatrell and Bailey, 1995)**. The spatial heterogeneity of populations and communities plays a central role in many ecological theories, for instance the ***theories of succession***, ***adaptation***, ***stability***, ***competition, predator-prey interactions, parasitism***, ***epidemics*** and ***other natural catastrophes*** and so on **(Legendre and Fortin, 1989**). In this research we adopted the **Hochachka et al. (2007)** view of a combination of data mining and statistical analyses method in the analyses of our ecological data so as to extract as much insight as possible this is accepted because according to **(Legendre and Fortin, 1989)** many of the fundamental statistical method used in ecological study are impaired on auto-related data. This preposition is supported by **Hochachka et al. (2007**) in adding that most ecologists use statistical methods as their main analytical tools (although the data mining method would have been more appropriate in circumstances where the researcher have little or no knowledge of the system of study) when

analysing data to ***identify relationships*** between a ***response*** and a ***set of predictors***;[18] thus, they treat all analyses as hypothesis tests or exercises in parameter estimation.

3.1 Spatial Analyses

3.1.1 Spatial Data Model:

Spatial data can be analysed by classifying them generally into two distinct categories, known as the ***raster*** and ***vector data models*** respectively; these classification is done based on similar characteristics/feature possessed by the entities of the spatial dataset. Detailed explanation of this has been given in ***section 2.5.2*** above.

3.1.2 Spatial autocorrelation

According to **Legendre and Fortin (1989)** Spatial autocorrelation frequently occurs in ecological data, and many ecological theories and models implicitly adopt an underlying spatial pattern in the distributions of organisms and their environment. **Autocorrelation** arises from the fact that elements of a given population or community (or even the geographic/social environment as a whole) that are close to one another in space or time are more likely to be influenced by the same generating process. According to **Chen et al. (2011)** spatial autocorrelation shows correlation of a variable with itself through space. In their own view, **Rossi and Queneherve, (1998); Legendre, (1993)** acknowledged that spatial autocorrelation measures the similarity between samples for a given variable as a function of spatial distance.

Spatial autocorrelation as seen by **Dale and Fortin (2009)** simply portrays self-dependence of spatial data (meaning that the individual observations made from the chosen samples include information present in other observations, so that the effective sample size, say *n*, is less than the number of observations, n); this dependence according to them poses a great problem that affects the significance rates of statistical test when it is positive and as such must be corrected in other to produce a better measurement of goodness-of-fit.

"Ecological phenomena often are patchy and give data with a wave structure, producing autocorrelation that cycles between positive and negative with increasing distance, further complicating the situation" (**Dale and Fortin, 2009).** Furthermore, **Koenig (1999)** added that

[18] We have used this to achieve the identification for ***cause*** and ***effect*** variables for our prediction model.

the simultaneous fluctuation of ecological variables over wide geographical area is the best explanation for spatial autocorrelation. Consequently we would therefore accept **Getis (2007)** proposition that the concept of spatial autocorrelation (***a special case of correlation - but differs in the sense that it goes ahead to show the correlation within variables across space***), is central to many concerns and is very evident and expressed especially in Regional Science

3.2 The knowledge discovery process

The process of KDD is interactive and iterative, involving several steps such as data selection, data reduction, data mining, and the evaluation of the data mining results.

3.2.1 Data selection:

Why use point pattern analysis:

We have chosen to use the point pattern analysis for our study, because according to **Booth et al, (2006)**, measuring per cent occurrence of objects from digital images can save time and expense relative to conventional field measurements also **Levy, (1927)** and **Levy and Madden (1933)**, ascertained that ecological assessments incorporating ground-cover (the area, usually expressed as a percentage, of ground covered by the vertical projection of vegetation, litter, and rock) measurements have relied on ***point sampling*** using point frames or according to **ITT (1996)** transect methods.

The measurement of ground cover from images has several potential advantages, including acceleration of field work, increased flexibility, repeatability, and convenience in the time and place actual measurements are made

Point pattern terminology:

- ***Point*** is the term used for an arbitrary location
- ***Event*** is the term used for an observation
- ***Mapped point pattern***: all relevant events in a study area **R** have been recorded
- ***Sampled point pattern***: events are recorded from a sample of different areas within a region

3.3 Data Mapping

A reasonable quantitative study of the **spatial** structure of an ecosystem will require a good mapping of the ecological variables. Good maps of environmental suitability for vegetation growth and retention have proved to be an important tool for analysis, and prediction of plant species in an ecological environment. The production of such maps relies on modelling to predict the vulnerability for most of the map, with actual observations of an "***event***" (the occurrence of a patch/bit of plant) usually only known at a limited number of specific locations. Estimation is complicated by the fact that there is often local variation of risk that cannot be accounted for by the known covariates and because data points of measured occurrence of a patch/bit of plant are not evenly or randomly spread across the area to be mapped. In most cases, these maps derive from samples obtained from a surface (like we used in these study), where by intermediate values are being estimated by ***interpolation.***

3.4 Data representation

We shall employ the raster data model for our data analysis because raster is well suited for representing data that changes continuously across a ***landscape (surface)***. Raster provides an effective method of storing the continuity as a surface. They also provide a regularly spaced representation of surfaces. We shall represent the **Elevation, temperature, precipitation** values (from our dataset) measured from the space around the ***Yunnan three parallel river as a surface maps raster so that we can spatially analysed them***. The raster below displays elevation—using green to show ***lower elevation and red, pink, and white cells to show higher elevation.***

3.5 Analyses

Legendre and Fortin (1989) discovered that the spatial structure we find in nature are most of the time gradient of patches, going by this we have based our analyses on the fact that area around the three parallel river are patches of regions containing bits of species of different plant in a given ecosystem.

3.5.1 Attribute analyses

The variation in specific properties of natural phenomenon e.g ***vegetation***, can be described using variables **(Stein et al 2002);** each variable relating to some properties. A variable therefore can take different values, on the basis usually of the properties at the earth surface. We distinguish two types of variables; ***Continuous variables (***variables that take values at a continuous scale); examples are temperature and rainfall. ***Discrete variables*** are variables that take only a limited number of distinct values; example - land suitability. Occurrences of these variables can be labelled and can then be given a name.

In measuring the average relationship between two (2) or more of these variables according to **Chikkodi and Satyaprasad (2010)** in terms of their initial units of data, we normally classify the variables into two (2) categories (***dependent*** and ***independent***). ***Independent*** variables (also known as ***explanator, predictor, or regressor***), possess the value that influences that value of the other variables while the ***Dependent*** variables (also known as ***explained, predicted, or regressed***- which in own case is the ***species***), depends on the independent variable to gain its value.

Attributes of a spatial data are grouped into three main types, which determine the nature of analysis and processing that could be carried out on the data. This classification is listed below

- Uni-variate (one variable or column)

The analysis could be done based on a uni-variate (single independent) variable. In this kind of analysis, we can easily calculate the mode, mean and median of the distribution as a sign of the central tendency; we could also calculate measures of dispersion which may include maximum value, minimum value or standard deviation of each observation point from the mean. **Anderson (2001)** added that another analysis that could be done with uni-variate data is the analysis of variance.

- Bivariate (relating two variables or columns)

Bivariate analysis considers attribute of two (2) data variables

- **Multivariate (more than two variables)**

Multivariate analysis basically deals with the situation where the dependent variable can be expressed mathematically as a combination of any number of independent variables, either linearly or non-linearly **(Kestin 2006).** In most cases, this kind of analysis would usually require external statistical packages such as ***SPSS or SAS***. The study we are involved in deals with more than two variables as such we have carried out multivariate analysis using SPSS (see chapter 6). **Anderson (2001),** acknowledged that the analysis of multivariate data in ecology is one of the major task ecologist face when testing the hypotheses concerning the effects of experimental factors a on a whole collections of species at simultaneously, this is why a strong statistical package like SPSS became useful.

Using the scientific method of project development according to **Riffenburgh (2006),** we have carried out this work in three (3) stages

(1) Describing the events (***using descriptive statistics***)

(2) Explaining these events (***using statistical testing***)

(3) Predicting their occurrence (***statistical modelling, regression and spatial analysis***).

3.6 Existing Solutions

The table below summarises some of the solutions for spatial data analysis and design and also for mining useful patterns from spatial data.

Table 4: Example of existing systems in spatial data mining (application of spatial data mining techniques in various disciplines)

Model	Input	Method	Variables	Output	Issue
GIS-based prediction model (**van Horssen et al.,1999)**	Spatial patterns	Geostatistical spatial interpolation - (kriging).	land use, soil type, and some hydrological processes (ground water data, surface water data and water quality)	Spatial	Response of wetland plant species
Species distribution model, process-based model and Habitat models (**Thuiller et at 2008)**	species' interactions, interaction between of **climate, land-use and d**emography	exploration of existing prediction models and implementation of model by incorporating species' migration into model	Climatic conditions (climate data) Environmental conditions (land-use data)	trailing edge response	Species distribution/migration and species probability
Review of existing models. **Pausas and Austin (2001)**		Measurement and comparism of multivariate environmental gradient by considering different types and lengths of gradients	Temperature, Rainfall Nutrients, Calcium, Water, light, Environmental heterogeneity, Disturbance, altitude, latitude, distance from the coast.		Patterns of plant species richness along environmental gradients
Niche-based model, **(Pearson et al., 2006)**		Extrapolation through model fitting	Min temperature, heat, evaporation, soil moisture,	Impact of niche based modelling to prediction,	Species **range** prediction
Generalized additive model and Generalized linear model **(Bio et al., 2002)**	Site conditions	Krigging (interpolation)	Soil type, management regime (site mowing), Groundwater,	Regression models	Species response to site conditions

Process-convolution model **Cressie et al (2006)**	Functional spatial variables (e.g. pH) and gross production of all variables Biological population and Communities	Geostatistics (ordinary and constrained kriging), Spatial moving average. Mantel test, mapping, statistical method	Water temperature, acidity, microinvertebrate index, oxygen concentration		Change of dissolved oxygen around a river network and its effect on exceedances.

3.7 Requirements Analysis

For the nature of our research, the type of requirement to consider is the non-functional requirements which could be summarised as follows:

- Physical environment (***event locations***, ***multiple sites***, etc.).
- Interfaces (***interaction medium*** etc.).
- Physical or human factors (***what variables would represent suitability factors***).
- Performance (***how well is the algorithm or model functioning in term of predictions***).
- Data (***qualitative substance***).
- Resources (***finding, physical space***).

Chapter Four

MATERIALS AND METHOD

4.0 System Development Methodology

The framework that is used to structure, plan, and control the process of developing an information system in software engineering is known as **software/system development methodology**. These specialised techniques can be utilised for finding scientific truth, making good interpretations of social phenomena, and designing effective systems.

SDLC Processes: the software development life cycle (**SDLC**) describes the various stages involved in every information system development project, from an initial feasibility study through maintenance of the completed application. The basic step/processes involved in every software development life cycle include:

- The existing system is **evaluated/assessed**
- The new system **requirements are defined**
- The proposed **system is designed**
- The proposed **system is developed**
- The system is **put into use/Implemented**
- The new system **is tested**
- The new system **is maintained**

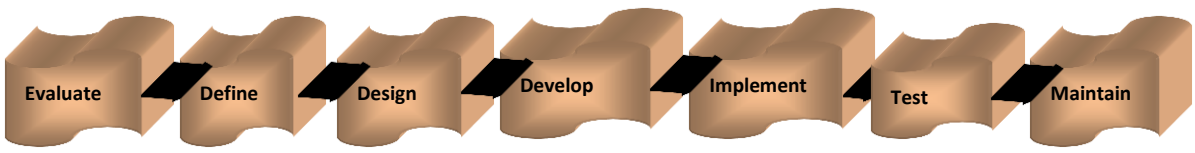


Figure 8: **diagram showing the main stages in a software developmental system**

4.1 SDLC Models:

The software development life cycle model is a framework that describes the activities performed at each stage of a software development project and there are various models that exist. Some of these models include:

- **Waterfall model**: a linear framework

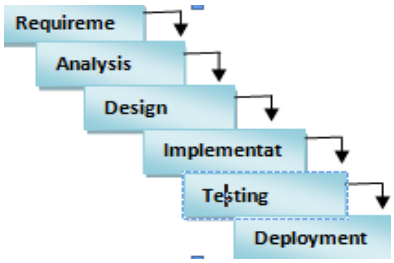


Figure 9: **diagram showing the stages in a waterfall model**

- **Prototyping**: an iterative framework

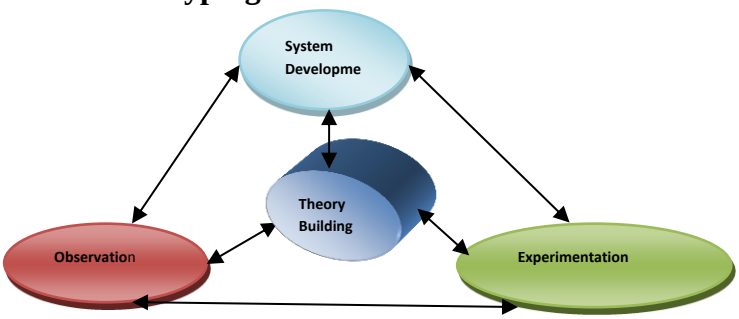


Figure 10: **diagram showing the stages in a prototype model**

The prototype software development methodology consists of series of bidirectional activities that constitutes the main body of the method. These activities include: **(a)** building theories using mathematical models **(b)** developing the system by defining all necessary fields **(c)** working out an experiments; for example through evaluation or by using field data and then **(d)** observation which may involve case studies, survey studies or field studies.

- ***Incremental model***: a combined linear-iterative framework

- ***Spiral model***: a combined linear-iterative framework

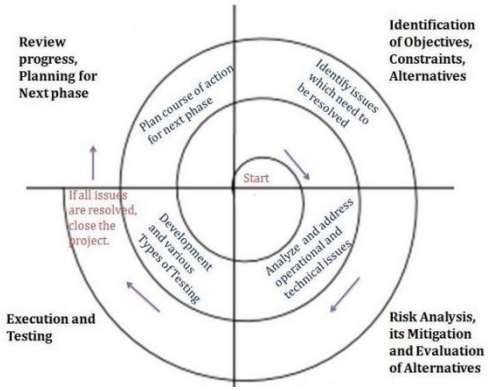


Figure 11: **diagram showing the stages in spiral model**

- ***Agile Methods*:** - **Nuevo et al. (2011)** pointed out that software development in a distributed way leads to multiple complications such as deteriorated communication; this has led to the development of another process development methodology known as the agile system. The Agile model is very useful in the handling of problems that are characterized by change, speed and turbulence **(Highsmith, 2002)**. The agile SDLC model is very expedient as we can see from the points listed below;
  - FDD (Feature Driven Development)
  - Crystal Clear
  - DSDM (Dynamic Software Development)
  - RAD (Rapid Application Development):
  - XP (Extreme Programming)
  - RUP (Rational Unify Process).

4.1.1 Our system development methodology:

The methodology adopted for this work is the **prototyping model**; this was chosen primarily because it well suits our objectives (***considering the explanation given in the prototype model description and figure 10 above***). We considered other methods including the traditional (**a)** water-fall model, **(b)** the agile model, **(c)** the spiral model and the incremental model and we have stated our findings in the table below.

Table 5: **software development models and their application areas**

Method	Stages	fit	Category
WATERFALL	6	Suitable for large scale plan driven project	Traditional hierarchical method
PROTOTYPING (chosen method)	4	*Suitable for building a working baseline model*	Traditional iterative method
SPIRAL	4	high risk projects	**A business management structure**
SCRUM	7	Business management	**Agile**
DSDM	7	High business management	**Agile**
XP	3	Small technical projects: (assumes that participants has interchangeable skills)	**Agile**

4.2: Choice of Software

The software adopted for this project is an open source application software known as **Arcgis 10.1**. This was chosen based on the fact that spatial data mining incorporates the features of classical data mining in its database creations but in addition to this it also considers ***space*** and ***spatial distribution***, as such a database management system that can manage and query the content of a geodatabase would be very necessary; that is why we chose ***Arcgis 10.1.*** more explanations about this has been given in section 1.8 above.

4.3: Data Preparation

Key Point: Tools used in this analysis are based on (1) **spatial analysis** using the **point interpolation** method and (2) **spatial statistics technique** based on modelling **the presence of auto-correlation**; we used the interpolation tool because it is one of the spatial analysis tools used to predict cell values for locations that are not included in a given sample points. Of the three main sites covering the three parallel rivers zone, we derived sample points from the **Lacang** site (this is logically correct because interpolation considers sample points for the prediction of non-sampled and infinite number of points). While spatial analysis bases on the location of the cell on the raster layer, statistics based analysis depends on the attribute value of each layer. Because of the presence of autocorrelation, we used spatial statistics to discover the nature of the pattern that exist among the various plant species and the ecological environment, and then we were able to establish the trend and relationship that exist among them.

Our work is based on the study and analysis of a ***given geographic surface*** using point pattern analysis. Surfaces represent phenomena that have values at every point across their extent (this forms the basis of our study of a spatial system i.e studying object that are related to space). In this case; an aerial photo of one of the major sites around the three parallel rivers (the **Lacang River** zone) was digitised and georeferenced and then some points were taken around some known and identified objects in other to map the land-cover around the river area for interpolation. Based on the fact that the values of points close to sampled points are more likely to be similar to each other that the points farther apart from each other, point interpolation was used to get the value of this set of sample points which was then used to derive the value of the points around the total surface area. **The underlying stimulation behind the operation above is the fact that elements of an ecosystem that are close to one another in space or in time are more likely to be influenced by the same generating process**. These is a way of mapping ecological variables in other to produce either a uni-variate map by interpolation, trend surface or krigging or to produce a map for multivariate data by constrained clustering **(Legendre and Fortin, 1989).**

4.3.1 Study area:

The study area used in this work is Located in the mountainous north-west of Yunnan Province in China (as shown in figure 12 below), it is known as the "Three Parallel Rivers of Yunnan Protected Areas". This area consists of eight geographical clusters of protected areas within the boundaries of the Three Parallel Rivers National Park, in the mountainous north-west of Yunnan Province, the 1.7 million hectare site features sections of the upper reaches of three of the great rivers of Asia: the Yangtze (Jinsha), Mekong (Lancang) and Salween Nu jiang run roughly parallel, north to south, through steep gorges which, in places, are 3,000 m deep and are bordered by glaciated peaks more than 6,000 m high. In addition, due to its location near the boundaries of three major bio-geographic realms, East Asia, South-East Asia and the Tibetan plateau, the park has 22 vegetation subtypes and 6,000 plant species (UNESCO, 2010).

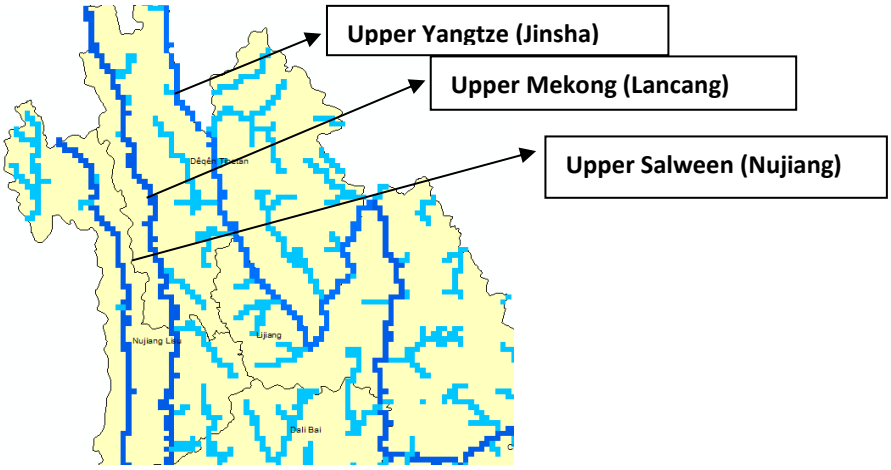


Figure 12: study area in eastern part of China

The Three Parallel Rivers of Yunnan Protected Areas is a natural serial property consisting of 15 protected areas, grouped into eight clusters. The Property contains an outstanding diversity of landscapes, such as deep-incised river gorges, luxuriant forests, towering snow-clad mountains, glaciers, and alpine karst, reddish sandstone landforms (*Danxia*), lakes and meadows over vast vistas (unesco, 2010).

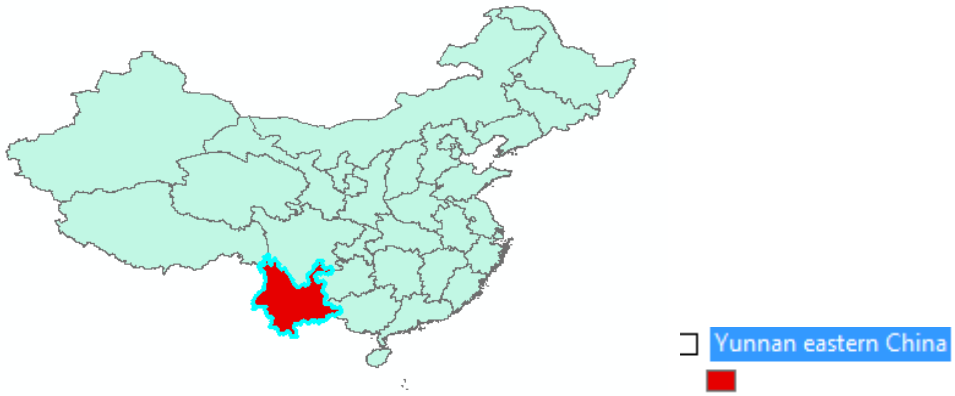


Figure 13: **study area in eastern part of China**

Figure 14: **study area in showing the north-western area of Yunnan province under study.**

According to **Wang et al. (2007)**, the ecological environments are believed to contribute to the plant species distribution and diversity around this river area; these environmental factors include climate (e.g temperature, precipitation), elevation, topography, e.t.c as shown in table 6.

This work is basically a contribution to an earlier study carried out by Wang et al. (2007) as we saw in table 6 below. Based on the data provided in that table, statistical analysis was applied to identify the existing spatial patterns that exist between the elements of our study area.

