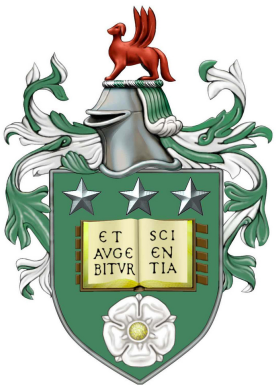
Predictive Spatial Models for Mineral Potential Mapping

Adamu Mailafiya Ibrahim



Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Computing

November 2016

The candidate confirms that the work submitted is his/her own, except where work which has formed part of a jointly authored publication has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. Some parts of the work presented in Chapters 4 has been published in the following articles:

Ibrahim, A. M., Bennett, B., (2015) The Optimisation of Bayesian Classifier in Predictive Spatial Modelling for Secondary Mineral Deposits Procedia Computer Science, held in conjunction with the Complex Adaptive Systems Conference San Jose, CA November 2-4, 2015.

My contributions: Main author, wrote all sections.

Other author contributions: Wording of the paper; discussion and suggestions of the formalism.

Chapters partly based upon work: Chapter four is also for the optimisation of Naive Bayes algorithm on Predictive Spatial Modelling using secondary mineral deposits predictive data approach.

Ibrahim, A. M., Bennett, B., (2014). The assessment of machine learning model performance for predicting alluvial deposits distribution. Procedia Computer Science, 36(0):637 642, 2014. Complex Adaptive Systems Philadelphia, PA November 3-5, 2014.

My contributions: Main author, wrote all sections.

Other author contributions: Wording of the paper; discussion and suggestions.

Chapters based upon work: Part of Chapter 4, PSM performance evaluation and dealing with Spatial Autocorrelation (SAC).

Ibrahim, A. M., Bennett, B., and Campelo, C. E. C. Predictive Expert Models for Mineral Potential Mapping, pages 3168. Mehdi Khosrow-Pour (Editor), Encyclopedia of Information Science and Technology, Third Edition, IGI Global Publishing Global, Hershey, PA, USA, 2015. ID: 112744.

My contributions: Main author, wrote all sections.

Other author contributions: Wording of the paper and discussion. Chapters based upon work: Chapter six

Ibrahim, A. M., Bennett, B., (2014) Point-based Model for Predicting Mineral Deposit Using GIS and Machine Learning- 1st International Conference on Systems Informatics, Modelling and Simulation, SIMS2014 Sheffield UK. pp.167,168, 29 April - 1 May 2014- 11.K.Intelligent Systems and Applications ISBN 978-0-7695-5198-2.

My contributions: Main author, wrote all sections.

Other author contributions: Wording of the paper; discussion and suggestions.

Chapters based upon work: Part of Chapter 4 of this thesis on the concept of designing point pattern analysis and Point-based PSM-MPM of mineral data deposits.

Ibrahim, A. M., Bennett, B., (2013) Predictive Model For Mineral Potential Mapping - Proceedings of the Conference on Spatial Information Theory (COSIT 2013) Doctoral Colloquium LNCS.

My contributions: Main author, wrote all sections.

Other author contributions: Wording of the paper; discussion and suggestions.

This thesis is dedicated to God almighty and to the loving memory of my late brother and mentor Lieutenant Commander Yusuf Mailafiya Ibrahim (Nigerian Navy) who passed away in 2009 on board Naval Helicopter while in active military service to our dear country Nigeria. May Allah "Subhanahu wata'alla" grant him "Al-jannatul firdausi" Amin.

I also wish to dedicate this work to my family (children) for enduring my long absence during the period of this study.

Acknowledgements

I wish to first and foremost acknowledge God almighty for the good health and wisdom he bestowed upon me in completing this great task.
I wish to also recognise the unflinching and untiring support of my supervisor Dr Brandon Bennett throughout the period of this research. Indeed, Dr Brandon is a man of great wisdom who knows how to manage people and situations even when all hope seems lost. I am truly grateful for his support both morally and financially.
I would like to thank the Federal Government of Nigeria for providing financial funding and support through their educational scholarship for this Ph.D.
I will also like to express my sincere appreciation to my viva reviewers, Dr Kassim Mwitondi and Dr John Stell, for agreeing to be the reviewers of this thesis.
I would like to thank my advisor and Transfer Chair, Prof. Anthony Cohn and my Transfer Examiner Dr Marc de Kamps for their valuable feedback and direction on how to achieve the set objectives of this work.
I would like to thank the members of the KRR group now called the Artificial Intelligence Theme for their enormous suggestion and contribution.
Special thanks go to members of students and staffs of the School of Computing for their feedbacks on my work. Specials to mention are, Elaine Duffin, Sam Wilson and Dr Mohammad Sulaiman Khan.
I would also like to acknowledge the input made by the unknown reviewers of my published papers for their comments and suggestions.
I would like to thank the School's administrative and support staff

for all the assistance provided. These include Judi, Charlotte, Teresa, Graham, Ian and Georgina Lambe; a former school of Computing student support personnel, who helped me sort out my admission to the University of Leeds and also gave me a very warm welcome on my arrival at Leeds.

I wish to thank Mr Thierry Hanser of Lhasa Limited Leeds for his tremendous guidance and practical support at the initial stage of this work.

I will also like to thank Professor O.E Osuwagu my MSc supervisor, for encouraging me to pursue further studies up to PhD level, and also gave a beautiful reference letter, in support of my admission to University of Leeds.

I would also love to say a big thank you to all my family members which include: Zeenat, Fatima, Nana Faiza, Fidaussi, Khalil, Faddila, Fauzuiyya, Fareeda, Khalifa .Specific to mention here is my beautiful wife Hajiya Zeenat for her patience, troubles (lol) and prayers throughout my academic career here in the University of Leeds, which encourages me to work harder to achieve the set goal and also for taking good care of the kids and enduring my absence for most part of this study, indeed she really deserved commendation. To my mother Hajiya Raliyat Ibrahim Mailafiya, I say thank you for giving birth to a worthy child.

A big thank you to my beautiful wife and also my reading partner Fatima Isiaka Mailafiya for her support and care throughout the period of this study. Finally, I would like to thank my friends both here in the UK and in Nigeria, especially Dr Shehu Magaji Suleiman of ABU Zaria, Abubakar Mohammed and Sam Danso in the UK for their friendship and support.

Abstract

Modelling and prediction of spatially distributed data such as the secondary cassiterite mineral distributions are often affected by spatial autocorrelation (SAC); a phenomenon that violates attributes data independence in space, which leads to type1 errors in classical statistics and overfitting or underfitting in machine learning (ML) classification respectively. The concept of overfitting and underfitting of spatially distributed datasets in an ML classification has not been properly addressed by the traditional random holdout technique of model validation, and this is a challenge to the assessment of predictive spatial model performance in spatially distributed datasets.

The thesis presents an approach to predictive modelling and performance evaluation of spatially distributed secondary mineral dataset, represented as points, using supervised machine learning (ML) classification. The work involves a systematic geological data survey of the existing mineral location coordinate points and other mineralisation attributes, in the Plateau Younger Granite Region (PYGR) of Nigeria. The predictive characteristics or values are extracted from a 2D space of discrete coordinate points using GIS into an ML acceptable format, consisting of 749 by 21 dimension (i.e., observational data points by the predictive attributes), with two classes of 0 & 1 representing mineralised and non-mineralised points respectively. The attributes describing the secondary mineral formation were used to build a *point based* predictive spatial model for mineral potential mapping (PSM-MPM) and using random holdout validation technique to assess its performance.

The thesis conducted predictive performance evaluation of the PSM-MPM to overfitting and underfitting by proposing a novel validation technique of *spatial strip splitting* (SSS) that *spatially splits* predictive data into training and testing; the proposed method reveals the detrimental effect of both the overfitting and underfitting associated with the conventional ML classification model validation of random holdout (RHO) or cross validation. The work also carried out a comparative analysis of PSM-MPM performance that involves the trio of performance evaluation techniques which include: attributes data preprocessing technique of principal component analysis (PCA); PCA-RHO with preprocessing that selects the best attribute subsets, the RHO without preprocessing and the novel SSS validation technique. The result showed that the SSS technique is the ideal method of assessing PSM-MPM performance because it shows clearly the detrimental effects of both overfitting and underfitting and provides more informative performance results when implementing PSM-MPM.

Abbreviations

AI	Artificial Intelligence
CDF	Cumulative Distribution Function
CSR	Complete Spatial Randomness
DA	Discriminant Analysis
DEM	Digital Elevation Model
DT	Decision Tree
GIS	Geographic Information System
KNN	K-Nearest Neighbour
LGR	Logistic Regression
MA	Measure of Arrangement
MD	Measure of Dispersion
NB	Naive Bayes
NYGR	Nigeria Younger Granite Region
PCA	Principal Component Analysis
PSM-MPM	Predictive Spatial Model for Mineral Potential Mapping
PSM	Predictive Spatial Model
PPA	Point Pattern Analysis
PYGR	Plateau Younger Granite Region
ROC	Receiver Operating Characteristics
RHO	Random Holdout
SAC	Spatial Autocorrelation
SSS	Spatial Strip Split
SVM	Support Vector Machine
TB	Tree Bagging
WOE	Weight of Evidence

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Overview of the Chapter

This chapter provides the general introduction of the thesis, highlighting the underlying concept and motivation behind the research as well as defining the research questions and objectives. The steps leading to research contributions in this thesis are also highlighted. Finally the chapter explains the general outline of the entire thesis.

1.2 Introduction

In many developing countries, such as Nigeria, lack of the complete geo-exploration dataset ideally required for mineral potential mapping seriously restricts economic development. The unavailability of such data is caused either by the absence of technical ability to deploy computer-based techniques in mineral exploration data surveying or lack of adequate data management systems required for the useful study of the mineral data components to build potential mineral deposit models.