Table 6: Basic data set for our ecological study (adopted from Wang et al., 2007)

Tuple-ID	mean temperature (monthly 0.1°C)	mean precipitation (monthly 0.1mm)	elevation (m)	topography	plant species
r_1	90, 100, 108, 130, …	0, 7, 21, 21, 307, …	[900,2000]	ascent	Camellia
r_2	80, 111, 130, 102, …	12, 13, 133, 55, …	[500,900]	ascent	Water-lily
r_3	99, 100, 144, 142, …	71, 205, 502, 330, …	[700,1100]	valley	Camellia
r_4	93, 115, 141, 165, …	0, 98, 171, 793, …	[200,700]	ascent	Camellia
r_5	77, 68, 116, 113, …	17, 228, 212, 453, …	[120,400]	valley	Water-lily
r_6	93, 105, 130, 145, …	36, 228, 679, 190, …	[200,800]	valley	Orchid
r_7	93, 103, 120, 151, …	40, 882, 46, 899, …	[600,1200]	basin	Orchid
r_8	67, 84, 81, 105, …	7, 62, 68, 184, 734, …	[1000,2000]	ascent	Water-lily

4.4: Method of Data Collection

Step 1: A real world presentation of the land cover classification based on the scope of our study.

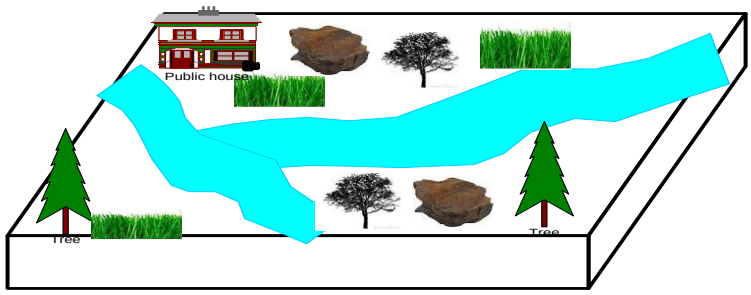


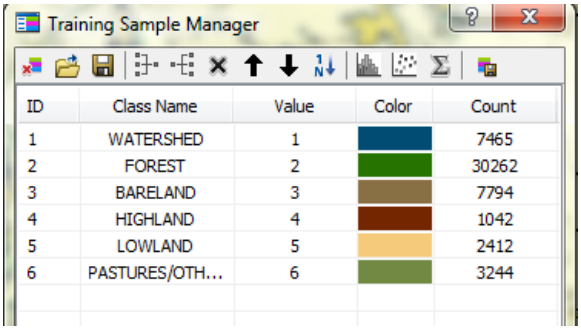
Figure 15: **real life representation of our study area**

Step 2: Preparing the data

While we used the table above to perform statistical analysis, we have chosen for spatial analysis of this study, a raster dataset derived from an image of an areal photograph of the study area as shown in figure 16 below.

Figure 16: **site representing the Arial view of the Lacang section of the three parallel**

Step 3: prepare a training set for the logical classification.



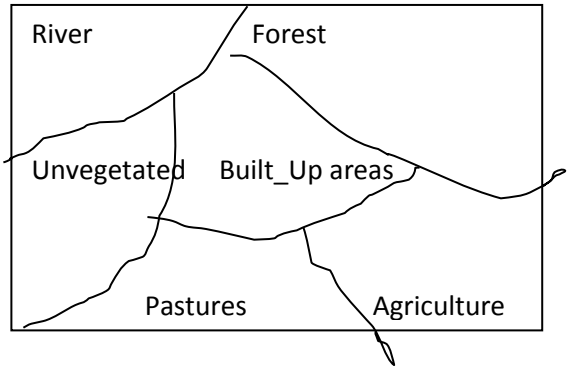


Figure 17: **supervised Classification**

Figure 18: **Classification of the study area according to land-cover**

Step 4: Using the training set, the image of the physical map in figure 14 was classified into six different classes as we can see below:

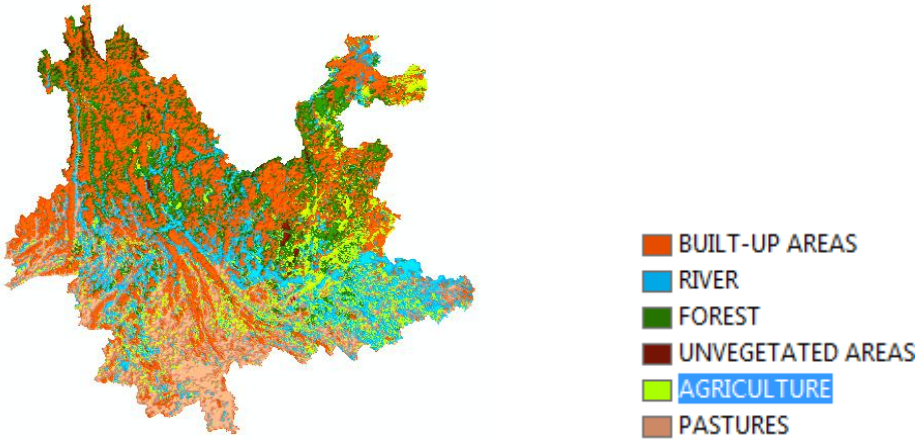


Figure 19: supervised classification of the Yunnan District according to Land Cover (based on geographical map)

Using a similar classification signature file, we are able to transforms the physical map of our study area to the following raster image as presented in our base map in **figure 20**;

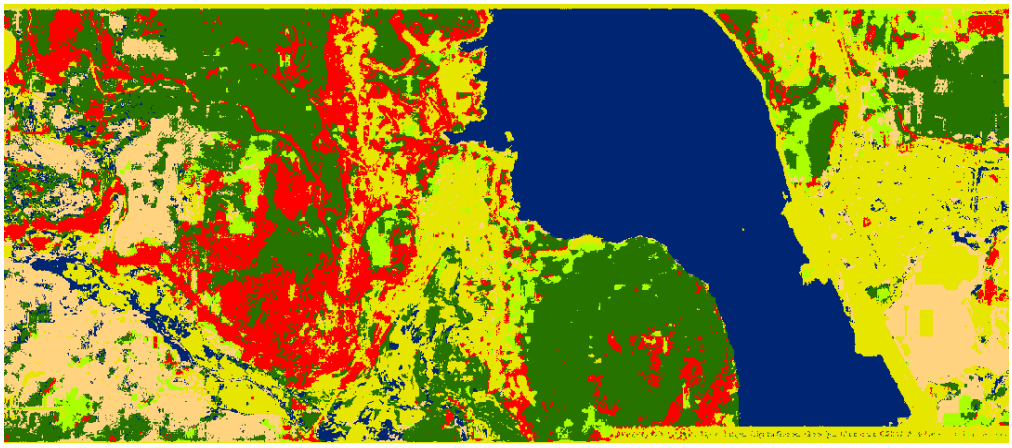
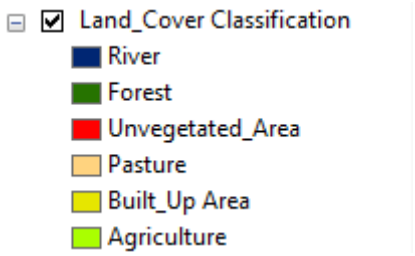


Figure 20: supervised classification of the aerial photo in figure 16 (Lacang zone of the three parallel river - based on satellite image fig 16 above).



Chapter 5:

Data Analysis and Design

5.1: Data Analysis

The relationship between **terrain, climate** and **vegetation** is the main concern of an ecological study **(Hao and Lu, 2010)** and the main goal of that study is to discover the existing association pattern between the plants and those ecological variables in other to be able to retain rare and endangered species of plants or even animals in that area of interest **(Wang 2008**). **Hara et al. (1996)** and **Qv (1984)** also added that the results of plant ecology, both **climatic factor** and **terrain factor** (as we have listed above) are main conditions of the spatio-temporal heterogeneity of vegetation. The former is largely water - precipitation, heat - temperature and their combination, while the latter takes effect by reallocating the combination of water and heat.

For our study of the distribution on plants species around the three parrarel rivers of Yunnan province china, we have classified the study are into **(i) water (ii) forest (iii) bare-soil (iv) pasture (v) roads (vi) river edges;** as shown in figure 15 above.

5.2 Finding the spatial pattern:

The **spatial pattern** of an ecosystem is the spatial regular distribution structure of ecological variables; this has proved to be one of the most embodiment patterns of spatial heterogeneity **(Wang, 1999; Wu 2000).** The first step of our analysis sets out to **describe** the spatial pattern of plant species patches with respect to some attributes which can be seen from the figure 21 below. These points were derived using Arcmap GIS software by sampling points from different locations on the aerial image of our study site given in figure 16 above.

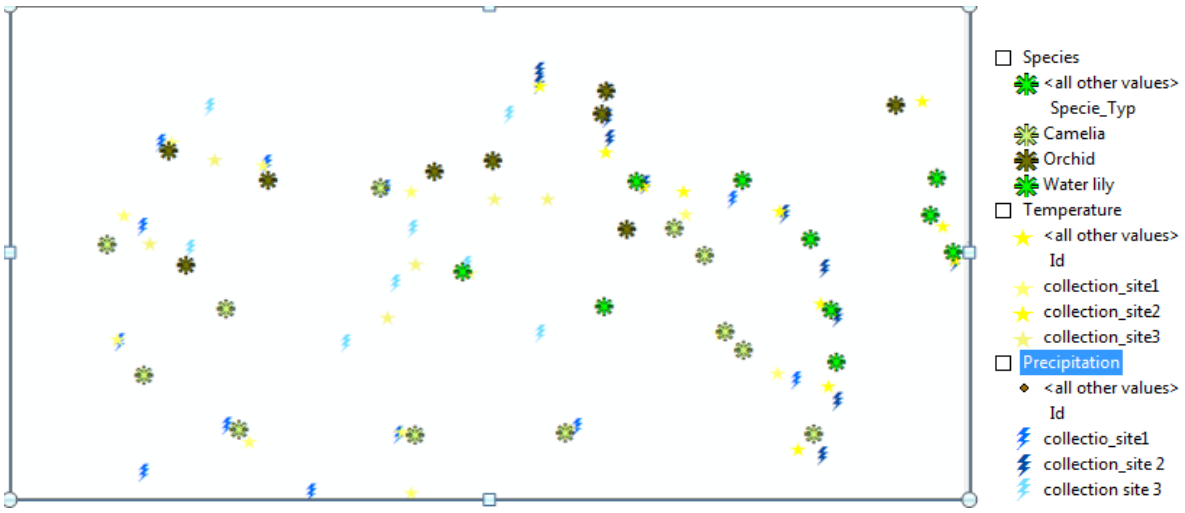


Figure 21: **showing attributes of plant species patches as located around the mountainous north-west of Yunnan Province in China – specifically, around the Lacang river**

As we can see from table 6 above, vegetation on and around the **zone** is dominated by ***Camellia and Water-lily at upper elevations***, and ***orchids*** at lower elevations. The temperate desert climate at the zone averages of 109.4 x 0.1^0c (as in table 6) monthly average precipitation sometimes falls as low as **0 x 0.1mm** (almost falling as snow) during extreme weather conditions.

We are interested in eight **(8)** species of the plants from the study site of the three parallel river protected zone (which is represented by $\mathbf{r_1}$ through $\mathbf{r_8}$ in table 6) and we have also considered two **(2) types** of criterion variables that characterises the heterogeneity of these vegetation types namely ***climatic*** and ***topographic*** as shown in **figure 18** below.

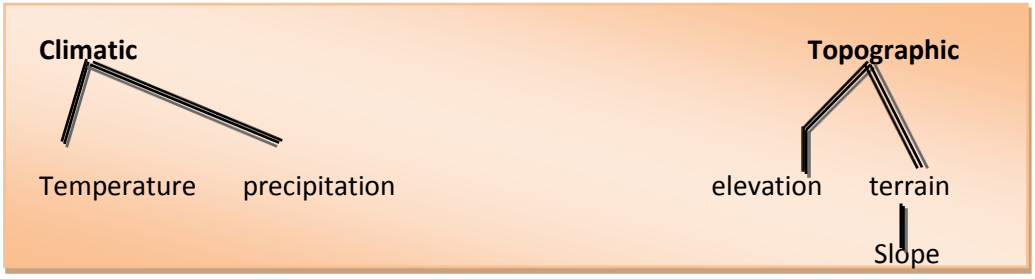


Figure 22: **showing main ecosystem variables.**

The figure below shows the various points on the study surface where point where collected, which gave rise to the figure

5.2.1 Points sampling:

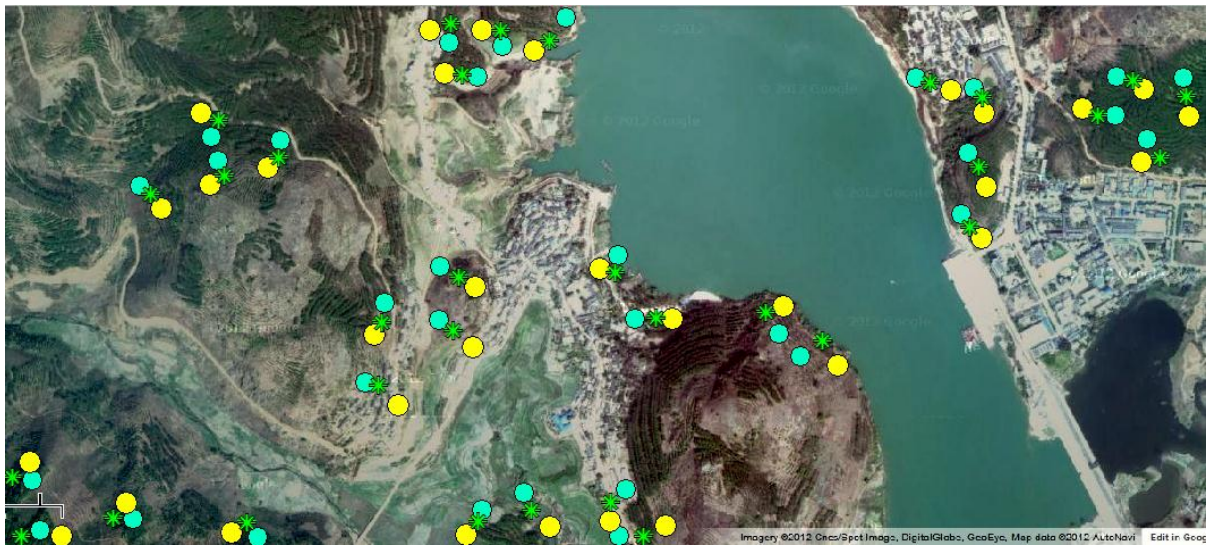


Figure 23: **showing locations where sample point where selected on our basemap**

5.2.2 Sampling methods:

According to **Chikkodi and Satyaprasad** (2010), sampling could be seen as the selection of typical and adequate fraction (finite subset) of the universe, population or bulk. This method or technique depends on the ***nature of the data***, ***source of the data*** and ***the purpose of the enquiry.*** **Ayala et al (2006)** acknowledged that spatial point patterns often arise as the natural sampling of information in many problems.

"The main aim of the analysis of mapped point data is to detect patterns (i.e., to draw inference regarding the distribution of an observed set of locations)" **Waller and Gotway (2004)**. There are basically two (2) types of techniques **Chikkodi and Satyaprasad (2010)** available for analysing collecting spatial data (a) census and (b) sampling. We have adopted the ***sampling*** method of data collection, because we are dealing with data that change across a surface over a period of time e.g temperature, precipitation, e.t.c.

Particular, what we want to achieve in this project is to detect whether the set of locations of plants around the three parallel rivers observed, contains clusters of events reflecting areas with associated increases in the likelihood of occurrence (example unusual aggregations of cases of a particular type of species; or whether these sets of locations contain outliers of events that are possess a large degree of spatial heterogeneity see figures 35 - 37.

5.3: Stating the research hypothesis

Research Hypothesis:

In this research project, our objective is to be able to **describe** the spatial pattern of plant or animal species in an ecosystem with respect to some ecological attributes, therefore going by the research methodology which we have adopted, we shall stating our research hypothesis thus:
H_o: - Plant species types around the three parallel river parks are not significantly auto-correlated with the environmental factors/predictors of that eco-site (if this is true, then we can use the parametric statistical test; **Legendre and Fortin (1989)**).

H_1: - Plant species types around the three parallel river park are significantly auto-correlated with the environmental factors/predictors of that eco-site; this means that there is a significant spatial autocorrelation thus, the value of the ***I*** coefficient would be significantly different from **E (*I*)** which is equal to **– (n-1)** $^{-1}$; which is approximately zero

5.3.1 Analysing pattern

Step 1: modelling the suitability of the **Lacang zone** for plant species through spatial analysis techniques. What we try to derive is a function that satisfies the typical scenario given below

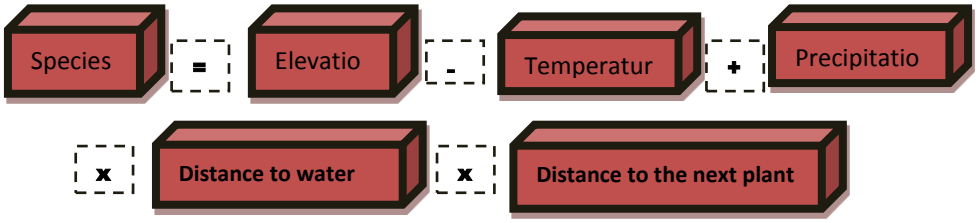


Figure 24: **structure of our Prediction model**

This evolve from deriving such variables as

- Slope from the elevation dataset
- **Creating the** land cover **dataset**
- **Calculated distance data for the** nearest species neighbour
- **Calculated distance data for** river
- **Solar radiation** derived from **temperature** data
- **Watershed** derived from **precipitation** data

The above derivation implies that the availability of a species **AS (**or the suitability of a given point location**)** is a function of the **Sum** of the **weight** of all the **predictors multiplied** by the **product** of the **distant** of the species from the river and the distance of the species to the next species as shown in the equation below:

$$\Rightarrow f(x) = AS = \sum_{i=1}^{n} w_i \, p_i \prod_{k=1}^{n} D_k \quad \text{.....(1)}$$

Where: AS = availability of a species **AS (**or the suitability of a given point location**)**

W_i = **weight** of the **predictors i (p_i)** as parts of the condition or criteria for suitability

P_i = **each ith predictor criterion for suitability**

D_k = all constraint including **space**

Expanding the first part of the equation $\sum_{i=1}^{n} w_i \, p_i$, we have

$$As = \left(w_s p_{s*} w_{lu} c_{lu*} w_{dist\_water} c_{dist\_water*} w_{dist\_neigb\_spec*} c_{dist\_nighb\_spec}\right) \text{..... (2)}$$

Where **s** stand for **slope** as derived from the elevation data shown above and w_s stand for the weight assigned to it. lu stands for land_use as one of the criteria, while w_{lu} stands for the weight assigned it, likewise, **dist_water** stands for the distance of the a given species occurrence from the nearest river network and $w_{\text{dist\_water}}$ is the weight of that distance. Similarly, **dist_neighb_spec** is the distance between any two neighbours of an instance of a given species, while w**dist_neighb_spec** is the weighting assigned to it.

The equation 1 above gives us a clear picture of the measurement of the suitability of an ecosystem for a given plant or animal species. This model can be applied to any form of spatial dataset in other to model or predict the occurrence of any event of interest.

For this analysis, we have chosen some areas where there would naturally be constraints of plant species as areas which include **rivers, road** and **built-up** areas which was represented in the equation 2 above as C_k.

If we then expand $\prod_{k=1}^{n} C_k$, we would have

$$C_k = C_{road} * C_{built\_up} * C_{river} \text{ ..... (3)}$$

Where C_{road} C_{built_up} C_{river}

C_k = constraints

C_{road} = constraint caused by road

C_{built_up} = constraint caused by built_up areas

C_{river} = constraint caused by built_up river

5.3.2 Deriving the constraint variables using spatial analysis

Table 7 below shows the output of creating a buffer zone for the constraint variable in other to determine the distance from each **point** (**representing the species sample)** to the nearest **water line**. This is a measure of the degree of unsuitability or suitability of a particular location (for the existence of a plant species) around the river zone based on distance.

Table 7: minimum and maximum buffer distance for the constraint

Constraint Source	Min Buffer Distance (m)	Max Buffer Distance (m)	Buffer for analysis (m)
Roads	20	200	20
River	10	150	10

Using the above table, we would be able to achieve a new matrix dataset of the form shown below. This in represents an identity raster (or - **(Boolean/Probability**);

Where:

1 represents a cell that is viable

0 represents a cell with constraint

And the resulting data is a **Boolean** raster.

Table 8: output matrix from constraint model

1	**0**	**1**	**1**
1	1	1	0
1	0	1	0

The raster represents the input to the mapping function used for spatial analysis and then it produces the new raster image in figure 31.

5.4 Building the Prediction Model

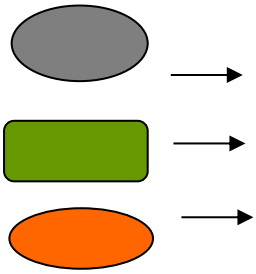
Models are simplifications of reality and they often contribute in system development or research process by helping researchers to formalize their understanding of a particular process or pattern of interest **(Thuiller et al., 2008)**. For example according to **Cressie et al. (2006),** spatially predicting whether nutrient loads exceed pre-specified limits involves an indicator function, which is nonlinear. So by creating a model, one can easily predict the outcome of such a process instead of having to run an experiment every time a similar result is desired.

5.4.1 Field data description

We included input maps for the variables of land cover (obtained from maximum likelihood classification of the base map), roads and water-line (which of which we obtained by digitizing information from existing maps – base map), then spatial maps of distance to water and distance to the closest neighbour were constructed manually from information obtained spatial point distribution of the base map using the spatial analysis tool – ***Euclidean Distance*** – this gave us the nearest neighbour value for two nearest species which cell were closest to each other (cell distributions were characterized by the number of occupied grid cells known as occupancy as illustrated by **Segurado and Arau´ jo, (2004)**. And then the distance of a given plant instance to the nearest water line (was calculated based on the straight-line distance between the two most distant occupied grid cells known as extent of occurrence (**Segurado and Arau´ jo, 2004)**). The other process in the prediction process is described below from steps 1- 4 of section 5.4.1 and steps 1 – 3 of section 5.4.2.