Known mineral deposits are sometimes represented by points in a particular region on a map. The occurrence of mineral distribution represented by the spatial pattern of such points can be characterised and analysed using point pattern analysis (PPA) (Boots & Getis, 1988; Diggle, 1983). The distribution pattern of mineral occurrences is the primary concern for geoscientists in mineral explo-

rations since these patterns are often non-random (Bishop, 1995) because, they are a result of the interplay between individual geological features such as lithological rocks, faults and proximity to the primary source of deposits, that genetically control their occurrence (Bonham-Carter, 1994). For mineral exploration, the study of the spatial association between existing mineral occurrence points and geological features is needed (Walker *et al.*, 2005) to determine the spatial distribution patterns of mineral locations and to appreciate the relationship between mineral deposit locations and geological features (Bonham-Carter, 1994). The analysis of the spatial associations of known particular mineral occurrence points with their geological features is very useful for weighing the relative importance of the type of geological feature that affects the presence of a particular mineral (Bonham-Carter, 1994).

Before defining *Predictive Spatial Model For Mineral Potential Mapping* (PSM-MPM), there is the need first to introduce some important fundamental concepts associated to PSM-MPM, as follows:

- *Mineral deposits*

The term *Mineral deposits* are the concentration or existence of one or more useful substances that are for the most part sparsely distributed in the Earth's crust (Bateman, 1951).

- *Mineralisation*

Mineralisation is the processes that lead to the formation of mineral deposits in a given location (Bateman, 1951).

- *Predictive Model*

A *Predictive Model* is a sort of computational process or set of mathematical equations that takes descriptor variables and calculates estimates for responses. The model tries to explain the relationship between the input descriptor variables and output response variables. The model is viewed regarding its usefulness to the set task rather than its perfection since models are merely a representation of reality. Predictive modelling is employed in a situation where estimates or forecasts are required. The design and implementation of a *Predictive Spatial Model* are the primary concerns of this thesis.

- Predictive Spatial Model for Mineral Potential Mapping (PSM-MPM)

A PSM-MPM is regarded as a form of computational process or mathematical representation of relations between the recognition criteria of mineral deposits in the form of spatial elements, geo-data features, and the target output mineral presence or absence (class). The model takes descriptor variables as inputs and tries to represent the relationship between the descriptors as input variables and to calculate an estimate of some unknown properties as output.

The PSM-MPM is expressed using empirical mathematical equations derived from the general definition of predictive modelling. The equation 1.1 designed in this work represents the relationships between mineralisation attributes or mineral deposits recognition criteria (i.e., descriptors/attributes) and the target mineral deposits or points, as represented in the predictive map of Plateau Younger Granite Region (PYGR) of Nigeria.

$$\text{PSM-MPM} = \langle\, \mathcal{S},\ \mathcal{A},\ \mathcal{V},\ \Pi, \rangle \tag{1.1}$$

Where,

- PSM-MPM is the *predictive spatial model for mineral potential mapping* ,
- $\mathcal{S}$ is a set of spatial elements which will typically be associated with a pair of spatial coordinates (e.g. for real number coordinates we would have $\mathcal{S} = \Re \times \Re$),
- $\mathcal{A} = A_i \times \ldots \times A_n$ where each A_i is a set of possible values of some attribute describing the geological domain of the area such as rocks, mineral location area etc. Thus each $\langle a_1, \ldots, a_n \rangle \in \mathcal{A}$ is a tuple giving the values of each of the n attributes of some data set,
- $\mathcal{V}$ is a set of possible output values indicating mineral presence or absence (e.g., a binary value in $\{0, 1\}$ or a real number in the range $[0 \ldots 1]$).
- $\Pi : \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{V}$, where $\mathcal{D} = \{d \mid d : \mathcal{S} \rightarrow \mathcal{A}\}$. Here, $\mathcal{D}$ is the set of all possible data maps, with a data map being a map from each spatial location in $\mathcal{S}$ such as geological maps to tuple attribute values in $\mathcal{A}$.

1. GENERAL INTRODUCTION

The PSM-MPM in this work is regarded as a representation of mineral deposit occurrence over space using mineralisation attributes. The machine learning (ML) classification technique is used to build PSM-MPM that takes the spatial relationship between mineralisation variables of the PYGR area as input and determines an estimate of one or more unknown variables as output that signifies the presence or absence of a target label. The targets of this work are the secondary mineral occurrence deposits consisting of the geo-data which is inherently spatially structured (Bonham-Carter, 1994; Dark, 2004; Liebhold & Gurevitch, 2002; Rahbek *et al.*, 2007).

Most mineral deposits geo-data exhibit some degree of correlation in space (Guisan *et al.*, 2006; Kissling & Carl, 2008). The similarity between objects values and events in space to other objects that are co-located or nearby is referred to as *spatial autocorrelation* (SAC) (Goodchild, 1987). The secondary mineral deposit structure obtained from PYGR, represented by points on a geological map of PYGR, is a real world example of mineral occurrences that portrays the existence of SAC, which affects the correlation values between mineral deposit positions (points) and other relatively close locations in space. Just as the concept of SAC affects the prices of houses and valuations of real estate market through associations between a house and comparable nearby houses (Griffith, 2013).