5.4.2 Computing the product of the constraint variables as shown by the equation $\prod_{k=1}^{n} C_k$

Symbol representation



Projected or proposed data

Tool applied to projected or proposed data

The derived data

Step1:

Create the buffer zone around the water_line and the road areas

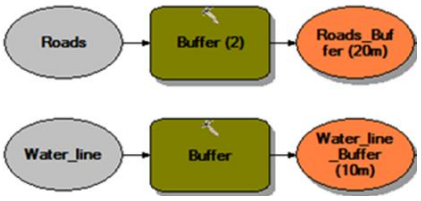


Figure 25: showing output of road and water_line buffer zone

Step2:

Convert the Road and Water_line features for the analysis

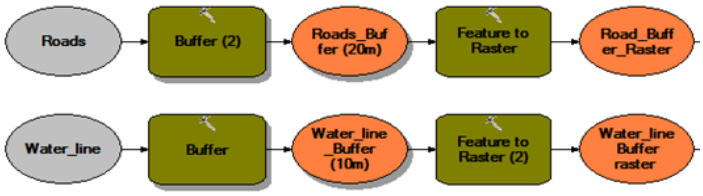


Figure 26: showing road and water_line buffer zone

Step3:

Convert the raster to a Boolean raster. In other to do this we convert **NoData** cells to 0 and convert the viable cells to 1 so as to achieve the matrix **(Boolean/Probability**) raster described above in section *5.3.2*.

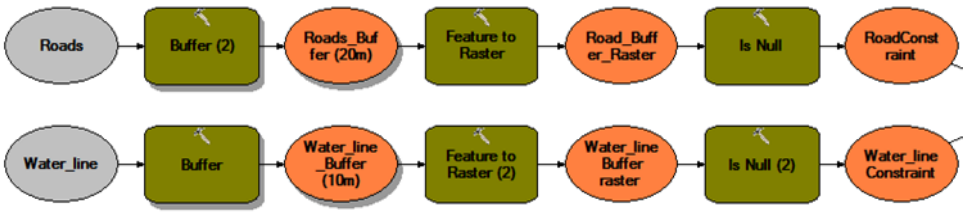


Figure 27: showing the computation of the NoData (null) cell

Step 4:

Multiply all the constraint to according to the constraint function $\prod_{k=1}^{n} C_k = C_{road} * C_{built\_up} * C_{river}$ from equation **(1) and (3)** above, this will give us the

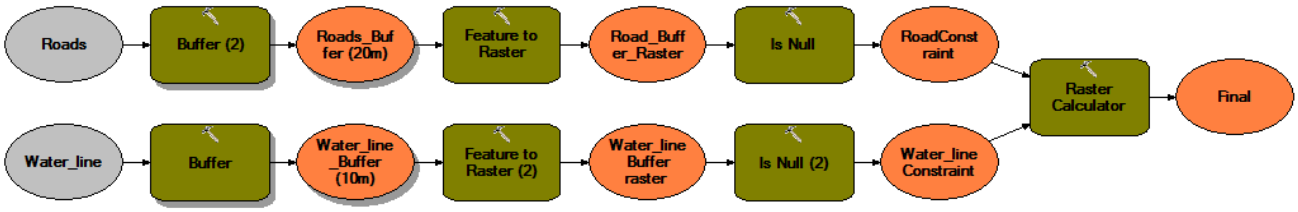


Figure 28: showing the end of the computational model with a final output raster that stand for the anticipated product of constraint as depicted in the map below.

5.4.3 Computing the sum of the prediction variables multiplied by the weight assigned to each as shown by the equation $\sum_{i=1}^{n} w_i p_i$.

Step 1: weighting all prediction variables

Our prediction model above has been based on the fact that the criteria for suitability for any given plant species around the lacing zone of the three parallel river of the Yunnan northwest district is assumed to be the following:

Table 9: prediction criteria weighting

Temperature around the area	***solar radiation***	***30%***
Precipitation rate	***watershed***	***30%***
Elevation	***slope***	***20%***
Distance to the **nearest water body**	***Dist_Water***	***10%***
Distance to the **nearest species** of common family	***Dist_Neighb_spec***	***10%***
		10%

Step 2: Scaling

The scaling range is based on the range of 1- 8 which is a typical classification of sample species chosen for the purpose of this study

Step 3: Overlay all layers based on weight and scale according to its important – so as to derive viability based on the criteria

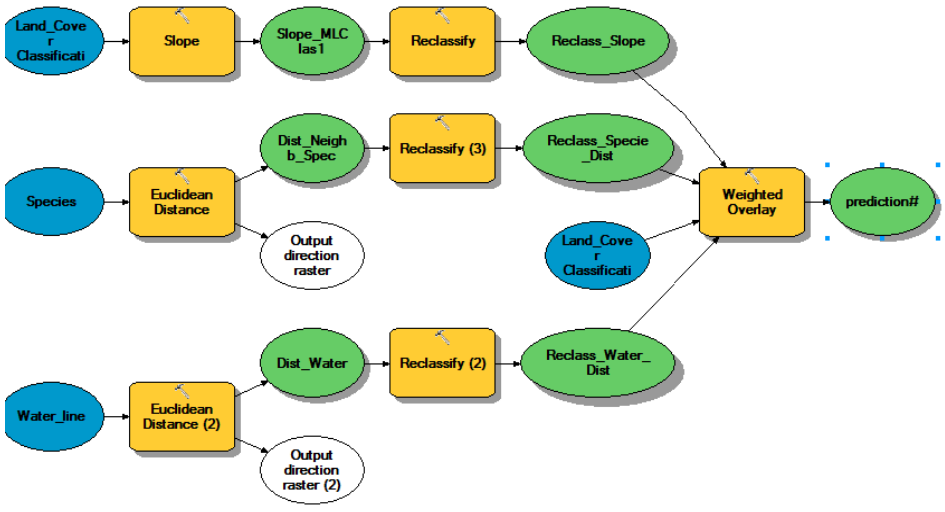


Figure 29: **showing the computation of the overlay (which is a form of superimposing a data against another)**

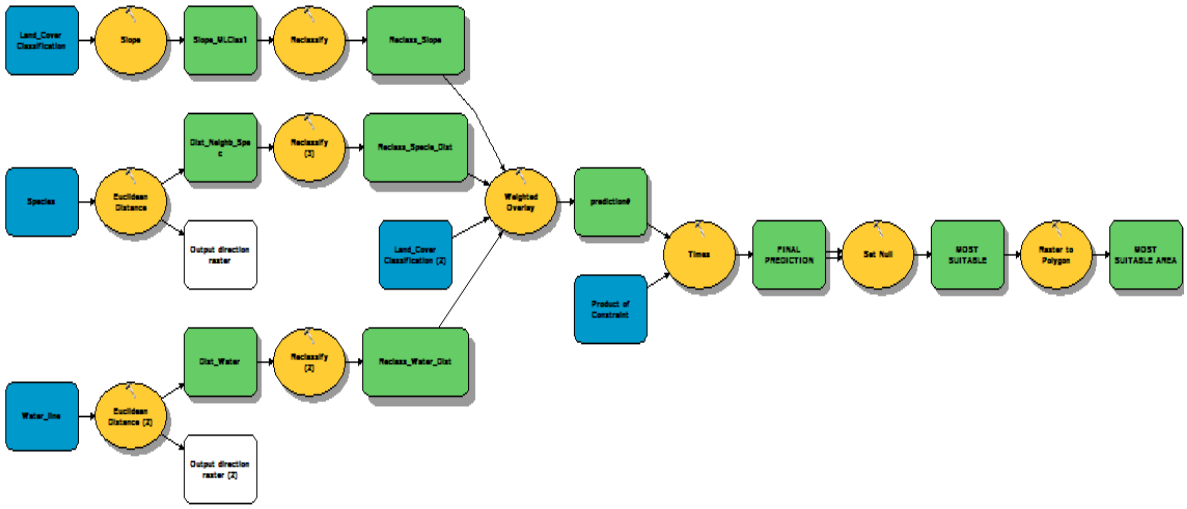


Figure 30: showing the final outcome of the prediction process with a map showing suitable area that plant can grow

CHAPTER 6

Generalisation and Interpretation (Analyses and Result)

6.1 Spatial Analysis

Although **Pausas and Austin (2001)** suggests that Patterns of species richness along some environmental gradients (such as altitude, latitude or distance from the coast/water) do not have any direct causal relationship to plant growth, they also acknowledged the fact that richness with temperature and water availability always show a tendency towards an increase in species in an ecosystem, as such we have considered such distances because what we are interested in is the spatial nature of the ecosystem and its effect to plant probability.

Thus, since the obvious regional difference of spatial heterogeneity of vegetation is induced by the complicated topographical terrain and monsoon climate system, which cause various river hydrology characteristics, soil types, vegetation types, etc as discovered by **Hao and Lu (2010),** we shall therefore conclude based on **Ritchie (2009)** Proposition that that predicting species diversity alongside its major patterns from underlying mechanism such as spreading and resource consumption is the main task to be carried out in the analysis of the study of an eco-system.

Based on the discussion above, we want to present the basis of our analysis at this point. What we are trying do is to carry out a surface analysis in this case, we shall use the calculation made from the surface distances between the species and the river as input for the constraints factors, this in conjunction with the input value from the computed variables from temperature, precipitation and elevation. It was discovered from this at the end of the model that the close distance from the stream is an important consideration when modelling the water-lily species and that a farther distance from the stream would be a factor to consider for species like camellia.

6.1.1 Result of spatial analysis

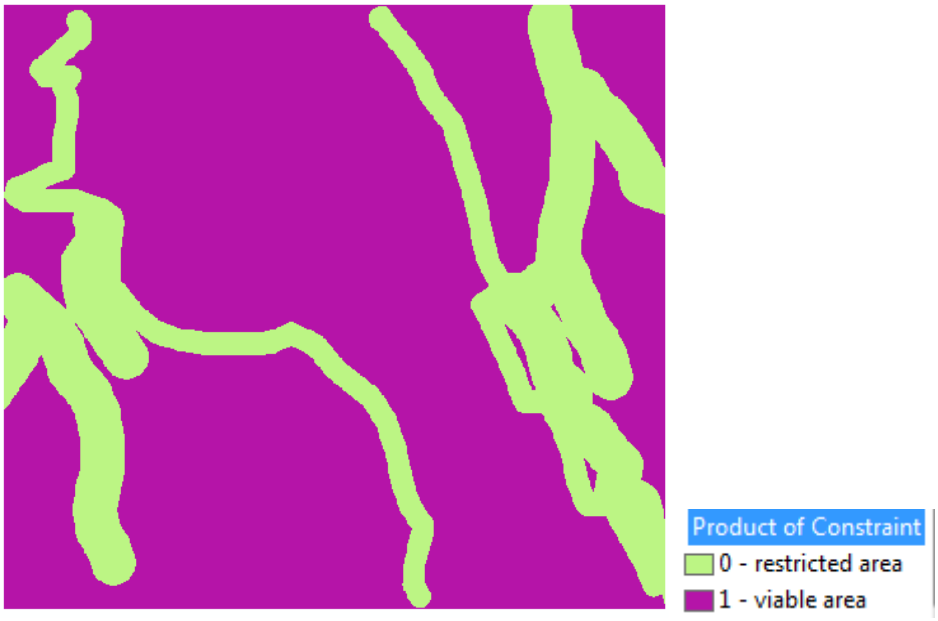


Figure 31: A raster image representation of the table in table 3 above, where 0 represent restricted area and 1 represent viable areas

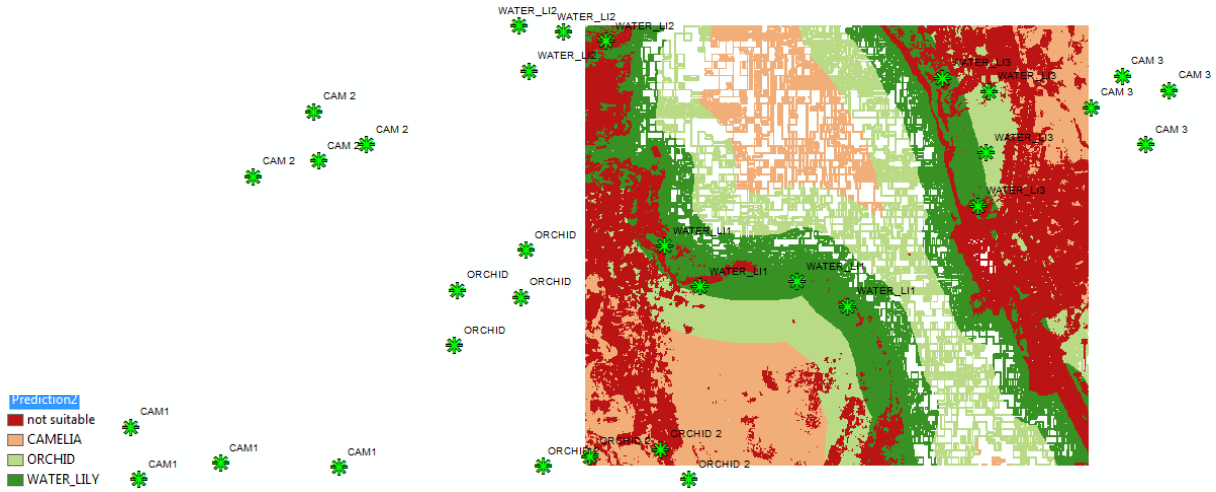


Figure 32: A general raster image representation of figure 27, showing the whole study area as classified by our model obtained by overlaying all the constraint and criteria variables

6.1.2 Result of the prediction model (predicting the presence of plant species around the river)

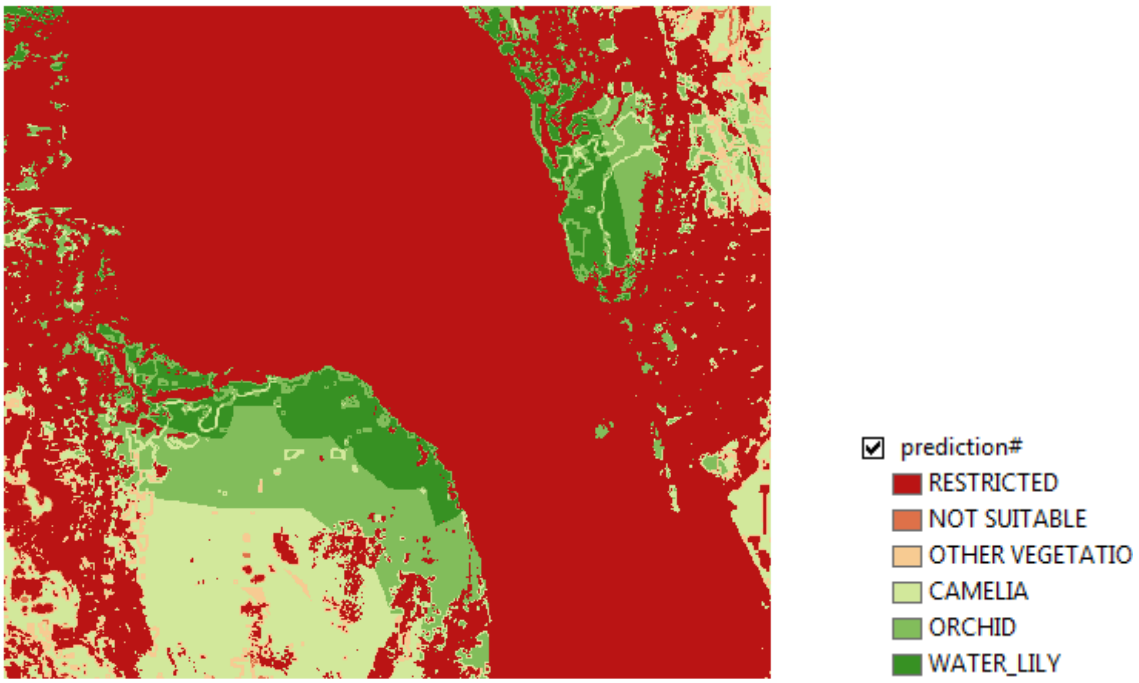


Figure 33: **A raster image representation of image 28 above showing a prediction value**

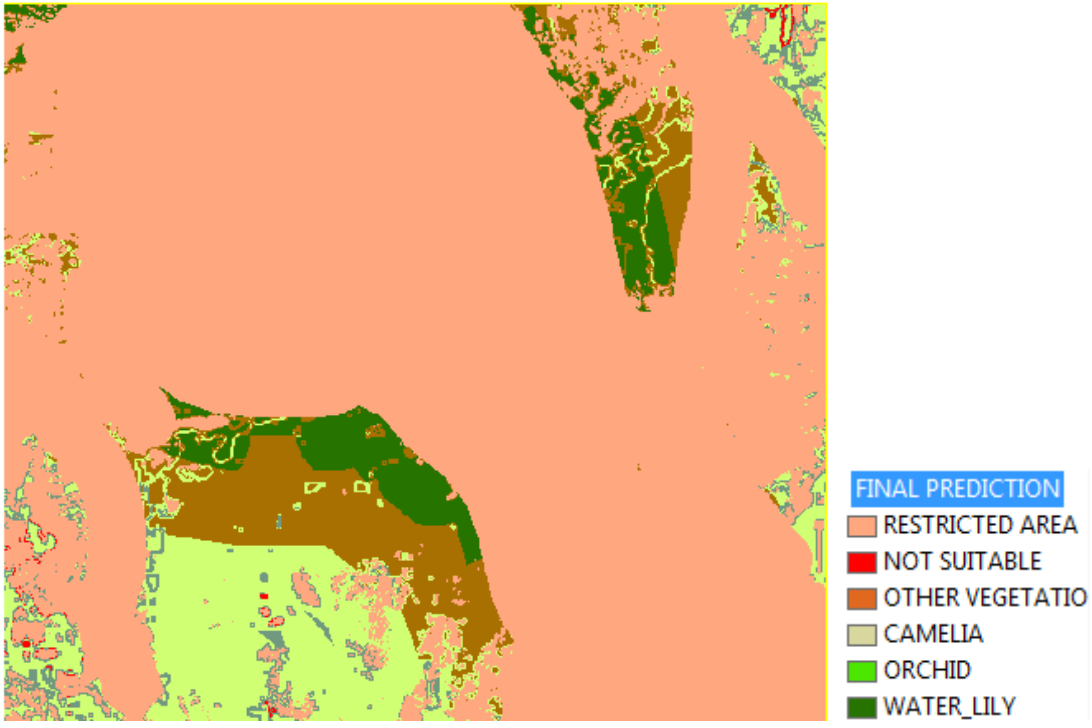


Figure 34 : **final suitability model showing only areas that is suitable for any plant species to grow**

6.1.3: Analysis

The maps above represents the outcome of the prediction analysis as carried out in chapters five (5) and six (6). The results are explained by the map legends besides each of them and they all show the suitability of the lacing zone of the parallel rivers initially represented in figure 16.

The next thing we shall embark on is the second stage of the analyses - ***statistical analysis;*** the statistical analysis helps us to identify patterns that exist among the variables of the elements of the three parallel rivers area.

Some of the variable as identified from the table 6 in chapter 4 above include; temperature, precipitation, and elevation. Using a statistical analysis tool like the SPSS, we wish to determine which of the attributes of an ecosystem affects the species of that ecosystem, the degree of effectiveness and non-effectiveness.

6.2 Statistical analysis

6.2.1 The resulting table from the process of spatial analysis (this will form the input data for our statistical analysis)

Table 10: prediction criteria weighting

FID	Shape	Id	Spec_Name	FID	Shape	Temp_Value	Spec_Name	FID	Shape	Id	Prec_Value	Spec_Name	FID	Shape	Id	Elev_value	Spec_Name
0	Point	1	CAM1	0	Point	90	CAM1	0	Point	1	0	CAM1	0	Point	0	900	CAM 1
1	Point	1	CAM1	1	Point	100	CAM1	1	Point	1	7	CAM1	1	Point	1	1200	CAM 1
2	Point	1	CAM1	2	Point	108	CAM1	2	Point	1	21	CAM1	2	Point	1	1600	CAM 1
3	Point	1	CAM1	3	Point	130	CAM1	3	Point	1	307	CAM1	3	Point	1	2000	CAM 1
4	Point	2	WATER_LI1	4	Point	80	WATER_LI1	4	Point	2	12	WATER_LI1	4	Point	2	500	WATER_LI1
5	Point	2	WATER_LI1	5	Point	111	WATER_LI1	5	Point	2	13	WATER_LI1	5	Point	2	600	WATER_LI1
6	Point	2	WATER_LI1	6	Point	130	WATER_LI1	6	Point	2	133	WATER_LI1	6	Point	2	800	WATER_LI1
7	Point	2	WATER_LI1	7	Point	102	WATER_LI1	7	Point	2	55	WATER_LI1	7	Point	2	900	WATER_LI1
8	Point	3	CAM 2	8	Point	99	CAM 2	8	Point	1	71	CAM 2	8	Point	1	700	CAM 2
9	Point	3	CAM 2	9	Point	100	CAM 2	9	Point	1	205	CAM 2	9	Point	1	790	CAM 2
10	Point	3	CAM 2	10	Point	144	CAM 2	10	Point	1	502	CAM 2	10	Point	1	850	CAM 2
11	Point	3	CAM 2	11	Point	142	CAM 2	11	Point	1	330	CAM 2	11	Point	1	1100	CAM 2
12	Point	4	CAM 3	12	Point	93	CAM 3	12	Point	1	0	CAM 3	12	Point	1	200	CAM 3
13	Point	4	CAM 3	13	Point	115	CAM 3	13	Point	1	98	CAM 3	13	Point	1	300	CAM 3
14	Point	4	CAM 3	14	Point	141	CAM 3	14	Point	1	171	CAM 3	14	Point	1	500	CAM 3
15	Point	4	CAM 3	15	Point	165	CAM 3	15	Point	1	793	CAM 3	15	Point	1	700	CAM 3
16	Point	5	WATER_LI2	16	Point	77	WATER_LI2	16	Point	2	17	WATER_LI2	16	Point	2	120	WATER_LI2
17	Point	5	WATER_LI2	17	Point	68	WATER_LI2	17	Point	2	228	WATER_LI2	17	Point	2	180	WATER_LI2
18	Point	5	WATER_LI2	18	Point	116	WATER_LI2	18	Point	2	212	WATER_LI2	18	Point	2	300	WATER_LI2
19	Point	5	WATER_LI2	19	Point	113	WATER_LI2	19	Point	2	453	WATER_LI2	19	Point	2	400	WATER_LI2
20	Point	6	ORCHID	20	Point	93	ORCHID	20	Point	3	36	ORCHID	20	Point	3	200	ORCHID
21	Point	6	ORCHID	21	Point	105	ORCHID	21	Point	3	228	ORCHID	21	Point	3	400	ORCHID
22	Point	6	ORCHID	22	Point	130	ORCHID	22	Point	3	679	ORCHID	22	Point	3	600	ORCHID
23	Point	6	ORCHID	23	Point	145	ORCHID	23	Point	3	190	ORCHID	23	Point	3	800	ORCHID
24	Point	7	ORCHID 2	24	Point	93	ORCHID	24	Point	3	40	ORCHID	24	Point	3	600	ORCHID
25	Point	7	ORCHID 2	25	Point	103	ORCHID	25	Point	3	882	ORCHID	25	Point	3	900	ORCHID
26	Point	7	ORCHID 2	26	Point	120	ORCHID	26	Point	3	46	ORCHID	26	Point	3	1000	ORCHID
27	Point	7	ORCHID 2	27	Point	151	ORCHID	27	Point	3	899	ORCHID	27	Point	3	1200	ORCHID
28	Point	8	WATER_LI3	28	Point	67	WATER_LI3	28	Point	2	7	WATER_LI3	28	Point	2	1000	WATER_LI3
29	Point	8	WATER_LI3	29	Point	84	WATER_LI3	29	Point	2	62	WATER_LI3	29	Point	2	1300	WATER_LI3
30	Point	8	WATER_LI3	30	Point	81	WATER_LI3	30	Point	2	184	WATER_LI3	30	Point	2	1800	WATER_LI3

6.2.2 Problem Statement:

What we are trying to do is to analyse the relationship between the time series data, and then see how they are related or how they affect the plant species.

Given:

Data → **Time Series (Temperature, Precipitation)**

Problem → **Non-stationarity)/Autocorrelation**

Test → **Statistic test using the *Durbin-Watson test Statistic***

6.2.3 Describing variable data

We would be using the table below for our analyses (the table is derived from the summary of the table in section 6.2.1, which contains only the Independent and Dependent variables -IVs and DVs- necessary for statistical analysis)

Table 11: prediction criteria weighting

Spec_Name	Id	Temp_Value	Elev_value	Prec_Value
WATER_LI1	2	80	500	12
WATER_LI1	2	111	600	13
WATER_LI1	2	130	800	133
WATER_LI1	2	102	900	55
WATER_LI2	5	77	120	17
WATER_LI2	5	68	180	228
WATER_LI2	5	116	300	212
WATER_LI2	5	113	400	453
WATER_LI3	8	67	1000	7
WATER_LI3	8	84	1300	62
WATER_LI3	8	81	1800	184
WATER_LI3	8	105	2000	734
CAM 2	3	99	700	71
CAM 2	3	100	790	205
CAM 2	3	144	850	502
CAM 2	3	142	1100	330
CAM 3	4	93	200	0
CAM 3	4	115	300	98
CAM 3	4	141	500	171
CAM 3	4	165	700	793
CAM1	1	90	900	0
CAM1	1	100	1200	7
CAM1	1	108	1600	21
CAM1	1	130	2000	307
ORCHID	6	93	200	36
ORCHID	6	105	400	228
ORCHID	6	130	600	679
ORCHID	6	145	800	190
ORCHID 2	7	93	600	40
ORCHID 2	7	103	900	882
ORCHID 2	7	120	1000	46
ORCHID 2	7	151	1200	899

If we assume that **q** independent variables (e.g **temperature, precipitation**) are potentially related to the a dependent variable (**species**), and If we have **N** sample points, we calculated the covariance of the **x, y** points (**temperature, precipitation** in our own case), using **T** to stand for temperature, **S** for species and **P** for precipitation, we need to calculate autocorrelation because we are trying to predict how these data change over time and how they affect the plant species around the zone. We present the result of the descriptive statistics as below;

Table 12: description of variables and their statistical description

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
FID	32	0	31	15.50	9.381
ID	32	1	8	4.50	2.328
TEMPERATURE	32	67	165	109.41	24.955
PRECIPITATION	32	0	899	237.97	278.156
ELEVATION	32	120	2000	826.25	507.281
Valid N (listwise)	32				

6.2.4 Finding Patterns

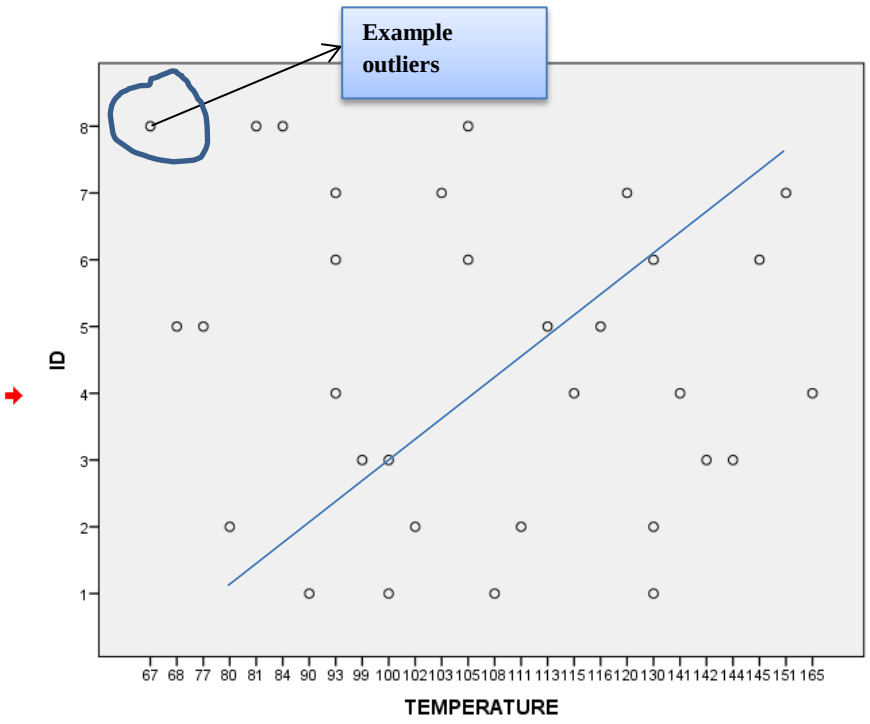


Figure 35: **diagram of the spatial auto-correlation of temperature data showing the nature of the pattern of temperature around the river zone – this explains phenomenon like outliers, collation, and association rule e.t.c.**

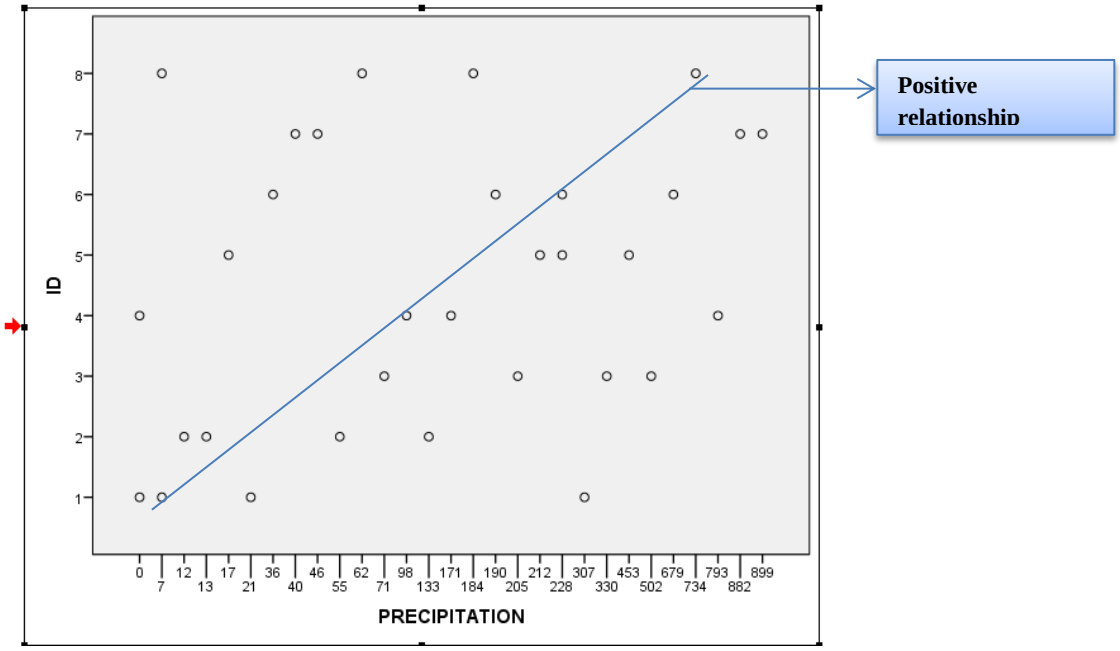


Figure 36: **diagram of the spatial auto-correlation of the precipitation data against itself – showing a linear positive relationship**

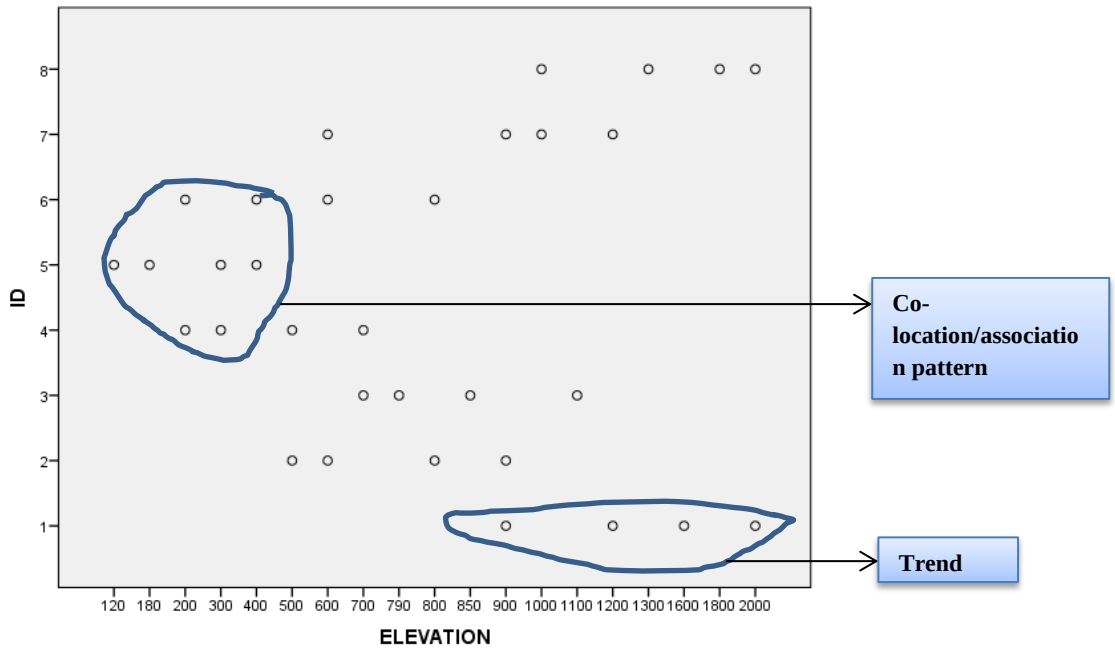


Figure 37: **diagram of the spatial auto-correlation of the elevation data, the curve depicts inconsistency suggest that there is no defined linear relationship among the data**

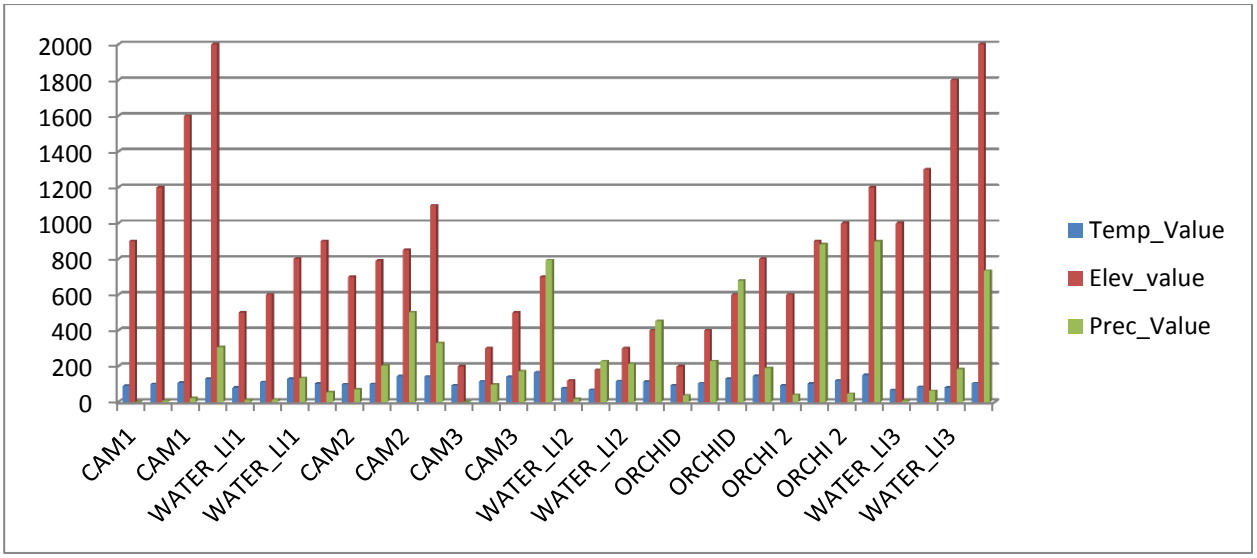


Figure 38: **using a chart to show patterns that exist between the ecological variables and the species type**

This chart above shows the basic trend that exists between the ecological variables and the species type around the three parallel river zones, with low temperature around the zone, high elevations and high precipitation can be perceived. The chart also shows that the Orchid grows basically around areas of average precipitation and low temperature while Camellia and Water-lily enjoy higher precipitation and higher elevation; see clarification of patter in figure below.