From a mathematical point of view, SAC means that a variable value observed at one location is significantly dependent on the values of the same variable in neighbouring regions, thereby violating the assumption of independence that characterises most statistical analysis. Irrespective of what the processes are that create the spatial structure of the data distribution, the presence of SAC is a significant challenge for standard statistical or model tests such as analysis of variance, correlation and classification modelling because, such statistical or modelling methods assume independently distributed errors (Legendre, 1993; Legendre & Legendre, 1998).

In this study, a novel approach to predictive spatial model performance evaluation strategy through validation has been established to address the problem of PSM-MPM overfitting and underfitting due to SAC in ML classification. Traditional methods of model validation such as; re-substitution, random holdout (RHO) and cross-validation (Porwal, 2006), have not adequately addressed the

problem of model overfitting and underfitting in predictive spatial modelling using the standard ML classification performance validation due to predictive attribute data dependence of each other in space or SAC. In other words, an attempt to address overfitting and underfitting as a consequence of the presence of SAC in modelling spatial distribution data is made in this work as the major challenge identified in PSM-MPM that affects predictive performance.

1.3 Research Motivation

To describe the research motivation, the motive and the inspiration behind this work needs to be explained first. The reason behind the research work stemmed from the problem of the availability of mineral resources in the PYGR area of Nigeria: specifically, the issue of the availability of secondary mineral deposits of cassiterite (tin) in the PYGR of Nigeria. In the 1970s, for instance, Nigeria was ranked as the sixth largest producer of tin by the International Tin Council due to its large estimated tin reserves. The established reserves of tin are known to reside underneath the cover of recent volcanic rocks and sediments. Since the early 1980s, however, there has been a steadily decreasing rate of tin production due to the exhaustion of the quickly discovered deposits. The sudden departure of foreign mining companies and mining expatriates from the PYGR of Nigeria in the late 1970s saw a drastic drop in cassiterite mineral production to less than 10% of its production capacity (Bowden & Kinnaird, 1978; Pastor & Turaki, 1985). The decline in the mineral production was also partly due to lack of adequate technical tools (e.g., advanced GIS tools) to capture and analyse the distribution pattern correctly, to design *predictive models* for mineral potential mapping. The lack of technical ability was evident in the new pockets of current mineral deposits discovered by the artisan or local miners that were motivated by the rise in prices of cassiterite, or tin ore, due to high demand. However, this was in contrast to the drop in the discovery and production of new cassiterite mineral deposits due to difficulties in discovering new deposits by the mining companies. The artisan miners used their local mining experience (i.e., qualitative spatial knowledge) without technological aids to prospect in areas where minerals had previously been discovered and mined by the departed mining companies. The

search for cassiterite minerals by these artisan miners recorded some success as well as failures in the discovery of new mineral (cassiterite) deposits in the PYGR. The number of success and failures was evident during the mineral occurrence geological survey of numerous mining pits in the PYGR, which indicated locations of both minerals found and minerals not found.

There is, however, the need to unravel the circumstances surrounding the fundamental inability of the mining companies to discover new mineral deposits leading to their departure from the PYGR, and the subsequent new discoveries of cassiterite deposits by artisan miners in the PYGR. One approach is to develop a robust predictive model that combines the knowledge of both the mineral discoveries by the mining companies and the artisan miners into a single model to represent the mineralisation component of the PYGR. The model should be capable of predicting the mineral potential of the PYGR area using existing mineralisation attributes dataset and generalised to other places.

The primary motivation is to gain insights into the distribution processes of spatial point dataset, such as secondary cassiterite mineral distribution and processes in the PYGR of Nigeria. The concern is in respect of the locations where secondary cassiterite minerals deposits are either present or absent within the PYGR, where cassiterite has been mined. The knowledge gained was then used to build robust *predictive models* that learn from the existing mineral distribution patterns and representation, to describe the current mineral deposits of the PYGR. The models were used to predict location or points where new mineral deposits might be present or absent.

1.4 Research Questions and Objectives

The research questions are as follows:

- How can one determine the significant relationships or correlations between mineral deposit occurrences and other geological attributes; and, how can one use these relationships to predict mineral potential? (In other words, how can one construct a predictive spatial model for mineral potential mapping (PSM-MPM))

- How can one recognise the effect of overfitting and underfitting caused by spatial attribute data dependence in PSM-MPM?
- How can one reduce or limit the detrimental effects of overfitting and underfitting on the accuracy of a PSM-MPM.

To answer the questions above, research objectives of this study are as follows:

- To determine geological features, that are spatially associated with or are indicative of mineralisation in the PYGR, and conduct spatial analysis of mineral occurrence points with associated geological features.
- To develop point-based predictive spatial models using ML classification algorithms on a spatially distributed dataset such as the secondary mineral potential mapping of the PYGR called the PSM-MPM and select the best performing classifier using predictive accuracy scores and ROC plot for further performance evaluation due to overfitting and underfitting.
- To determine the effect of spatial attributes and spatial distribution or Spatial autocorrelation (SAC) to the performance and generalisation of the PSM-MPM.
- To design and implement a method of attribute data preprocessing that optimises the performance of ML classifiers used for developing PSM-MPM.
- To develop an ML classification performance evaluation measures that best reveal the causal and detrimental effects of over-fitting and under-fitting in PSM-MPM, and to offer an ideal approach to PSM-MPM validation assessment that challenges the well-established traditional RHO or cross-validation approach.

1.5 Steps Towards Contribution

The following measures were taken to accomplish the set objectives:

- Designed and implemented a unique and systematic method of geodata collection, by first conducting a geological survey of all the mining points in an attempt to tackle data paucity in the PYGR. The desired data collected included the geological map of the area, the coordinate location of points for all the existing mining areas (i.e., latitude, longitude and elevation), by using some specialist equipment of Global Positioning System (GPS) tools and the historical mining information of the area; especially in relation to the presence or absence of minerals. Other steps taken include the digitisation of all analogue map of the PYGR using GIS to obtain the digital coordinate points of the mining locations alongside the corresponding geological and geographic features, such as type or size of rocks, spatial distances between the mining points and the geological features to be assembled in an ML classification acceptable format.

- Conducted an exploratory data analysis; using statistical and spatial data analysis techniques of point pattern and distance distribution analysis respectively, to determine relationships as well as the distribution pattern of mineral occurrences as point data needed for efficient modelling.

- Conducted an intensive literature survey on some of the established area of secondary mineral data distribution and ML classification. Other interdisciplinary areas such as geospatial data analysis were also studied to keep abreast with the state of the event to understand the gaps in modelling and prediction of spatial distribution data, specifically on causes and effect of overfitting in ML classification performance.

- Implemented a PSM-MPM using standard ML classification algorithms capable of capturing the spatial relationships among mineralisation attributes by learning the distribution pattern of the mineral deposits of the PYGR and uses the technique of RHO section for validation. The classifiers involved are Naive Bayesian (NB), Tree-Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) and Logistic Regression (LGR). The classifiers suspected to be either

overfitting or underfitting were carefully selected for further evaluations and optimisation.

- Investigated the importance of predictive spatial attributes and SAC to the predictive performance of PSM-MPM by comparing the predictive accuracy scores of each classifier produced using firstly, all the attributes datasets; then secondly, without the spatial attributes and finally using only the attribute datasets, to show the importance of spatial characteristics in PSM-MPM. The evaluation technique also used the simulation of mineral distribution datasets obtained from the PYGR that eliminates all the space attribute in the datasets and validates the PSM-MPM produce with the results of the model developed containing spatial attributes in the datasets. The investigation seeks to gain insight into the effect of spatial components of the dataset to model performance affected by SAC causing overfitting and underfitting.

- Deployed a new technique of PSM-MPM performance validation evaluation approach using attributes data sampling, which included re-substitution, RHO, half longitudinal spatial split and quartered or longitudinal *spatial strip split* (SSS) techniques. The method of model validation deployed was adapted from the work of Bahn & McGill (2013) which account for SAC by allowing predictive characteristics data to be more heterogeneous and truly correlated during modelling. The technique has been used to test the effect of only overfitting in the past, but in this work, it investigated the effect of both overfitting and *underfitting* and in a different data domain (Ibrahim & Bennett, 2014a).

- Finally, conducted a comparative analysis of the SSS techniques with the method of RHO on original datasets and PCA-RHO; data pre-processing using PCA that selects the best predictive attribute subset data in an ML classification RHO split. The aim is to identify the most efficient technique and the best classifier for determining the detrimental effect of overfitting and underfitting and optimise the predictive performance of PSM-MPM, through the performance of the classifiers.

1.6 Thesis Outline

The thesis is organised as follows:

Chapter 1 provides the Introduction into the study. The chapter highlights the concept of PSM, the research motivation, research questions, objectives of the work as well as the summary of steps taken to accomplish the set objectives..

Chapter 2 discusses the related literature and background of the current state of developments in the area of predictive spatial modelling on spatially distributed datasets, particularly for secondary mineral potential mapping of a given area. The chapter highlights the gaps and challenges in the modelling and determine the right approach to asses PSM performance validation using secondary mineral distribution dataset. The work tries to highlights steps taken to overcome the challenges of building PSM-MPM, beginning from problems of data acquisition and research area selection with justification. Other areas include; the mineralisation attributes extraction and selection of accurate predictors to deploy in building mineral deposits. It also elaborates on the concept of predictive model overfitting in ML classification due to SAC, which is inherent in the data structure of spatially distributed dataset such as the secondary cassiterite mineral distribution attributes and suggests ways of mitigating the detrimental effect.