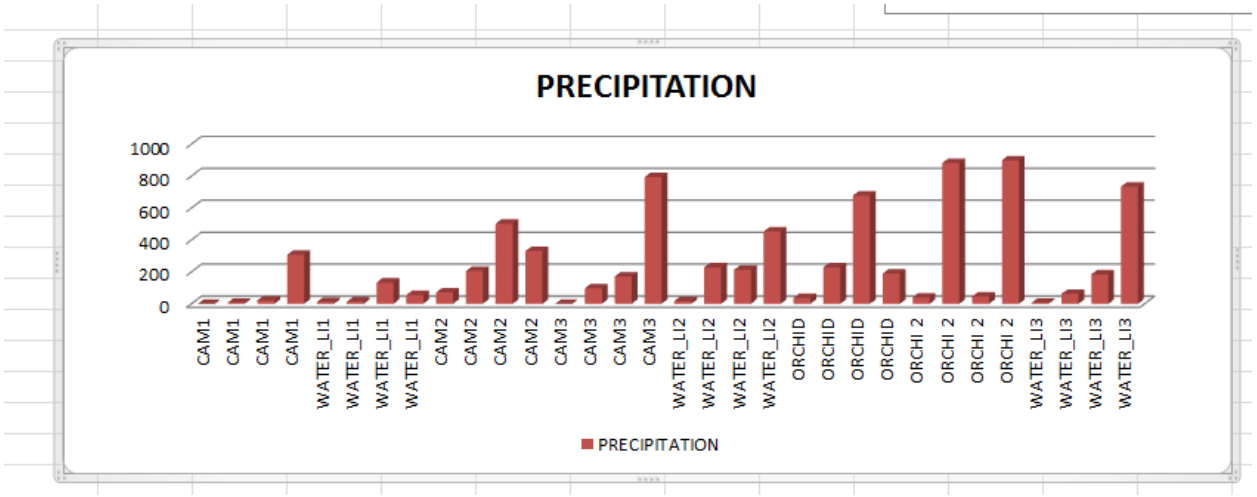


Figure 39a: how precipitation affects the species around the river zone

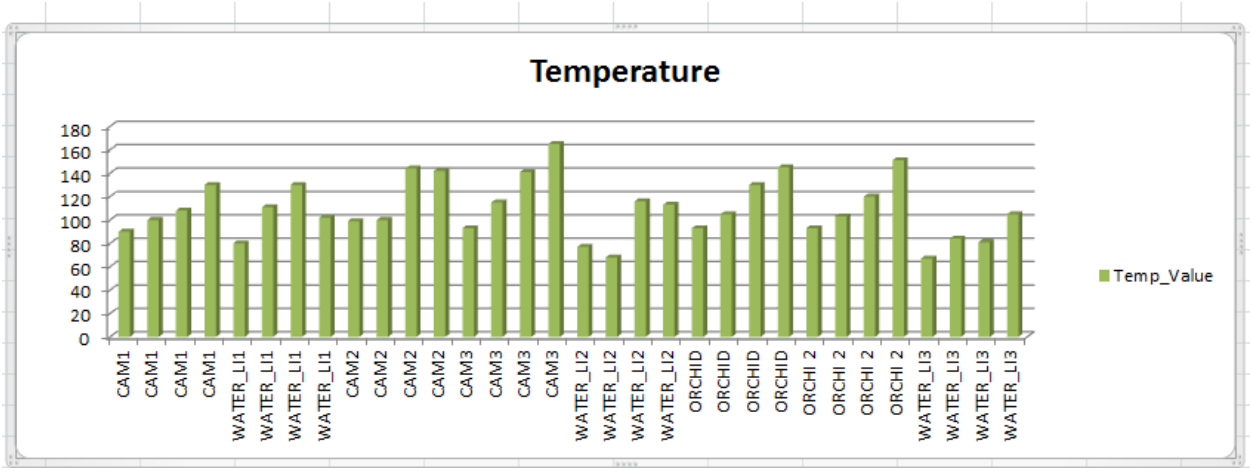


Figure 40b: how Temperature affects the species around the river zone

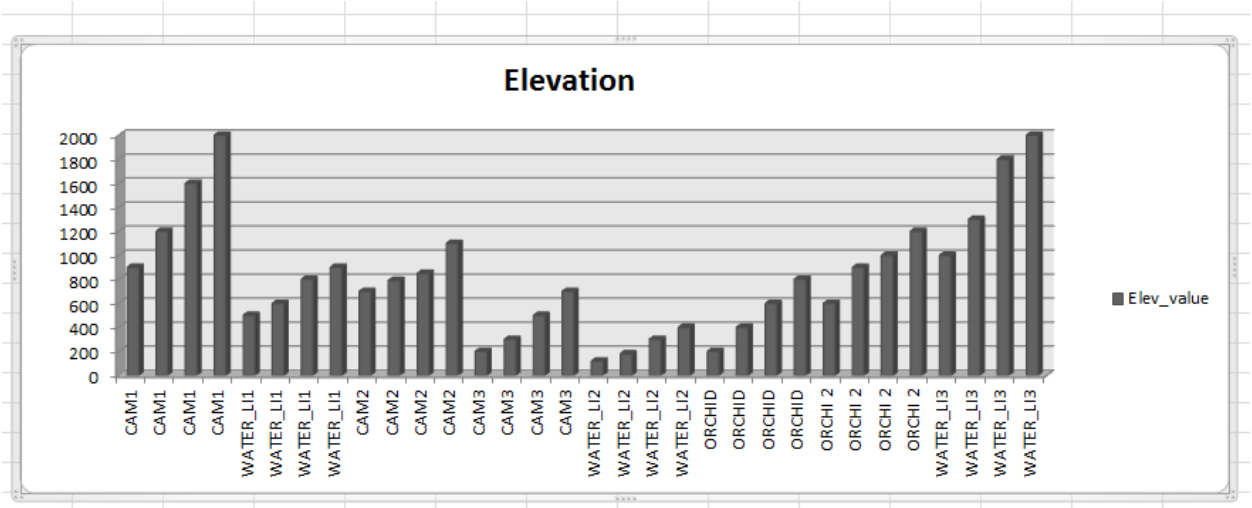


Figure 41c: how elevation affects the species around the river zone

6.2.5 Graphical representation of the time series data

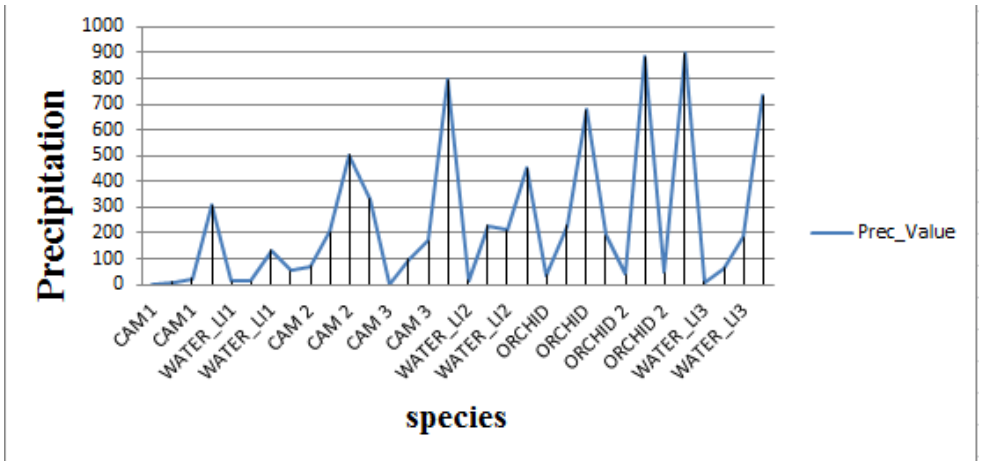


Figure 42: **time series diagram for precipitation data**

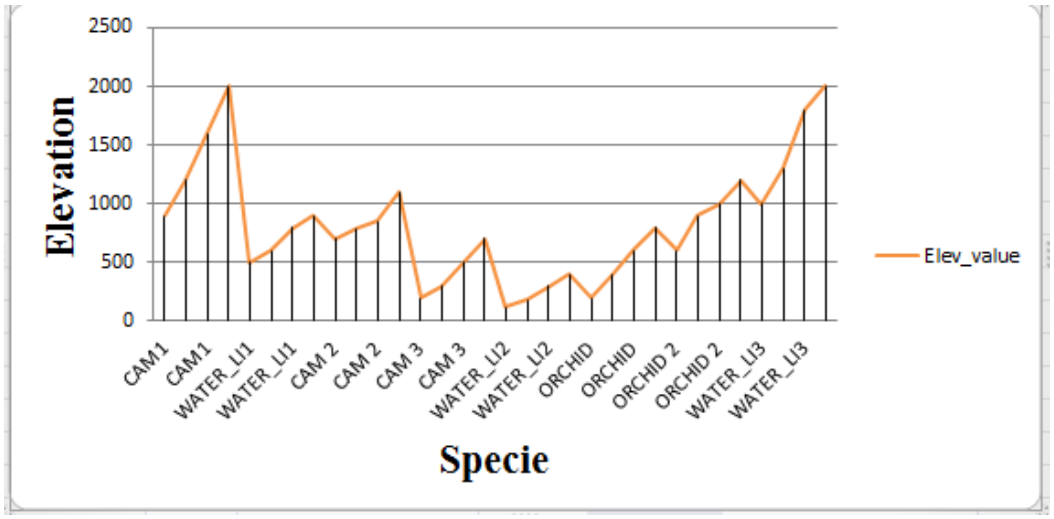


Figure 43: **time series diagram for elevation data**

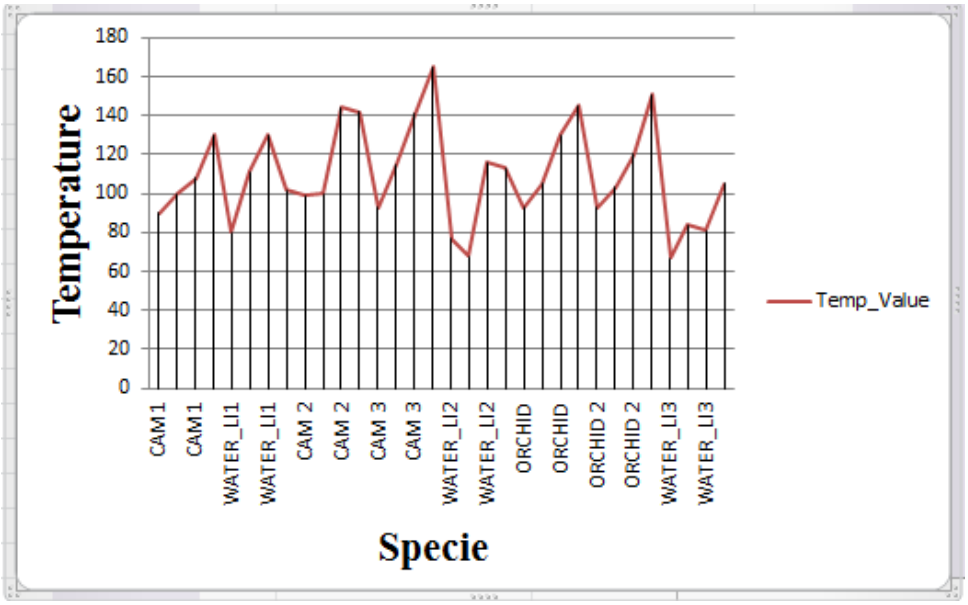


Figure 44: **time series diagram for temperature**

6.2.6 Testing for autocorrelation among the variables

The correlation of a time series data with its own historical and future value can simply be termed **autocorrelation**. This has been presented using the diagrams below; each diagram explains the nature of the pattern that exists in a given eco-site – showing outliers and trends.

According to **Nerlove and Wallis (1996),** in other to use the ***Durbin-Watson test Statistic*** for testing auto-correlation, the null hypothesis states that there is a significant serial independence among the residuals of a regression analysis but the alternative hypothesis states that there is a positive autocorrelation among the regressor variables – although this power is being limited by the presence of lagged dependence among the regressor variables (this explains the concept of spatial outliers – a case of error variables which are not considered in the mining process).

We used the ***Durbin-Watson test Statistic*** for testing for the presence of autocorrelation. Because we are dealing with multiple time series data which always shows sign of positive autocorrelation, the test is considered significant in our study because it considers the fact that residuals from a multiple regression analysis are independent. This helps us accept our ***null hypothesis*** below and then reject the alternative or vice versa. Generally, the Durbin Watson test is of the form:

H_0: q = 0

H_1: q > o

This means that **H_0**, then we are saying that the residual of the regression analysis **q** equals 0

H_0: - Plant species types around the three parallel river parks are not significantly auto-correlated with the environmental factors/predictors of that eco-site (if this is true, then we can use the parametric statistical test; **Legendre and Fortin (1989)**).

H_1: - Plant species types around the three parallel river park are significantly auto-correlated with the environmental factors/predictors of that eco-site; this means that there is a significant spatial autocorrelation thus, the value of the *I* coefficient would be significantly different from E (*I*) which is equal to – $(n-1)^{-1}$; which is approximately zero.

We can summarise the test as below according to **white (1992):**

Given that **t** is the position of each species that occurs at a given location around the study space, **y(t)** is the value of the response variable obtainable in position **t**, which is affected by the value

$\mathbf{x_t}$ at a specific point in time in $\mathbf{x_i}$ where $_{\mathbf{i}}$ the number of predictor variables = 1….. **k** and **n** is the total number of observations obtainable by number of sample points, then the ***Durbin-Watson test Statistic*** can be given as:

$$d = \frac{(\sum_{t=2}^{n}(e_t - e_{t-1})2}{\sum_{t=1}^{n} e_t^2} \quad \textbf{eqn. (1)}$$

Where $\mathbf{e_i}$ the i_{th} residuals equals (=) the value of the observed y_t at the i_{th} observation of the response variable minus (-) the predicted y_t at that observation, *d* is the value of the test of significance which is ***Durbin Watson's*** value in this case. We shall reject the ***null hypothesis*** that there is ***no significant autocorrelation*** between environmental factors of an ecosystem and the existence of a given plant species, if the value of the test statistics **q** is less that the significant confident level which we have chosen as ….

This means:

Reject $\mathbf{H_0}$ if *d* is less than *d* (<) (dL)

Accept $\mathbf{H_0}$ if *d* is greater than (>) (dU)

Otherwise result will remain inconsistent.

Table 10: summary statistics of proposed model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.515[a]	.265	.186	2.100	.696

a. Predictors: (Constant), ELEVATION, TEMPERATURE, PRECIPITATION

b. Dependent Variable: ID

Our model accounts for 26.5% of the variance in species

If we apply the model to similar situation then the predicting power reduces from 26.5 to 18.6%

R = 69.6 shows that there is a significant autocorrelation among the predictor variables

It is very obvious from the value of the regression analysis above that each value of the I^{th} coefficient is significantly different from the "Expected $I'' = 0.00323$

Where the value of the I^{th} *coefficient = 0.515.* Thus this shows that Plant species types around the three parallel river park are significantly auto-correlated with the environmental factors/predictors of that eco-site.

Starting with

k = 3 (number of predictors)

D (***Durbin Watson test statistic***) = 0.696

N = 32

α =0.01

dL = 1.01, dU = 1.42

Error rate = 1% type 1 error rate

The p value of 0.01 is a good line of demarcation for us to make a judgment. Confidence = 1-p = 1 – 0.01 = 0.99. So we have a 99% confidence that we are making the correct decision.

Note: the type of test we carried out was the ***Durbin Watson test statistic for spatial auto-correlation*** with our hypothesis as stated above, the significance level we have taken as 0.01 giving our level of significance at 99%. So we reject the null hypothesis that there is no significant auto-correlation between the ecological factors f an ecosystem and the plant species present in the region. We computed sample test statistic and come up with the value 0.696.then we choose and computed our **p** value to be 0.01 and then the model was developed based on the decision made in the outcome.

6.2.7 Building the prediction model

Hypothesis:

H_0: - Plant species types around the three parallel river parks are not significantly auto-correlated with the environmental factors/predictors of that eco-site.

H_1: - Plant species types around the three parallel river park are significantly auto-correlated with the environmental factors/predictors of that eco-site; this means that there is a significant spatial autocorrelation. Thus the value of the *I* coefficient would be significantly different from E (*I*) which is equal to $-(n-1)^{-1}$; which is approximately zero (**Legendre and Fortin, 1989**).

Also from **champion et al. (1998)**, then if we take the positions **t** of the species in a sample of **n** observations, i.e **n** = 32 in our own case, we can then obtain a series in the form:

$[t, y(t)]$; $t = 1,2,3\ldots\ldots n$ ……**eqn. (2)**

Where y(t) is a function based on t position.

If we superimpose eqn. (2) into a regression model, we would have that:

$y(t) = g_{(i)}t + e_t$ for $i \geq 0$, $t \geq 1$……**eqn. (3)**

Where e_t is the expected value of the i^{th} coefficient, which is equal to $-(n-1)^{-1}$

Then for series i = 0…k, **k** = 3 (three (3) predictor variables) and i = each value of k at t position, t = 1…. n and **g** is the correlation coefficient (**R**) of the predictor variables of we would then have:

$y(t) = g_{(0)}t + g_{(1)}t + g_{(2)}t + g_{(3)}t + e_t$ for $i \geq 0$…… **eqn. (4)**

This fits into the general regression formula **y = gx +a…….eqn. (5)**

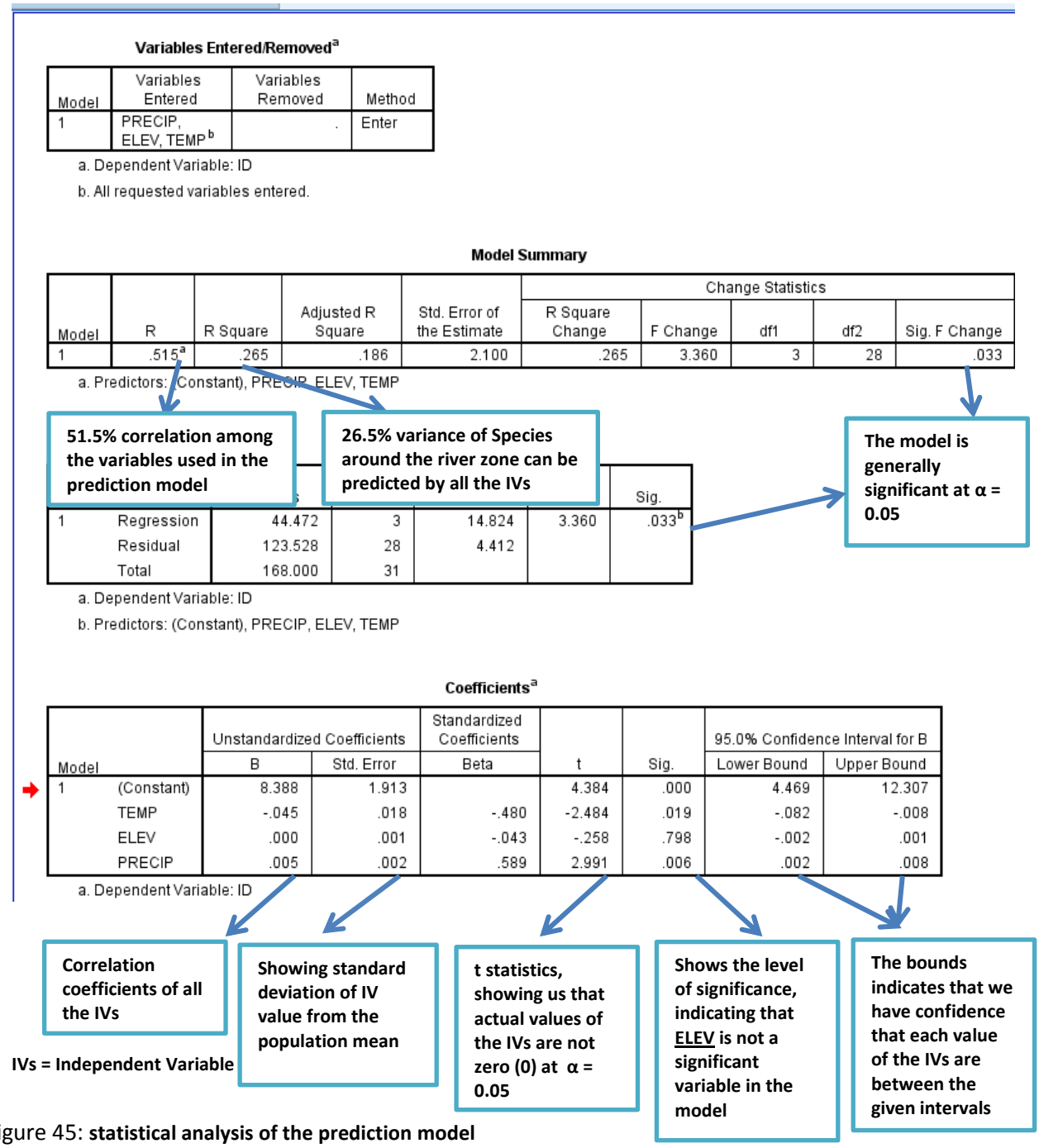
➟ Here **y** = predicted (response) variable,

g = slope of line – which in this case is the correlation coefficient of ith value at position t

x = known variable (predictor),

a = y intercept of the linear regression line

Using statistical analysis method, we were able to test our hypotheses statistically by three standard methods; anova, correlation and regression and the result is shown below.



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	PRECIP, ELEV, TEMP[b]	.	Enter

a. Dependent Variable: ID

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.515[a]	.265	.186	2.100	.265	3.360	3	28	.033

a. Predictors: (Constant), PRECIP, ELEV, TEMP

51.5% correlation among the variables used in the prediction model

26.5% variance of Species around the river zone can be predicted by all the IVs

The model is generally significant at α = 0.05

						Sig.
1	Regression	44.472	3	14.824	3.360	.033[b]
	Residual	123.528	28	4.412		
	Total	168.000	31			

a. Dependent Variable: ID

b. Predictors: (Constant), PRECIP, ELEV, TEMP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	8.388	1.913		4.384	.000	4.469	12.307
	TEMP	-.045	.018	-.480	-2.484	.019	-.082	-.008
	ELEV	.000	.001	-.043	-.258	.798	-.002	.001
	PRECIP	.005	.002	.589	2.991	.006	.002	.008

a. Dependent Variable: ID

Correlation coefficients of all the IVs

Showing standard deviation of IV value from the population mean

t statistics, showing us that actual values of the IVs are not zero (0) at α = 0.05

Shows the level of significance, indicating that ELEV is not a significant variable in the model

The bounds indicates that we have confidence that each value of the IVs are between the given intervals

IVs = Independent Variable

Figure 45: **statistical analysis of the prediction model**

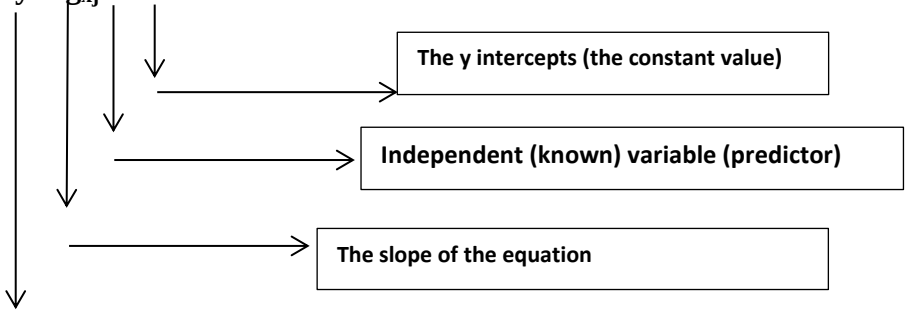
The regression analysis has indicated that the ***elevation (though not zero from the t- statistic)*** of the three parallel river geographical locations does not have a significant relationship with the existence of plant species around the three parallel rivers, as we can see in figure 39 above.

Table 13: the correlation between precipitation, temperature and species

Statistics	**Elevation (R -coef)**	**Precipitation (R- coef)**	**Temperature (R - coef)**
Coefficient	**0.00001**	**0.005**	**-0.045**
Confidence Level	**95%**	**99%**	**95%**
Significance	**0.05**	**0.01**	**0.05**

Applying the value of the unstandardized coefficients to our own regression equation, we obtain the new values for equation (**5**) –

y = g_{xj} t +a becomes**:**



Dependent variable (predicted)

6.2.8 Model interpretation

<u>The raw equation</u> (shows the effects of the predictors to the predicted variable but does not show which is stronger):

Y (species) = bx + a

- ⇨ *Species (id) = 0.00001 elevation + 0.005 precipitation – 0.045 temperature + 8.388*

*The standardized equation shows which of the IVs has the strongest effect on the species = **Precipitation***

Y (species - id) = βx

oR

Y (species) = bx1 +bx2+bx3….

⇨ *Species = .000 elevations + 0.005 precipitations – 0.045 temperature……*

The model is generally significant at α =0.05 with a 95% confident level and F = 0.033 (remember according to the explanations in section 3.1.2, the poor significance rate of 0.033 is as a result of the problems posed by the self-dependent nature of spatial data which gives rise to autocorrelation and thus making the effective sample size less than the number of observations – especially because F is positive).

We shall accept the alternative hypothesis (**with type 1 error**) that the environmental factors of an ecosystem are significantly auto-correlated with the species around the ecosystem. This is so because all the values of the I^{th} coefficient were significantly different from E (*I*) which is equal to – $(n-1)^{-1}$; which is approximately zero.

We therefore conclude that the plant species types around the three parallel river park (using the **Lacang zone**) are significantly dependent on some environmental factors/predictors (such as temperature and precipitation – we saw that elevation on the areas around that site has no significant influence on the species) of that eco-site; this means that there is a significant spatial autocorrelation among the independent variables (IVs) and the dependent variables (DVs).

Based on the above conclusion our final model for the prediction of plant species around the three parallel rivers will be stated as:

⇨ ***Species (id) = 0.00001 elevation + 0.005 precipitation – 0.045 temperature + 8.388 *dist to the water * dist to neighbour***

Chapter 7:

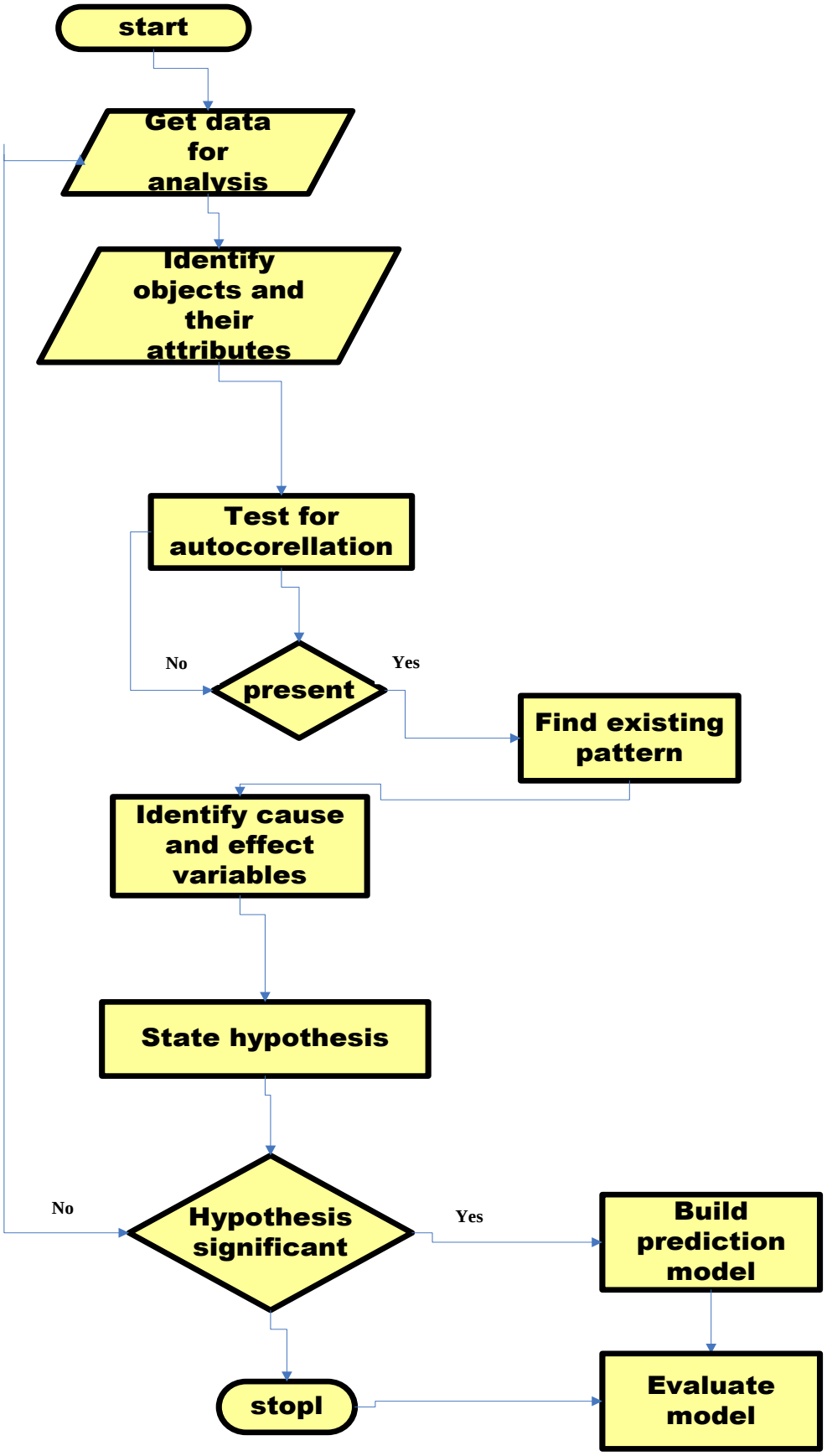
System Design

7.1 Basic Algorithm for Mining a Complex Spatial Dataset

The 8 steps below are suggestions of basic steps to mining spatial data; these would be improved on in further research:

1. Get data from survey, observation or digitized map layer
2. Identify all attribute data present in all objects
3. Test for auto-correlation; using moran I for single variables and statistical packages for multivariate data
4. If autocorrelation is present, identify the nature of pattern that exist
5. Using appropriate tool derive variables that are most likely the event predictors
6. Test for cause and effect impacts
7. Derive the prediction model by explaining some **events** occurrence through analysis and exploration of data
    (a) form a set of h**ypothesis** about these variables which are likely to cause these events
    (b) test statistical significance of the hypothesis
    (c) model more precisely, any quantitative nature of the relationship that may exist using linear regression or any other tool for multivariate data.
8. Evaluate your model.