Chapter 3 focuses on the design methodology for the work done; this includes data collection, method of data analysis, conversion and implementation of the dataset. First, it shows how the study area and the data were collected from the PYGR and how the predictive attributes were extracted using Geographic Information Systems (GIS) and analysed for model building using the followings: Firstly, using statistics to identify point data patterns and determine spatial autocorrelation (SAC) in the secondary mineral occurrence data obtained from the PYGR. Secondly, the chapter presented an explanation on how geographic and geological data are processed, and numeric attribute values are extracted using GIS software and transformed into standard supervised ML algorithms or classifiers acceptable format to build a PSM-MPM using some ML software such as MATLAB, R and WEKA. The ML classification design was systematically explained using the concept of conventional model building method; from the point of data collection, data assembly, model building, model validation or evaluation

and model selection. Some procedures for the selection of ideal model based on the performance of individual classifier were equally explained. These methods formed the foundation of a process for developing the PSM-MPM using a standard ML classification approach. The chapter also discussed the proposed approach to PSM-MPM performance assessment. Suggesting different approach that includes the survey of different model performance validation evaluation that detects the presence of overfitting and underfitting in model performance due to SAC in the datasets and to selects the best approach that offers ideal predictive model performance accuracy by the classifiers.

Chapter 4 involves the implementation of geospatial and statistical data analysis before developing a PSM-MPM using ML classification. The chapter explains the conventional tools used for the pre-modelling, modelling and post modelling process of spatial mineral occurrence data. The pre-modelling process includes statistical and spatial data analysis to determine the distribution pattern and establish correlations among predictive attributes used for modelling. The standard classification algorithms were used to build PSM-MPM and assessed the individual predictive performance to select the classifiers that are suspected to show overfitting and underfitting for further evaluations. The seven classifiers used include Naive Bayes (NB), Bagged Decision Tree or Tree-Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) and Logistic Regression (LGR). The predictive spatial attributes data selection and justification for inclusion in building PSM-MPM were conducted in this chapter to include, a method of using simulated secondary mineral distribution data to justify the importance of SAC components in PSM-MPM. Other techniques of predictive model performance evaluation due to overfitting and underfitting conducted in this chapter include a four-way model validation test, which includes; re-substitution, random hold-out (RHO), half longitudinal split and quartered longitudinal strip split (SSS).

The chapter also showed a new approach to predictive performance assessment test, through a comparative analysis of the methods that involve RHO, PCA-RHO; a pre-processing of attributes data selection using a method of attribute data dimension reduction and best subsets selection against the novel technique of SSS. The ideal method of model performance evaluation leads to the optimisation

of PSM-MPM performances as well as identify the presence and adverse effect of overfitting and underfitting in ML classification.

Chapter Chapter 5 summarises the entire findings of this thesis, highlights the major achievements, the limitations of the work, conclusion of the thesis and suggests some possible new directions to future work of this research.

Chapter 2

Review of Related Literature

2.1 Overview of the Chapter

This chapter describes the general background of the research by looking at the various literature on the predictive modelling of mineral distribution, to identify the challenges in knowledge, regarding the different methods employed. Section 2.1 is the overview of the chapter's layout. Section 2.2 introduces the chapter, while 2.3 explains the background of the research involving the use of spatial statistical analysis, GIS and ML techniques to automate the prediction of secondary mineral deposits represented in a point pattern. The secondary mineral deposits occurrence process and distribution data points is also contained in this section. Section 2.4 highlights some of the existing methods and literature reviews on modelling and prediction of the mineral occurrence and their challenges. Section 2.5 discusses the general concept of a predictive spatial models, particularly on mineral potential mapping. Section 2.6 discusses the problem of overfitting and underfitting in the random selection of training and testing dataset to validate the performance of PSM-MPM. Section 2.7 investigate the effect of spatial autocorrelation (SAC) in a dataset, which present a challenge to modelling spatial data such as the secondary mineral distribution attribute data and proffer methods of detecting and testing to deal with it, as presented in secondary mineral distribution data obtained from PYGR.

Section 2.8 introduces the use of ML classification algorithms to model and automate the prediction of mineral potential in a given area represented by points,

and the ability to measure its performances and the extent to which it can be generalised to other contexts. The model validation method that determines the generalisation of PSM-MPM that checks for predictive accuracy performance due to overfitting and underfitting was also discussed. The section also highlights the method of selecting a PSM-MPM based on a comparative analysis of the performance of the competing classifiers in terms of their ability to predict well on unseen datasets. Finally, section 2.9 summarises the chapter.

2.2 Introduction

The evolution of Machine Learning (ML) classification has significantly improved the scientific methods for modelling and predicting phenomena that go beyond human capabilities (Lary, 2010; Lary *et al.*, 2016). This chapter discusses the general background of this research thesis, highlighting the ability of ML; a branch of Artificial Intelligence (AI). The ML technique of classification was deployed due to its ability to learn patterns in spatially distributed data (i.e., secondary mineral deposit distribution) based on existing information (data), or training data to build models that can predict the occurrence of similar data, called the test sets. The supervised ML classifiers such as: Naive Bayes (NB), Bagged Decision Tree or Tree-Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) and Logistic Regression (LGR) were used to build predictive spatial models, using mineral recognition criteria (predictive attributes) derived from predictive map layers, created in geographical information system (GIS). The mineral deposit occurrence represented as points in the geological map of the PYGR were analysed to generate spatial predictive map data showing all predictive attributes in a spatial frame of reference to build predictive models for mineral potential mapping using ML classification.

The work highlights previous work done in the area of mineral deposits modelling or prediction, and proffers a new direction in this area of research through the adoption of a systematic combination of statistics, GIS and ML classification respectively. The aim is to model the distribution of secondary mineral deposits

presented as a discrete data points obtained from the PYGR area; which is currently an under-researched area (Ibrahim & Bennett, 2014a; Ibrahim *et al.*, 2015a).

2.3 Background

The thesis presented here is deemed to be in the area of *Machine Learning* (ML) classification a branch of *Artificial Intelligence* (AI) (Russell & Norvig, 2005) and spatial distribution data modelling and analysis. The research uses GIS, statistical analysis and ML classification to build a predictive model using secondary mineral distribution data obtained from the Plateau Younger Granite Region (PYGR) of Nigeria. The ML classifiers are used to learn about the distribution patterns of the mineral data as points to build a predictive spatial model for the mineral deposit potentials (PSM-MPM) of this area and test for the ability to generalise to other contexts. The ML classification algorithms or classifiers uses a range of data mining programs, capable of learning general rules from the behaviour of secondary mineral data distribution by gradually discovering patterns in the dataset as it progresses (Mitchell, 1997). The primary reason for choosing this method was the need to automate information extraction from data through computational methods to make intelligent decisions based on the existing mineral deposit datasets obtained from the PYGR. The ML classification brings together the power of computer and statistics to exploit smart decision.

Many real life problems present a well-structured input or output, such as gene function prediction, image processing and geospatial datasets, this has made it easier to model the structure of the input to predict the output. Several predictive models have been used to predict mineral deposits potential, the output can either be a discrete variable (classification) or a continuous variable (regression). The secondary mineral data obtained from the PYGR is a typical example of a spatially distributed discrete dataset; that is presented as an ML classification problem and a predictive modelling task with structured outputs (supervised ML). The modelling process begins with the geospatial and statistical data analysis of mineral occurrences, represented as points within a geographical space of the PYGR. A combination of the geospatial elements of the individual mineral occurrence represented as points, and other geospatial data as the predictive attributes

considered for modelling. The statistical and spatial data point analysis is an everyday type of analysis used to determine point distribution patterns and feature spatial relationships or correlations using GIS tools (Bonham-Carter, 1994; Boots & Getis, 1988).

In mineral prospecting similar to mineral potential prediction, one of the primary goals is discovering new mineral deposits. A new mineral discovery can be obtained by analysing and modelling the spatial distribution of known mineral occurrences represented as points (Bonham-Carter, 1994; Carranza & Hale, 2003). As the concept of mineral potential modelling becomes more established, various approaches to predicting mineral deposits through geographic information systems (GIS) have been developed. Recently, there has been a significant paradigm shift towards research in data mining, using machine learning and GIS. This paradigm change has been stimulated by an increase in the volume of heterogeneous geospatial data (geographical and geological), this is due to large geo-datasets that can be obtained from different maps (digital and analogue) using GIS to identify distribution patterns, correlations and any other explanations from the datasets. This shift includes the use of machine learning to predict (with some degree of uncertainty) the occurrence of mineral deposits in a given area.

Spatial data analysis has been an active area of research for the last two decades. It has improved through the use of different kinds of computer applications, such as GIS, computer-aided design (CAD), multimedia information systems, data warehousing and earth observation systems (Shekhar *et al.*, 2001). Satellite images and digital maps are examples of spatial data because the information is extracted from them by processing the data with respect to a spatial frame of reference about the earth's surface. Computer aided spatial data analysis, mapping and modelling techniques have been used in applied geosciences for many years to detect patterns in the distribution of natural phenomena (Bonham-Carter, 1994).

The choice of PYGR area was because it is rich in secondary mineral deposits. The ring-complexes that formed the province are of high level sub-volcanic orogenic intrusions exposed over a distance of about 400km (Bowden *et al.*, 1981). The mineral sediments connected with the ring complexes of the region have been the primary motivating force behind the geological study in the province, ever

since the first discovery of alluvial tin deposits (Falconer, 1912; Ibrahim & Bennett, 2014b).

The independent manner in which data was sourced from the research area due to data paucity in the PYGR has made the use of GIS very important to deploy the various mineralisation attributes data needed for ML classification modelling. The challenge of using ML to predict natural phenomena such as mineral occurrence has always been in the collection of the right datasets that can be deployed in ML. While it is very common to implement statistical or GIS approaches on their own, this research uses GIS as a tool for data collection, processing and analysis, while the ML classification is used to build a predictive spatial model for mineral potential mapping (PSM-MPM). GIS were used to provide a computer-based tool for capturing, managing, analysing, and displaying the spatial geographically referenced information and present them in either a vector or raster format. The GIS tools were used for analysing all the geo–data (spatial and non-spatial). The geodata analysis enables the mineral occurrence data to be handled spatially in a spatial reference that will allow for easy extraction and analysis of predictive attribute values.