7.2 Process model for mining data in a spatial dataset in a programming context



7.3 System Design based on Data Mining Methods

In a brief description, what we have achieved in this process could be easily described using a fish bone diagram ***(figure 43 below***) which shows a **cause** and **effect** scenario (***Category 1 and 2***) of the factors that can contribute to the existence of a plant species in any eco-habitat:

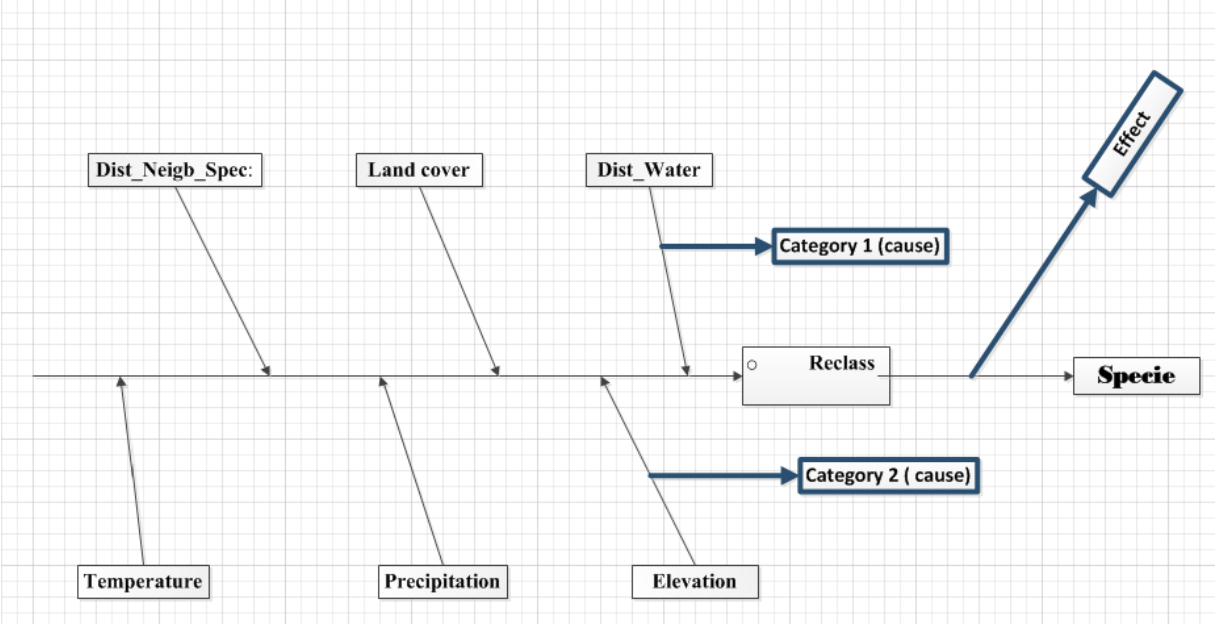
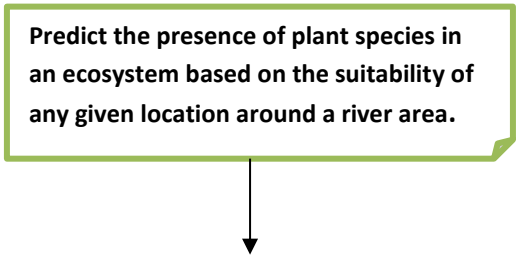


Figure 46: **Fish bone diagram showing cause and effect**

Category 1 causes as shown in the diagram was used for the data mining technique based analyses (for the purpose of prediction), while the section in **category 2** was used for statistical analyses (for finding patterns and relationships and also for determining the presence of absence of autocorrelation).

7.4 THE CONCEPTUAL MODEL:

Step 1: Stating the problem



Step 2: Breaking the problem down

Measurement:- what is the best point for a particular species to grow, would it preferable breed near or far away from the river, what degree of slope would be more suitable, what intensity of solar radiation and quantity of precipitation will determine viability or otherwise. Other objectives that could be included in this example could be plants existing in an area with highest density of similar species. In addition, consideration should also be made of areas where plants cannot breed like un-vegetated areas, rivers and built up parcels.

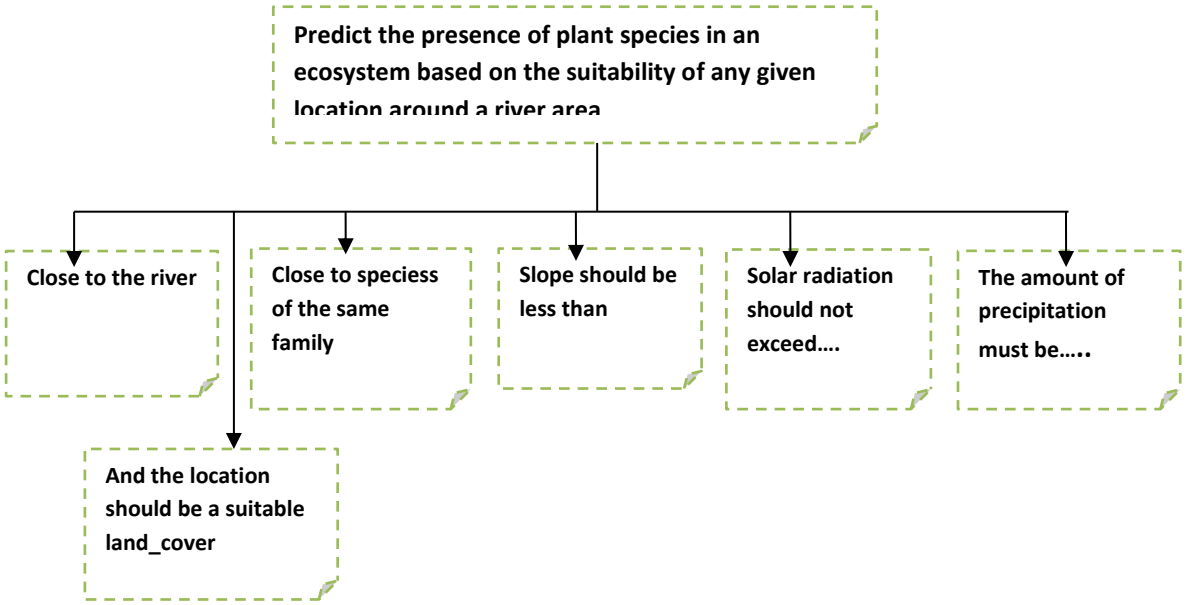


Figure 47: using a conceptual model to understand the sub-systems involved

Step 3: Defining classes and process, and Computation of variables

Where are locations lower slope

- Input the **elevation** data set

Does the land cover in the above locations suitable for plant breeding?

- **Input the** land cover **data set (but you need to consider land cover suitability factor e.g forest, pastures, agricultural and river areas would do better than un-vegetated and built-up areas).**

Are they (the locations) near enough to the location of other family members

- **Input** location **dataset for the** nearest species neighbour

Are they (the locations) near enough to the river area?

- **Input the** location **of the** river (and create a buffer for restriction)

Is there enough sunlight for maximal radiation?

- Input **solar radiation**

Is there enough sunlight for maximal radiation?

- Input **watershed**

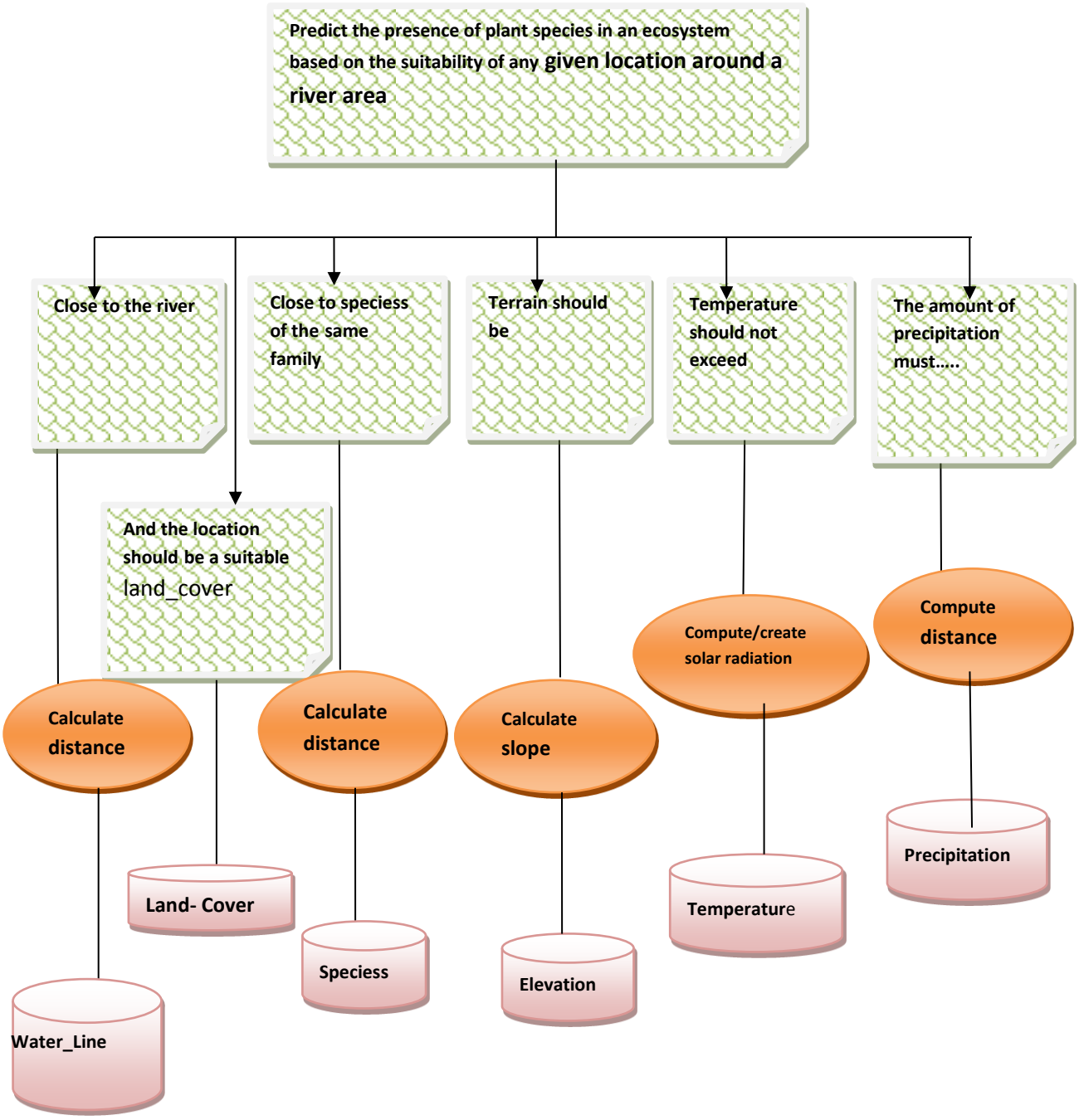


Figure 48: **A conceptual model to define classes and process, and computation of variables**

Step 4: Describing the Dataset:

- **Elevation**: - Dataset representing the elevation of the area
- **Land cover**: - Dataset representing the land-cover types over the area
- **Dist_Water:** - Feature class representing the of river network which was calculated by creating a 10 meter buffer zone around the **water_line** feature)
- **Dist_Neigb_Spec**: - Feature class representing point locations of species sample sites (basically three types chosen for our). This is used to show the nearest species neighbour within the shortest distance and it was calculated using the **euclidean distance** tool in arcgis **spatial analyst** extension.
- **Temperature**: - Feature class representing sampled point locations sites for temperature data collection
- **Precipitation**: - Feature class representing sampled point locations sites for precipitation data collection
- **Reclass**

Step 5: Reclassification:

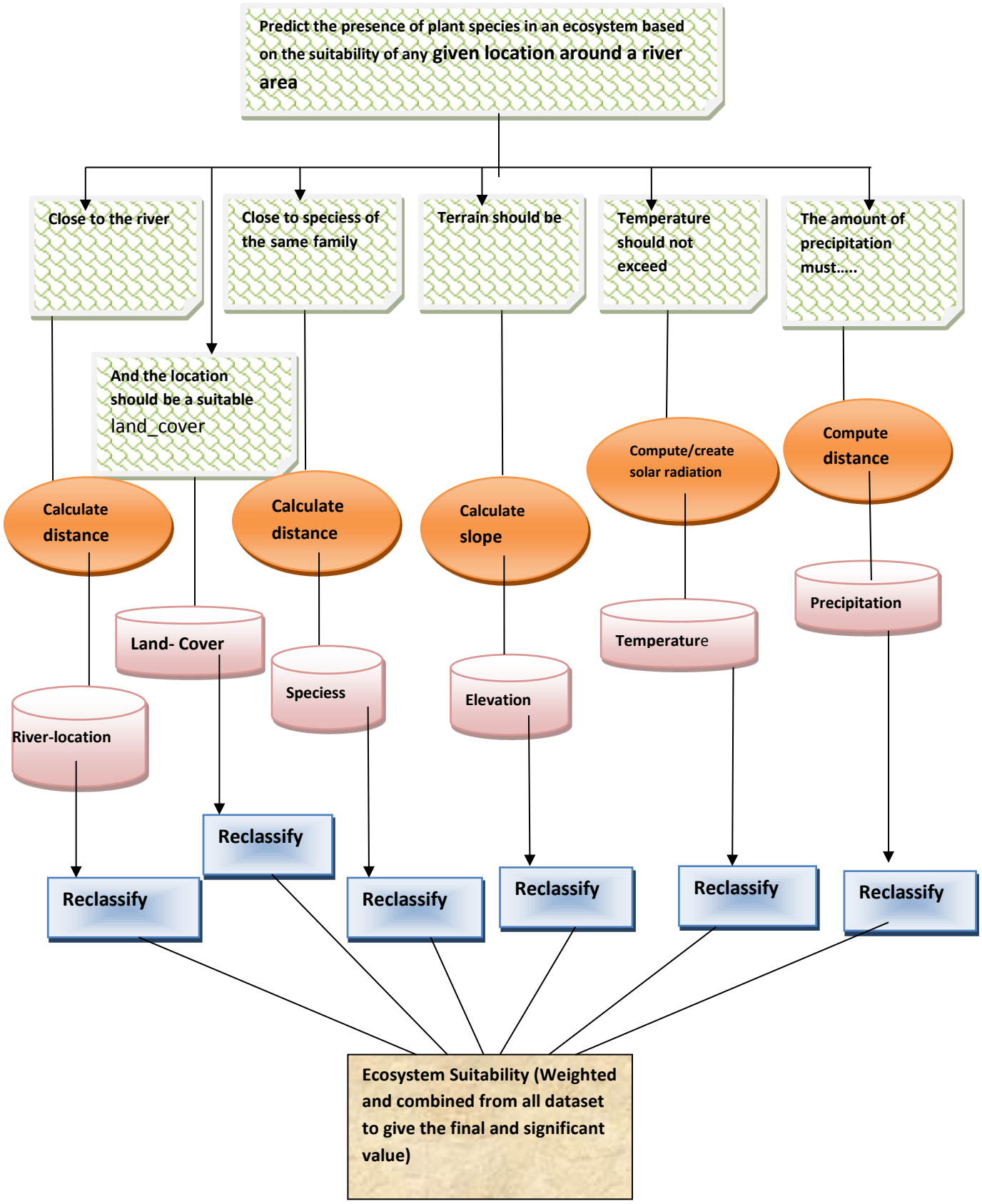


Figure 49: **A conceptual model to showing reclassified classes**

7.2.1 Reclassification explained:

The process of reclassification applied in spatial data analysis involves weighing or grouping of values based on certain criteria; this can be done going one value at a time or groups of values at once. These criteria could be specified intervals or specified areas. The functions used for reclassification are designed to allow you to easily change many values on an input raster to desired, specified, or alternative values.

7.3 CLASS DIAGRAM

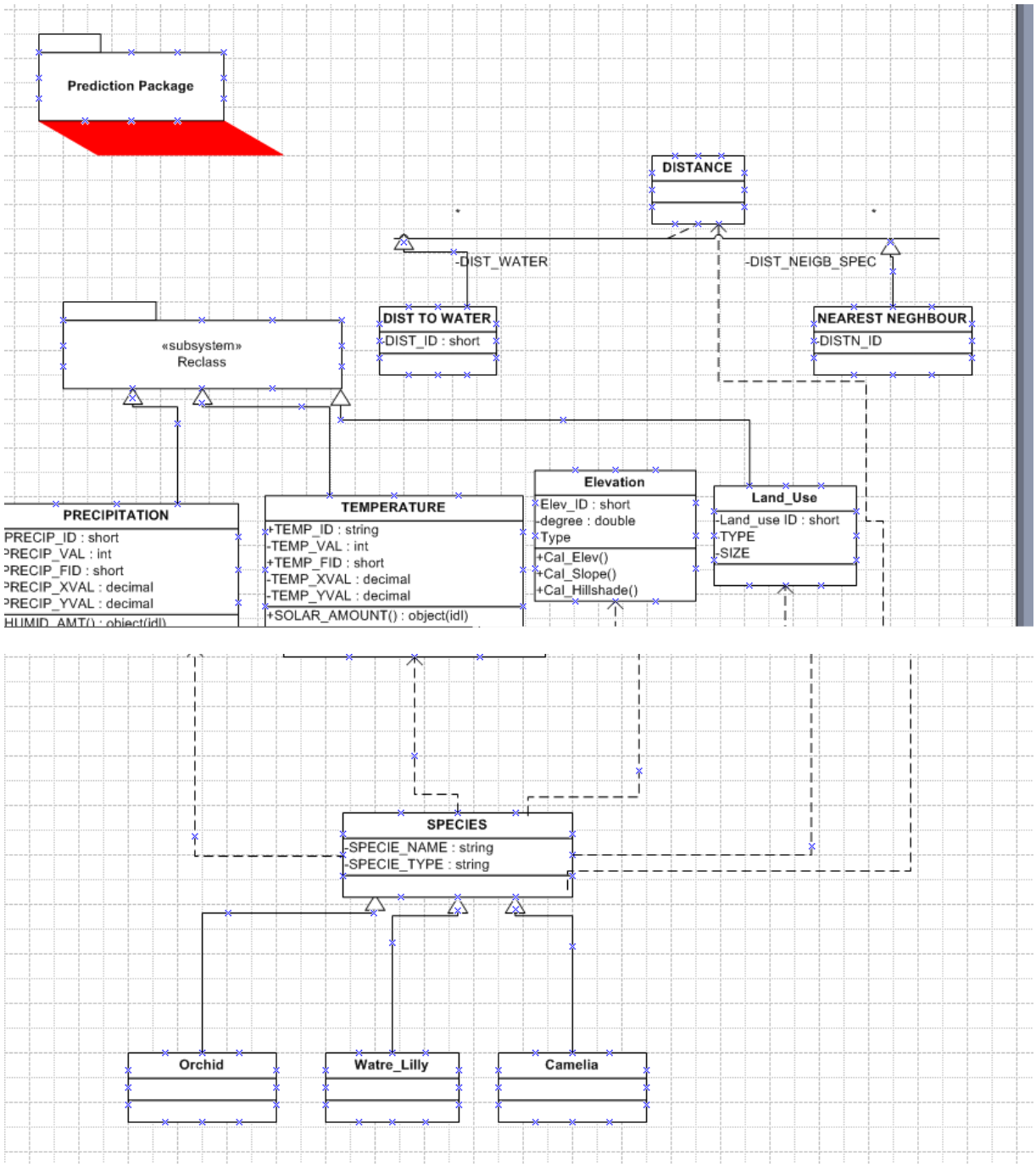


Figure 50: **class diagram showing classes and their attributes**

7. 5 Activity Diagram (in context)

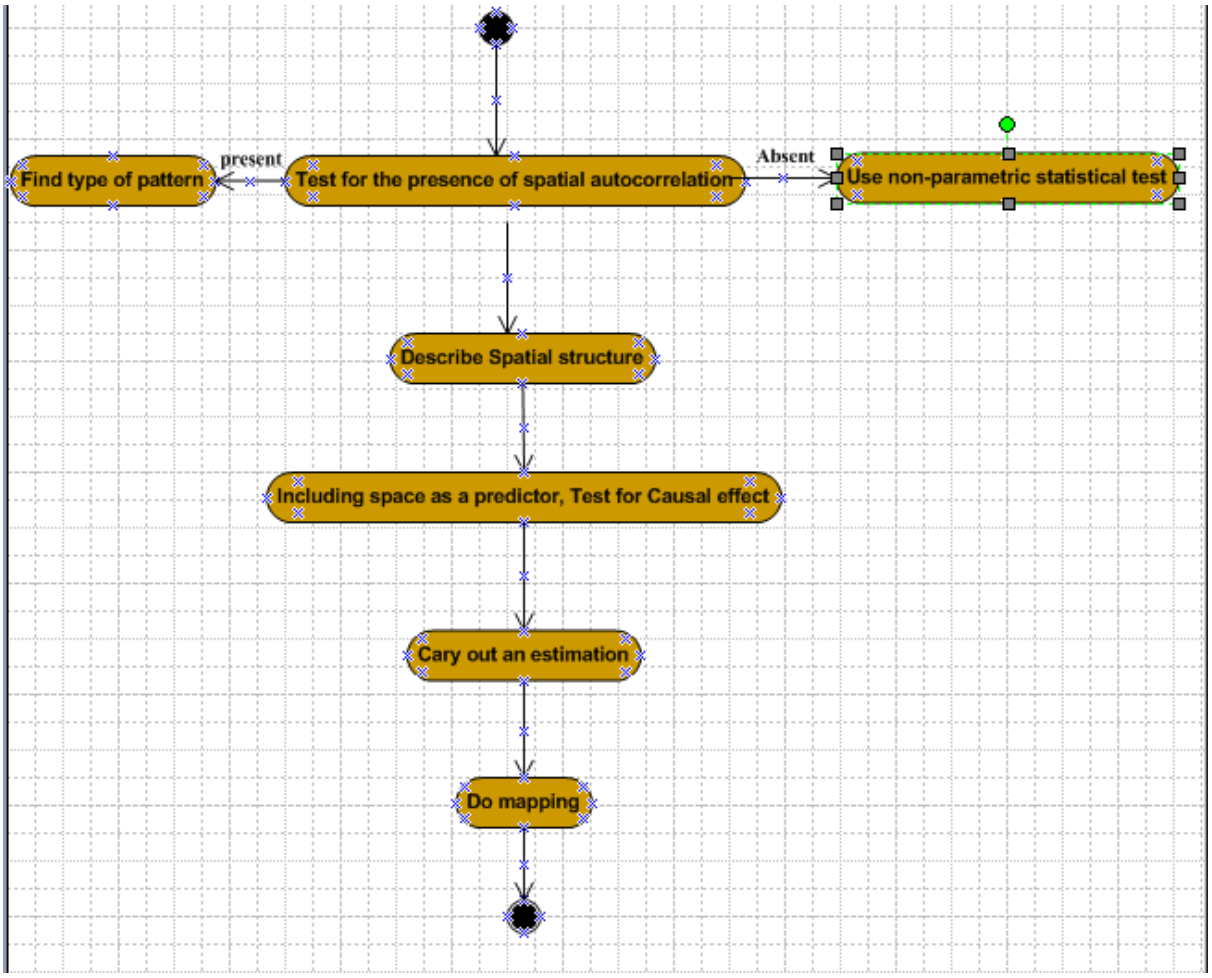


Figure 51: Use Case Diagram for Programming Aspect

Chapter 8:

Professional and Ethical issues in scientific project development

Professional and ethical issues are very important for the success of a project, and it is always a good practice to make sure that projects are done within the law without causing harm to the users and the environment in which it would be used

8.1 Ethical issues

Adu-Gyamfi and Okech (2010) stated that most scientific domains have guidelines concerning ethics in research. These guidelines are designed to enable researchers to conduct good research while avoiding potential harm to research participants. **Leedy and Ormrod (2005)** summarize ethical issues in research in four (4) main categories considering the major stake holders of any form of research (***Human and Data***):

(a) ***Protection from harm*** – participant must not be exposed to harm

(b) ***Informed consent*** – participant must know the nature of the research

(c) ***Right to privacy*** – participant right of privacy must be protected

(d) ***Honesty with professional colleagues*** – researcher must be honest in their report

8.2 Professional issues

Professional issues in computing and information has to do with the fact that information systems, software or computing based project has to conform to the codes of conduct of relevant legislations and regulations, making sure that systems e.g projects are written professionally.

Some of these codes of conduct include the standard regulations set by:

- British Computer Society (BCS)
- Association for Computing Machinery (ACM)
- The Institute of Electrical and Electronics Engineers (IEEE)

In writing the research project (based on the nature of study that we are undertaking), we have ensured that there was no clash of interest created by our mode of data selection especially the data from the ***study area,*** which has been handled with maximum integrity and confidentiality. Furthermore, based on the four point BCS standards for professional writing in information technology as can be seen in (**BCS 2011**) code of conduct report, we have also ensured that the data for the purpose of this research was only used to achieve the scientific research objectives.

Some other codes of good practice as enabled by BCS professional ethics incudes the fact that the project report would be:

- Written in clear brief English without any slangs.
- Presented in contents page and appendices.
- Presented in a clear layout, in line with module handbook regulations.
- Referenced in Harvard style.
- Presented with limited grammatical or spelling irregularities.

Chapter 9

Evaluation of product by self and by stake-holders

Evaluation is a way of testing effectiveness of a developed system. This will help us to analyse the developed model and then make any necessary adjustment.

9.1 Evaluation of product by self

9.1.1 Statistical evaluation

Based on statistical inference, the first evaluation of our model in predicting the plant species around the three parallel river Lacang zone was done based on statistical significance and the predicting power of our model. This is shown in the table below where the highest error is 4. This prediction is based on the actual values as we have in the table (11) at the beginning of this model building;

Spec	id	Elev	Temp	Prec	pred. y	error
CAM1	1	900	90	0	4	-3
CAM1	1	1200	100	7	4	-3
CAM1	1	1600	108	21	4	-3
CAM1	1	2000	130	307	4	-3
WATER_LI1	2	500	80	12	5	-3
WATER_LI1	2	600	111	13	3	-1
WATER_LI1	2	800	130	133	3	-1
WATER_LI1	2	900	102	55	4	-2
CAM 2	3	700	99	71	4	-1
CAM 2	3	790	100	205	5	-2
CAM 2	3	850	144	502	4	-1
CAM 2	3	1100	142	330	4	-1
CAM 3	4	200	93	0	4	0
CAM 3	4	300	115	98	4	0
CAM 3	4	500	141	171	3	1
CAM 3	4	700	165	793	5	-1
WATER_LI2	5	120	77	17	5	0
WATER_LI2	5	180	68	228	6	-1
	5	300	116	212	4	1

WATER_LI2						
WATER_LI2	5	400	113	453	6	-1
ORCHID	6	200	93	36	4	2
ORCHID	6	400	105	228	5	1
ORCHID	6	600	130	679	6	0
ORCHID	6	800	145	190	3	3
ORCHID 2	7	600	93	40	4	3
ORCHID 2	7	900	103	882	8	-1
ORCHID 2	7	1000	120	46	3	4
ORCHID 2	7	1200	151	899	6	1
WATER_LI3	8	1000	67	7	5	3
WATER_LI3	8	1300	84	62	5	3
WATER_LI3	8	1800	81	184	6	2
WATER_LI3	8	2000	105	734	7	1

The values with the _____ line has been predicted correctly, which means given another 4 sample of a same species with varying ecological variables like temperature, precipitation and elevation (although the elevation has little or no significance in the prediction), there is a 26. 5% guaranty that the IVs can predict the value of the species as has been indicated by our model.

The high error rate of 4 accounts for the outliers which we can identify from the diagram in figure 35 through 37.

Generally at 95% confidence level, our model is model is significant.

9.1.2 Evaluation using non-parametric bootstrapping

Another way we could evaluate the efficiency of the model developed in this project, is based on the ideas suggested by **Todem et al. (2010),** here since they acknowledge that, because in most prediction models, some of the model characteristics (which may have most perverse effect) are always non-identifiable from observed data, thus one best approach to evaluate the statistical hypothesis is to fix a minimal set of sensitivity parameters conditional upon which hypothesized parameters are identifiable. They believe that in most multivariate statistical modelling, there is always likelihood that outliers are never ignorable when evaluating covariate effect on the model's behaviour or performance. The bootstrapping tool is an evaluation tool that is basically used when the normal traditional assumptions are violated, to accurately test or adjust the model. This evaluation at this stage is beyond the reach of this project scope (because of time constraint) but will be considered if further research works

9.2 Evaluation of product by stakeholder

The major stake holder in context for this project is the project supervisor. In evaluating the product of this project work, the supervisor used the heuristic method of information system projects evaluation. This mode of project evaluation entails the process of finding satisfactory solution using intuitive judgement, educated guess or common sense. **Judea (1983)** also highlighted that heuristics strategies uses readily accessible but loosely applicable information to establish problem solving inhuman beings.

In other words the supervisor's evaluation of the project work was based on already known features of prediction models, which helps to identify some of the main key item that must be present in a prototype model of this nature based on the fact that the project work is a research framework for a more advanced research on mining complex systems using spatial data mining techniques.

Chapter 10

CONCLUSION

The project work was based on mining of complex spatial databases. After conducting a deep, thorough and reflective research, we were able to come up with a prototype model for predicting the plant species around an ecosystem based on the features and influences of some predictor variables. Our research output is the development of spatial mining methodologies, basic algorithm and tools that can address the following problems:

- ❖ Regional patterns discovery – Interesting places and their associated patterns (take our case study for instance)
- ❖ Spatial clustering and outlier effects in a spatial data
- ❖ Co-location and correlation mining
- ❖ Mining predictions for complex spatial systems.

Spatial data mining is a branch of data mining where space and location of object is an important factor. In this advanced research based project, we have carried out an extensive research on the field of data mining and we have managed to develop a framework for spatial data mining which is suitable for further expansion and research. We looked at the various branches and tools for data mining and we had a detailed study of spatial data mining; tools techniques, methods, and tasks. We also looked at the various application areas of spatial data mining and the nature of specific pattern that could exist in a given spatial dataset. Using a two stages methodology, we developed a prototype generalised prediction model, the first stage was the analysis of the spatial data based on spatial analysis tools such as **distance, overlay, con, isnull e.t.c** this spatial analysis produced a prediction map showing parts of the areas around the river zone which are suitable for plant species to grow and other areas which are not suitable,; this was derived using a prediction model as we can see in figure **31 through 34 in chapter 6 above.**

The other analysis we did was based on statistical implication, this was basically used to test for autocorrelation and to find pattern. Through that analysis we deduced that the elevation of any geographical location does not have any significant impact on the plant species around that location, it was also deduced that the precipitation around that area has the greatest impact on the

plant species whereby increase in precipitation always yield a positive impact on the species whereas lower temperatures are more preferable.

Finally, we evaluated our model using statistical inference and bootstrapping and our model proofed approximately 27% effective in predicting the existence of a plant species around the three parallel rivers (**Lacang zone**)

FUTURE WORK**:**

Despite the time constraint of this project, we have tried to produce a prototype model for the prediction of a complex spatial dataset, thus we plan to carry out much detailed and complex analysis of various models of real-world problems as a future contribution to this work. Basically the main investigation to be done on this study area, is the development of ***novel, genetic and generic algorithms*** for complex spatial data mining; this would make the spatial data mining field very versatile as it would possess the capability of solving problems from various range of field including analysing human related complex processes such as applying the study to the building and modelling of ***human cognitive ability, human – environmental physiognomies and many other***.

More practically, our major future plan is to develop a generalised model for spatial pattern mining capable of analysing data from a complex spatial system and then produce information that would be useful in various disciplines where spatial data form the basis of general interest. In essence, our main aim is to be able to create solutions for the following projects based on our developed algorithm:

- Modelling the spatio-temporal patterns of human cognition
- Spatial analysis of natural disturbances: the effect of air pollution on water or weather condition
- Modelling and simulation of inflation rate and control: a case of the Nigerian economy
- Simulating the effect of Climate on the culture of a people

More so, one of the problems encountered in this analysis is a way to solve the problem of poor measurement of goodness-of –fit caused by the self-dependent nature of spatial data, as such in our future research, this will be one of the main tasks in our future research.

APPENDICES:

Appendix 1: Proposal

DESCIPTION OF PROJECT

To: Prof. Lu
From: Grace Samson (U1251405 – MSC Computer Science)
Date: 5th June 2012
Subject: Research proposal

Proposed Research Topic

An investigation in Efficient Spatial Patterns Mining: Mining Complex Spatial data

Key words:
Complex Spatial Systems (CSS), Predictive modelling/ Knowledge Management, Spatial Patterns Mining, Querying CSS, Mathematical CSS Modelling, Event Prediction.

Purpose

To develop a generalised model for spatial pattern mining capable of analysing data from a complex spatial system and then produce information that would be useful in various disciplines where spatial data form the basis of general interest. As acknowledged by Wilson (2002), complex spatial systems are defined as those systems described by many variables, with high levels of interdependence between elements, governed by non-linear processes and having significant spatial structures. One would have noticed that the major challenge in trying to build a general complex spatial system model would be; to be able to integrate the elements of these complex systems in a way that is optimally effective in any particular case.

As highlighted by Shekhar et al (2005), the explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. This is what motivates our research interest. Although there some general purpose data mining tools such as Clementine and Enterprise Miner which are designed to analyse large commercial databases according to them, general purpose tools for spatial data mining (especially in the case of a complex spatial data) need also to be develop because extracting interesting and useful patterns from spatial data sets is more difficult than the patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. As a result, we seek to develop a predictive model output patterns for spatial data mining.

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as, droughts, vegetation diseases, and earthquakes.

We seek to create an explicit spatial model for event prediction using basic spatial data mining algorithms and not any of the general purpose data mining algorithms. In essence we aim to look at modelling (predictive modelling/ knowledge management of complex spatial systems), querying and implementing a complex spatial database (using data structure and algorithms).

Background

There are basically three types of complex systems as noted by Weaver (1948, 1958). These include the simple, the organised complexity and the disorganised complexity systems respectively. For the purpose and scope of our research work we are going to be considering the organised complex system and then move further into the ***disorganised complex system and complex adaptive systems*** in further research works. Organised complex systems are described by many variables, and all variables have strong interdependencies. *Human beings, brains, economies, cities, ecosystems, and*

*language all provide examples of organised complex systems. Organised systems are also characterised by the presence of nonlinearities (*consider figure 1.0)

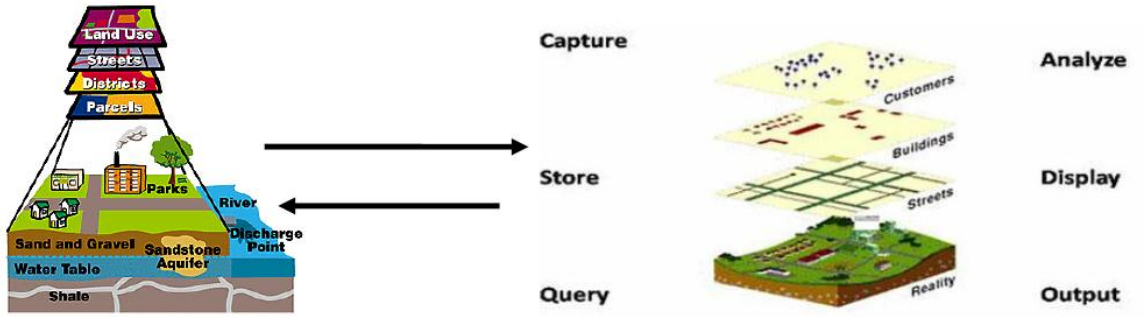


Fig 1.0: A diagrammatic view of different spatial data layer and data from such a system is handled for knowledge management in a complex organised system.

Because complex spatial systems are those with significant spatial structures, we shall concern ourselves majorly with ***three main tasks***:
1. We shall investigate, examining and analyse the different range of disciplines where complex spatial data has a significant function and then we shall link some of these functions to the development of complexity theory.

2. We shall investigate existing mathematical models for modelling these kinds of systems and then try to develop a better analytical/predictive model for such complex systems, by implementing (using data structure and algorithms) and querying a given complex spatial database.

3. We shall then take an example from one of the various existing models (e.g. Urban) which form an important antecedent of the programme of building general models of complex systems within the field of complexity theory.
4. Finally we shall make a conclusion based on what we have discovered from the proposed extensive study of complex spatial systems.

However, the four (4) basic areas of interest of spatial data mining as listed below, would be a major term of reference to what we intend to achieve (and our major task would be to develop an algorithm for some of these tasks and then identify their application areas).
1. Predictive modelling/ Knowledge Management (for event prediction)
2. Spatial outlier detection

3. Spatial co-location rule/patterns mining
4. Spatial clustering.
5. Spatial trend
6. Spatial classification

Theoretical framework:
We shall be guided most generally by the interpretive perspective, and more specifically by Wilson's (2002) '*Models of cities'* approach. The interpretive perspective places the focus on interpreting the meanings and perspectives of modelling a complex spatial system based on three major range of method/approaches; Non – linear mathematics, Computer simulation and Visualisation of result. Exploration of the meanings of cities and region as complex spatial systems, identification of elements of a complex spatial system, conventional models for CSS and relationships that could exist between elements of CSS as well as investigating a conceptual frame work shall be based on the work of Wilson (2000). Shekhar et al (2005) will guide us through the basic components, elements, techniques and features of a spatial pattern mining task; as this forms the basis of our argument. Some of these issues include
1. Spatial data types (points, lines and polygons/regions)

2. Spatial attributes (latitudes, longitudes shapes…….)

3. Implicit spatial relationships among variables

4. Observations that are not independent, and

Method
1. we shall start by Conducting an intensive literature review on Data mining, Spatial data mining, Complex systems, Spatial patterns mining, Complex spatial system, Methods for modelling and querying a spatial database, Predictive modelling/ Knowledge Management and Creating mathematical models for computer simulation.
2. Then we would undertake a thorough analysis of existing models and algorithms for predictive modelling/ knowledge management of CSS.
3. We shall then try to work out a new algorithm for modelling a complex spatial system
4. Design a simulated system for complex spatial system prediction
5. Design a system for visualization of event prediction result using Java programming language
6. Write a research report that combines our understanding of the relevant theory and previous research with the results of our empirical research and
7. **Finally, we hope to be able to come with a system for capturing, storing, checking, integrating, manipulating, analysing and displaying data which are spatially referenced to the Earth**

Limitations:
Time constraints of the semester require less time than may be ideal for the study of complex interactions among elements of a complex spatial system. And also being an early scholar in the field of data/spatial data mining may also limit the expected output of this research process (though we would give in all our best to achieve an effective system).

Conclusion

In summary, spatial data mining organises by location what is interesting thus, the major issue around its study include:

1. Analysing spatial autocorrelation
2. The fact that space is continuous
3. Existence of complex spatial data types
4. The need for regional knowledge
5. Availability of large dataset and many possible patterns
6. The importance of map as summaries e.t.c……..

For the sake of argument, we would try to see how much of these issues listed above that we could handle. And this leads to the conclusion that our intended research output would be the development of spatial mining methodologies, algorithm and tools that can address the following problems:

1. Regional patterns discovery – Interesting places and their associated patterns in spatial dataset.
2. Spatial clustering algorithm with some fitness function
3. Co-location and correlation mining
4. Change analysis in spatial dataset and
5. Mining predictions for complex spatial systems.

References

Weaver, W. (1948) 'Science and complexity', *American Scientist* 36, pg. 536-544.
Weaver, W. (1958) A quarter century in the natural sciences, Annual Report, The Rockefeller Foundation, New York.
Wilson, A. G. (2002) 'Complex Spatial Systems: Challenge for the Modeller', *Mathematical and Computer Modelling*. 36, pg 379 – 387
Wilson, A. G. (2000) *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Pearson Education
Shekhar, S., Zhang, P. & Huang, Y. 2005, "Spatial Data Mining" in Springer US, Boston, MA, pp. 833-851.

Appendix 2: Project Timespan (derived from methodology)

Project Timescales

1st May – 29th May: Project discussion with prospective supervisor including proposal; writing. (4wks)

1st June – 29th June: Collecting facts and reviewing literature. (4wks)

2nd July: **milestone 1 <Report from literature>.**

4th July – 31st July: Theory building and model derivation. (4wks)

3rd August: **milestone 2 <working theory>**

6th August – 10th September: Analysing cause and effect factors. (5wks)

11th September: **milestone 3 <working causal theory>.**

13th September - 11th October: Development of prediction model. (4wks)

15th October: **milestone 4 <analysed prediction variable>.**

19th October – 18th November: report writing, Project review, presentation preparation (4wks)

20th November – 28th November: Project review with supervisor. (1wk)

30th November: **milestone 5 <project submission>.**

14th December: **milestone 6 <project presentation>.**

Appendix 3: Project Time-plan diagram (derived from methodology)

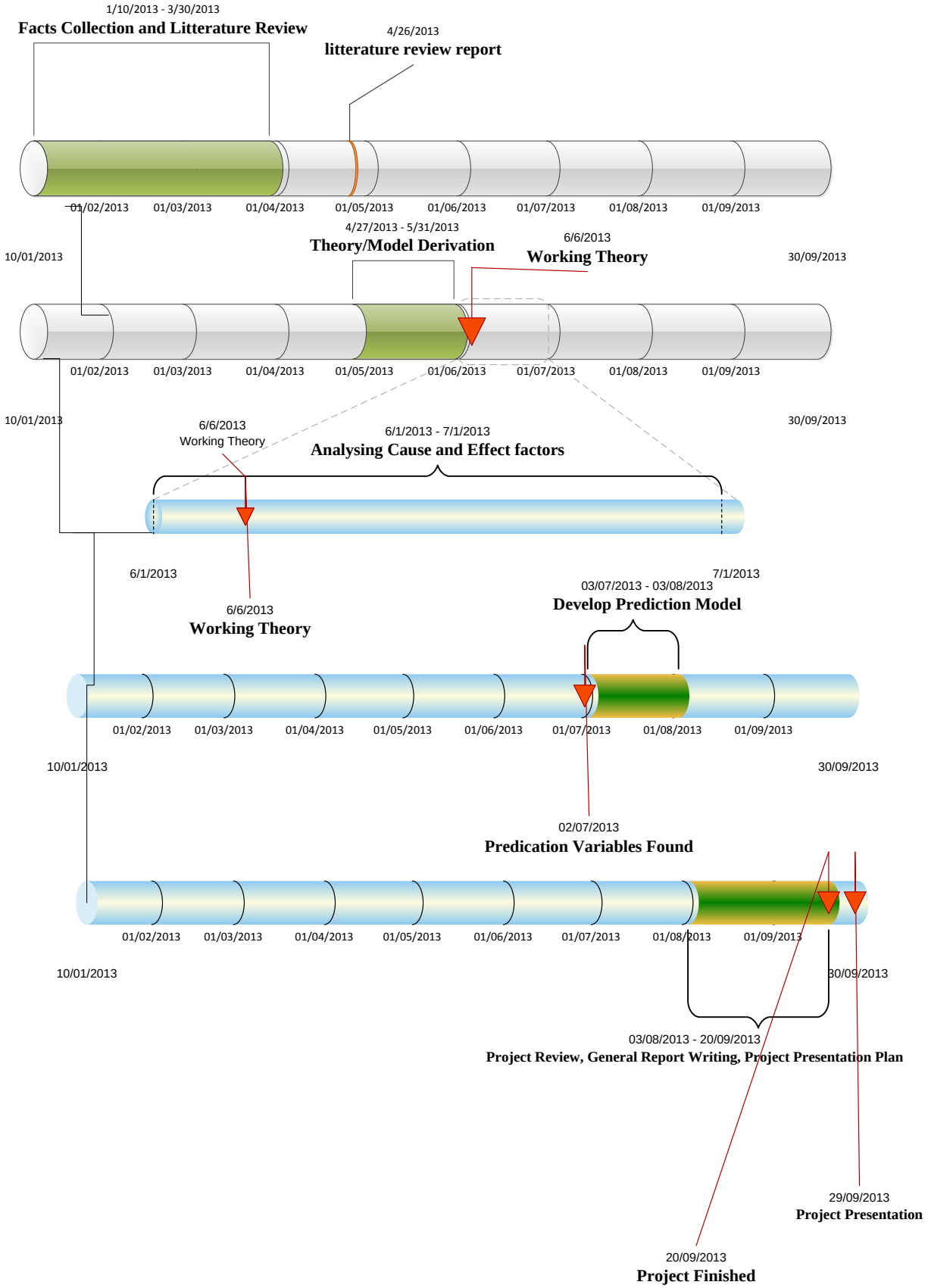


Figure 52: Project timeline

References:

Adhikary, J. (1996) Knowledge Discovery in Spatial Databases Progress and Challenges. Canada: *School of Computing Science, Simon Fraser University*

Adu-Gyamfi, K. and Okech, A. (2010) "Ethics in Research in Mathematics Education", *Journal of Academic Ethics.* 8 (2), pp. 129-135

Agrawal, R., Imielinski, T., and Swami, A. (1993) "Mining association rules between sets of items in large databases." In ACM SIGMOD international conference on management of data. pp. 207–216

Anagnostopoulos, T., Anagnostopoulos, C. and Hadjiefthymiades, S. (2012) "Efficient Location Prediction in Mobile Cellular Networks", *International Journal of Wireless Information Networks.* 19 (2), pp. 97-111.

Andrienko, G. and Andrienko, N. (1999) Data Mining with C4.5 and Interactive Cartographic Visualization. In *Proceedings of the 1999 User Interfaces to Data Intensive Systems* (UIDIS '99). IEEE Computer Society, Washington, DC, USA, p.162.

Ayala, G., Epifanio, I., Simó, A. and Zapater, V. (2006) "Clustering of spatial point patterns", *Computational Statistics and Data Analysis.* 50 (4), pp. 1016-1032.

Babovic, V. (2000) "Data Mining and Knowledge Discovery in Sediment Transport", *Computer-Aided Civil and Infrastructure Engineering.* 15 (5), pp. 383-389.

Bailey-Kellogg, C., Ramakrishnan, N. and Marathe, M. (2006) "Spatial data mining to support pandemic preparedness", *ACM SIGKDD Explorations Newsletter.* 8 (1), pp. 80-82.

BCS. (2011) *British Computer Society.* [Online] Available at : http://www.bcs.org/upload/pdf/conduct.pdf > [Accessed 2 December 2012].

Bembenik, R. and Rybiński, H. (2009) "FARICS: a method of mining spatial association rules and collocations using clustering and Delaunay diagrams", *Journal of Intelligent Information Systems.* 33 (1), pp. 41-64.

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., and Ramanujam, K. (19997) "Brief Application Description Advanced Scout: Data Mining and Knowledge Discovery in NBA Data" *Data Mining and Knowledge Discovery*. 1, pp. 121–125

Bhramaramba, R., Allam, A.R., Kumar, V.V. and Sridhar, G.R. (2011) "Application of Data Mining Techniques on Diabetes Related Proteins", *International Journal of Diabetes in Developing Countries.* 31 (1), pp. 22-25.

Bio, A.M.F., De Becker, P., De Bie, E., Huybrechts, W. and Wassen, M. (2002), "Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions", *Biodiversity and Conservation.* 11 (12), pp. 2189-2216.

Bolstad, P. (2002). *GIS Foundamentals: A Fisrt Text on GIS*. Eider Press.

Booth, D. T., Cox S. E., and Berryman, R. D. (2006) "Point Sampling Digital Imagery with 'Samplepoint'" *Environmental Monitoring and Assessment* Springer. 123, pp. 97–108

Brendan, J. F. and Delbert, D. (2007) "Clustering by Passing Messages Between Data Points" *Science*. 315, pp. 972-976.

Brinkhoff, T., Kriegel, H.P. and Seeger, B. (1993) Efficient processing of spatial joins using R-Trees. *Proceedings of SIGMOD – 93* May.

Brown, D. G. (1994) "Predicting Vegetation Type at Treeline using Topography and Biophysical disturbance variables" vegetation science. 5, pp. 641-656

Chang, T. (2004) "Spatial Statistics", *Statistical Science.* 19 (4), pp. 624-635

Chen, T. and Chen, C. (2010) "Application of data mining to the spatial heterogeneity of foreclosed mortgages", *Expert Systems with Applications.* 37 (2), pp. 993-997.

Chen, J., Chen, Y., Yu J., and Yang, Z. (2011) "Comparisons with spatial autocorrelation and spatial association rule mining", IEEE, pp. 32.

Chidanand, A., Bing, Liu., Edwin, P. D., Pednault, and Padhraic, Smyth. (2002) "Business applications of data mining", *Association for Computing Machinery, Communications of the ACM.* 45 (8), pp. 49.

Chou, K. and Shen, H. (2007) "Recent progress in protein subcellular location prediction", *Analytical Biochemistry*, 370 (1), pp. 1-16

Cios, K.J. (2000) "Medical data mining and knowledge discovery", IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society. 19 (4), pp. 15.

Clementini, E., Sharma, J., and Egenhofer M. J. (1994) "Modelling Topological Spatial Relations: Strategies for query processing" *Computer and graphics*. 18 (6), pp.815 – 822.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006), "Spatial Prediction on a River Network", *Journal of Agricultural, Biological, and Environmental Statistics.* 11 (2), pp. 127-150.

Dale, M.R.T. and Fortin, M. (2009) "Spatial autocorrelation and statistical tests: Some solutions", *Journal of Agricultural, Biological, and Environmental Statistic.* 14 (2), pp. 188-206.

Dasarathy, B.V. (2003) "Information fusion, data mining, and knowledge discovery", *Information Fusion.* 4 (1), pp. 1-1.

Deng, M., Liu, Q., Cheng, T. and Shi, Y. (2011) "An adaptive spatial clustering algorithm based on delaunay triangulation", *Computers, Environment and Urban Systems.* 35 (4), pp. 320-332.

Ding, W., Eick, C.F., Yuan, X., Wang, J. and Nicot, J. (2011) "A framework for regional association rule mining and scoping in spatial datasets", *GeoInformatica.* 15(1), pp. 1-28

Estivill-Castro, V. and Lee, I. (2002) "Multi-level clustering and its visualization for exploratory spatial analysis" *GeoInformatica*, 6 (2002), pp. 123–152

Fang, G., Wei, Z. and Yin, Q. (2008) "Extraction of spatial association rules based on binary mining algorithm in mobile computing", IEEE. pp. 1571

Fangju Wang and Yunli Sun (2002) "Spatial object clustering and buffering", *IEEE Multimedia.* 9 (1), pp. 26-42.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) "From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P.Smyth and R. Uthurusamy (Eds.)," *Advances in data mining and knowledge discovery (*pp 1-34). Cambridge MA: MIT Press.

Flentje, P., Stirling, D., Palamara, D. and Chowdhury, R.N. (2007) "Landslide susceptibility and landslide hazard zoning in Wollongong", Proceedings 10th Australia New Zealand Conference on Geomechanics: Common Ground. Brisbane, Australia, October 21-24, 2007. (2) pp 392-397.

Franklin, J. (1995) "Predictive Vegetation Mapping: Geographic Modelling of Bio-spatial Patterns in Relation to Environmental Gradients", Progress *in Physical geography* 19 (4), pp.474-499

Fröhlich, P., Simon, R., Baillie, L., Roberts, J. and Murray-Smith, R. (2007) "Mobile spatial interaction", ACM. pp. 2841.

Furong, J., Junhui, Y. and XieFu, W. (2010) "Pseudo-Association Rules algorithm in data mining", *IEEE*. 7, pp. 3070 - 3074

Gatrell, A.C. and Bailey, T.C. (1995) *Interactive spatial data analysis,* Harlow: .Longman Scientific &Technical.

Gaixiao, L., Rencan, P., Yidong Z., and Jidong, Z. (2010) "Spatial Data Mining and its application in Marine Geographical Information System", IEEE. pp. 514.

Getis, A. (2007) "Reflections on spatial autocorrelation", *Regional Science and Urban Economics.* 37 (4), pp. 491-496.

Ghosh, A. and Freitas, A., A. (2003) "Guest editorial data mining and knowledge discovery with evolutionary algorithms", IEEE *Transactions On Evolutionary Computation*. 7 (6), pp. 517 - 518.

Gregory, D., Johnston, R. and Pratt, G. Eds. (2009) *Dictionary of Human Geography.* 5th ed. Hoboken, NJ, USA: Wiley-Blackwell. [Online] Available at :http://site.ebrary.com/lib/uoh/Doc?id=10308208&ppg=816 [Accessed 18 August 2012]

Guo, D., and Mennis, J. (2009) "Spatial data mining and geographic knowledge discovery: An introduction" *Computers, Environment and Urban Systems.* 33, pp.403–408

Güting, R., H. (1994) "An introduction to spatial database systems" *The International Journal on Very Large Data Bases*. 3 (4), pp. 357 – 399

Gunther, O. and Buchmann, A. (1990) "Research Issues in Spatial Database" SIGMOD RECORD. 19 (4), pp.61-68

Hammawa, M., B. and Sampson, G., L. (2011) "Applying Data Mining Research Methodologies on Information Systems", *Orient. J. Comp. Sci. and Technol.* **4** (2), pp. 241-251. [Online] available at: http://www.computerscijournal.org/OJCSTAbsArchive.asp?Vol=4&issue=2 [Accessed 6 August 2012]

Hara, M., Hirata, K., Fujihara, M. and Oono, K. (1996) "Vegetation Structure in Relation to Micro-landform in an Evergreen Broad-leaved Forest on Amami Ohshima Island South-West Japan," *Ecological Research*. 11, pp. 325-337.

Hao, C. and Lu, X. (2010) "Spatial Heterogeneity of Vegetation and Its Causes in Southern Yunnan Province", IEEE. pp.126.

Qv, Z. X. (1984) *Plant Ecology*. Beijing: Higher Education Press: pp. 45-51.

Highsmith, J. (2002) *Agile Software Development Ecosystem*. Addison- Wesley Professional

Hengqing, T. and Li, G. (2008) "Application of Data Mining in Psychological Evaluation", *IEEE.* pp. 343

Hochachka, W. M, Carunna, R., Fink, D., Munson, A., Riedewald, M., Soroka, D. and Kelling, S., (2007) "Data Mining for Ecological Discovery" *Journal of Wildlife Management*. 71 (7), p.503

Huang, C., Yang, Y., Yang, D. and Chen, Y. (2009) "Applications Of Data Mining Techniques To Automatic Frog Identification", *Applied Artificial Intelligence.* 23 (7), pp. 553-569.

Imielinski, T., and Mannila, H, A (1996) "database perspective on knowledge discovery". Communications of the ACM. 39 (11), pp. 58-64.

Interagency Technical Team (ITT): 1996, Sampling Vegetation Attributes, Interagency Technical Reference, Report No. BLM/RS/ST-96/002+1730. Denver, CO: U.S. Department of the Interior, Bureau of Land Management – National Applied Resources Science Centre. [Online] Available at: http://www.blm.gov/nstc/library/pdf/samplveg.pdf. [Accessed 22 Sept. 2012]

Jain, A. K., and Dubes, R. C. (1988) *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Judea, P. (1983) *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. New York: Addison-Wesley. pp.7

Kakamu, K., Polasek, W. and Wago, H. (2008) "Spatial interaction of crime incidents in Japan", *Mathematics and Computers in Simulation.* 78 (2), pp. 276-282.

Koenig, W.D. (1999) "Spatial autocorrelation of ecological phenomena", *Trends in Ecology & Evolution.* 14 (1), pp. 22-26.

Kohavi, R. and Provost, F. (2001) "Applications of Data Mining to Electronic Commerce", *Data Mining and Knowledge Discovery.* 5 (1), pp. 5-10.

Koperski, K. and Han, J. (1995) Discovery of spatial association rules in geographic information databases. In Proceeding of the 4th Int'l Symposium on Large Spatial Databases (SSD'95): Portland, Maine, Aug. pp 47-66.

Krzysztof, K., Junas, A., and Jiawei, H. (1996) "Spatial Data Mining: Progress and Challenges" Survey Paper

Leedy, P. D. and Ormrod J.E (2005*) Practical research: planning and design,* Upper Saddle River, N.J.: Merrill Prentice Hall

Lee, A.J.T., Hong, R., Ko, W., Tsao, W. and Lin, H. (2007) "Mining spatial association rules in image databases", *Information Sciences.* 177 (7), pp. 1593-1608.

Legendre, P. and Fortin, M-J. (1989) "Spatial Pattern and Ecological Analysis" Vegetation. 80, pp107-138

Legendre, P. (1993) "Spatial Autocorrelation: Trouble or New Paradigm?" *Ecology*. **74**, pp. 1659–1673.

Leung, Y. (2010) Knowledge Discovery in Spatial Database: Advances in Spatial Science. Berlin Heidelberg: Springer Verlag

Levy, E. B. (1927) "Grasslands of New Zealand", *New Zealand Journal of Agriculture* 34, 143–164.

Levy, E. B. and Madden, E. A. (1933) "The point method of pasture analysis", *New Zealand Journal of Agriculture*. 46, pp. 267–269.

Li, Y., Liu, X and Zhu F. (2010) "Application of data mining in intrusion detection", *IEEE.* 10, pp. V10-153 - V10-155

Esther, M., Kriegel, H. and Sander, J. (2001) "Algorithm and Application for Spatial Data Miming" Geographic *data Information and Knowledge Discovery Research Monograph in GIS.*

Mennis, J. and Guo, D. (2009) "Spatial data mining and geographic knowledge discovery—An introduction",*Computers, Environment and Urban Systems.* 33 (6), pp. 403-408

Miner, A.S., Vamplew, P., Windle, D.J., Flentje, P. and Warner, P., (2010) "A comparative study of Various Data Mining techniques as applied to the modeling of Landslide susceptibility on the Bellarine Peninsula, Victoria, Australia*" International Association of Engineering Geology and the Environment (2010), Geologically Active (11),* Auckland, New Zealand: Research Online.

Nelson, T.A. and Boots, B. (2008) "Detecting Spatial Hot Spots in Landscape Ecology", *Ecography,* 31 (5), pp. 556-566

Neuman, A., Freimark, H. and Wehrle, A. (2010) "Geodata Structures and Data Models" [online] Available at: https://geodata.ethz.ch/geovite/ -Version September 2010. [Accessed 12th August 2012]

Nisbet, R., John, Elder. and Gary, M. (2009) "Basic Algorithms for Data Mining — A Brief Overview". *Handbook of Statistical Analysis and Data Mining Applications*.Academic Press. [Online] available <*http://common.books24x7.com.libaccess.hud.ac.uk/toc.aspx?bookid=37310*> [Accessed August 10, 2012]

Nooghabi, M. J, Nooghabi, H. J. and Nasiri, P. (2010) "**Detecting Outliers** in Gamma Distribution" *Communications in Statistics - Theory and Methods*. 39 (4), pp. 698 - 706

Nuevo, E., Piattini, M. and Pino, F.J. (2011) "Scrum-based Methodology for Distributed Software Development", *IEEE*, pp. 66

Papadopoulos, A.N., Manolopoulos, Y. and Vassilakopoulos, M.G. (2004) *Spatial databases: technologies, techniques and trend.* US: Idea Group

Pausas, J.G. and Austin, M.P. (2001) "Patterns of plant species richness in relation to different environments: An appraisal", *Journal of Vegetation Science.* 12 (2), pp. 153-166.

Pearson, R.G., Lees, D.C., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P. and Dawson, T.P. (2006) "Model-based uncertainty in species range prediction", *Journal of Biogeography.* 33 (10), pp. 1704-1711.

Peng, C., Hongye, S., Lichao, G., and Yu Q. (2010) "Mining fuzzy association rules in data streams", *IEEE.* 2nd International Conference on Computer Engineering and Technology. 4, pp. V4-153 - V4-158

Pérez-Ortega, J., Miranda, F., Reyes-Salgado, G., Santaolaya, R., Pazos, R., Rodolfo, A. and Mexicano, A (2010), "Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico", *International Journal of Combinatorial Optimization Problems and Informatics.* 1 (1), pp. 61-67.

Perry, J. N., Liebhold, A. M., Rosenberg, M. S., Dungan, J., Miriti, M., Jakomulska A., and Citron-Pousty S. (2002). "Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data" *ECOGRAPHY*. 25, pp. 578–600

Pudi, R. and Krishna, P. R. (2009) *Data Mining*. India: Oxford University Press

Raza, K. (2012) "Application of Data Mining In Bioinformatics*", Indian Journal of Computer Science and Engineering*. 1 (2), pp.114-118

Razin, S.V. and Larovaia, O.V. (2005) "Spatial Organization of DNA in the Nucleus May Determine Positions of Recombination Hot Spots", *Molecular Biology*. 39 (4), pp. 543-548.

Ren, J. and Yin. S. (2010) "Research and improvement of clustering algorithm in data mining", *IEEE*. pp. V1-842. 2nd International Conference on Signal Processing Systems (ICSPS).

Ritchie, M. E. (2009) *Monographs in Population Biology: Scaling, Heterogeneity and the Structure of Ecological Communities*. United States: Princeton University Press. pp. 122 [Online] Available at http://site.ebrary.com/lib/uoh/Doc?id=10359226&ppg=132. [Accessed 22, Nov 2012]

Rossi, J.P. and Quénéhervé, P. (1998) "Relating Species Density to Environmental Variables in Presence of Spatial Autocorrelation: A Study Case on Soil Nematodes Distribution". *Ecography*. **21**, pp. 117–123.

Schneider, M. (2002) *Spatial Data Types: Conceptual Foundation for the Design and Implementation of Spatial Database Systems and GIS*. [Online] Available at: http://www.cise.ufl.edu/~mschneid/Service/Tutorials/TutorialSDT.pdf [Accessed 23, July 2012]

Segurado, P. and Arau´ jo, M.B. (2004) "An evaluation of methods for modelling species distributions", *Journal of Biogeography.* 31, pp. 1555–1568.

Shekhar, S., Schrater, P.R., Vatsavai, R.R., Weili Wu and Chawla, S. (2002) "Spatial contextual classification and prediction models for mining geospatial data". pp. 174.

Shekhar, S., Zhang, P., Huang, Y., and Vatsavai, R. R. (2003). "Trends in spatial data mining", *Data mining: Next generation challenges and future directions*, pp. 357-379.

Shekhar, S., Zhang, P. and Huang, Y. (2005) *Spatial Data Mining.* US: Springer. pp. 833-851.

Sloan, C. D., Duell, E. J., Shi, X., Irwin, R., Andrew, A. S., Williams, S. M., and Moore, J.H. (2009) "**Ecogeographic genetic epidemiology"** *Genet Epidemiology,* **33, pp.** 281-289

Smyth, P., Hand, D.J. and Mannila, H. (2001) *Principles of data mining,* Cambridge, Mass: The MIT Press.

Soares, C., Peng, Y. and Meng, J. (2008) *Applications of Data Mining in E-Business and Finance,* Amsterdam: IOS Press,

Strachan, S. and Murray-Smith, R. (2009) "Bearing-based selection in mobile spatial interaction", *Personal and Ubiquitous Computing.* 13 (4), pp. 265-280.

Sumathi, N., Geetha, R. and Bama, S. (2008) "Spatial Data Mining - Techniques Trends and its Applications" *Journal of Computer Applications* 1 (4), pp. 28-30

Thirumurugan, S. and Suresh, L. (2008) "Statistical spatial clustering using spatial data mining", pp. 26.

Thuiller, W., Schurr, F.M., Sykes, M.T., Zimmermann, N.E., Albert, C., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, A., Hickler, T., Midgley, G.F. and Paterson, J. (2008) "Predicting global change impacts on plant species' distributions: Future challenges", *Perspectives in Plant Ecology, Evolution and Systematics.* 9 (3), pp. 137-152.

Todem, D., Fine, J. and Peng, L. (2010) "A Global Sensitivity Test for Evaluating Statistical Hypotheses with Nonidentifiable Models", *Biometrics.* 66 (2), pp. 558-566.

Urbach, D. and Moore, J.H. (2011), "The spatial dimension in biological data mining", *BioData mining.* 4 (1), pp. 6-6.

UNESCO. (2007) [ONLINE] Available at: http://whc.unesco.org/en/list/1083. [Accessed 27th, August 2012].

Wang, Z., Q. (1999) *Geo-statistics and Its Application in Ecology*. Beijing : Sciences Press. pp. 168-169.

Wang, J. eds. (2003) *Data Mining: Opportunities and Challenges.* US: IGI Global

Wang, L., Lu, J., and Yip, J. (2007) An Effective Approach to Predicting Plant species in an Ecological Environment, In Proceedings of the 2007 International conference on Information and Knowledge Engineering (IKE' 07), Las Vegas Nevada, USA, June 25-28, 2007, pp. 245-250.

Weaver, W. (1958) A quarter century in the natural sciences, Annual Report, New York : The Rockefeller Foundation

Weaver, W. (1948) 'Science and complexity', *American Scientist.* 36, pg. 536-544.

Wilson, A. G. (2000) *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Pearson Education

Wilson, A. G. (2002) 'Complex Spatial Systems: Challenge for the Modeller', *Mathematical and Computer Modelling*. 36, pg 379 – 387

Wilson, A.M., Thabane, L. and Holbrook, A. (2004) "Application of data mining techniques in pharmacovigilance", *British journal of clinical pharmacology.* 2 (57) pp. 127-127.

Witten, I.H. (2008) *Data mining algorithms Part 1*, Henry Stewart Talks, London.

Wu, C. and Sharma, R. (2012) "Housing submarket classification: The role of spatial contiguity", *Applied Geography*, 32(2) pp. 746.

Wu, J. G. (2000) "Landscape Ecology-concepts and Theories," Chinese
Journal of Ecology, 19, pp. 42-52.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J. and Steinberg, D. (2008) "Top 10 algorithms in data mining", *Knowledge and Information Systems.*14 (1) pp. 1-37.

van Horssen, P. W., Schot, P. P., and Barendregt A. (1999). "A GIS-based plant prediction model for wetland ecosystems" *Landscape Ecology:* Kluwer Academic Publishers. **14, pp.** 253–265.