GIS tools were very useful in manipulating both the quantitative and other additional qualitative information (Bennett, 1996) needed to be deployed to ML classification algorithms to build a PSM-MPM. The PSM-MPM developed used the secondary mineral occurrence data obtained from PYGR was validated and tested to generalise well. Predictive spatial modelling (PSM), and geospatial statistical analysis about mineral potential mapping have been well documented (Agterberg *et al.*, 1990; Bonham-Carter, 1994; Porwal, 2006) but the specific approach proposed in this research has not previously been attempted namely:

- Combination of GIS and ML.
- The application of ML and GIS technique to secondary mineral deposits of cassiterite or tin.
- The evaluation of various validation techniques of PSM-MPM due to presence of spatial autocorrelation (SAC) in the data sets that leads to overfitting or underfitting.

2.3.1 Secondary Mineral Deposit Points Distribution

Secondary mineral deposits originate from external processes caused by the environment, as well as material or chemical events, thereby instigating ore materials to concentrate at the regolith (i.e., transported by stream or river from the source to place of deposit). The physical processes involved include, erosion and weathering. Behind the secondary mineral deposit formation, the theory of *ore genesis* describes the composition in three different components, namely: source, transport (conduit) and trap (deposit Point) (Ibrahim & Bennett, 2014b). Mineral deposits are often classified based on their type (Bowden & Jones, 1978; Falconer, 1912), although classification tends to be difficult because for the multiple causes of their formation.

The secondary casiterite mineral deposits or placers are derived from the weathering and erosion of the primary casiterite deposits. Casiterite is a typical example of mineral deposits found at a secondary level of occurrence and is the principal mineral mined in the PYGR. Hence, the focus of this research. Cassiterites are chemically resistant, heavy metals and readily form residual concentrations. These levels may develop over a primary deposit (eluvial) and on slopes below the deposit (colluvial). When the cassiterite reaches a drainage system, it may be transported to a river channel and concentrated into an alluvial placer deposit. A placer deposit buried by younger sediments or lava is known as a deep lead. Deposits in oceanic submerged river channels are important sources of tin or casiterite. More than half of the world's tin production is currently from deposit in mainly in Malaysia, Indonesia and Thailand (Geoscience Australia, 2007, 2013).

Figure 2.1 shows the formation of residual (primary) and eluvial or alluvial placer (secondary) mineral deposits. The primary cassiterite mineral deposits are found within and on the rocks layer only, while the secondary mineral deposits are normally dispersed through a process of chemical weathering, and transported by river or stream to areas within the primary source (i.e., rocks), and buried beneath the earth surface. The mineral occurrence structure symbolises a typical spatial data distribution.

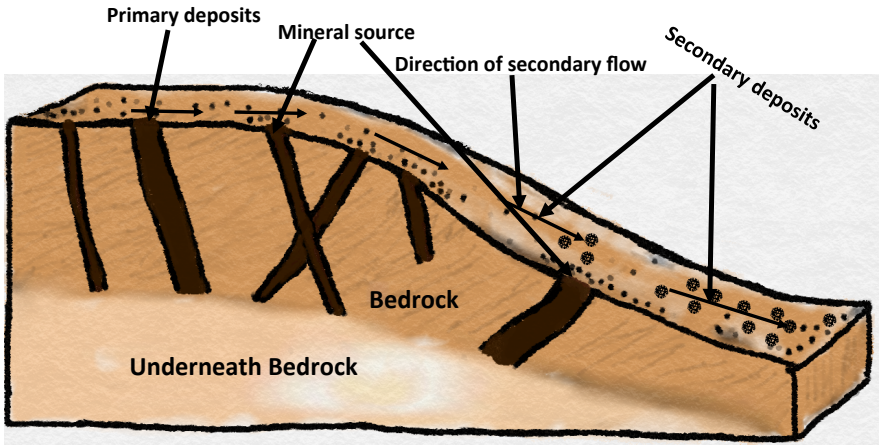


Figure 2.1: A modified diagram of secondary mineral deposit formation obtained from Haldar (2013).

The secondary mineral when represented as a set of points of a particular type can be characterised and analysed through *point pattern analysis* (PPA) (Boots & Getis, 1988; Diggle, 1983). Through cartography, real world objects are represented as points, polygon, lines in a two-dimensional plane. A common representation used for secondary mineral deposit analysis is by points; the analysis is often referred to as PPA. Part of the reason for studying the points pattern is to determine the following: a source of information about a certain phenomenon and the process responsible for its occurrence (Boots & Getis, 1988). In some cases, opinions or information about a certain event represented as point is enough to develop an explanatory model for it. Even if available information about a phenomenon is very elementary, the information gained from the point pattern analysis enables an initial insight into how the secondary mineral cassiterite process occurs in the given area.

2.4 Existing Approaches to Mineral Potential Mapping

An Expert system named *Prospector* was developed at Stanford Research Institute for evaluating mineral prospects (Duda *et al.*, 1995). It uses Bayesian inference

networks to evaluate mineral deposits by generating an estimate of the most likely mineral deposit type and an explanation of the results obtained. Even though the original Prospector did not support the use of spatial data, like geophysical surveys or geological maps, later versions were improved to support the regional assessment of mineral modelling.

Duda *et al.* (1995) used Prospector to combine predictor patterns in a study of the Island Copper deposits in British Columbia. Information was propagated in the networks by first using Bayesian updating of prior to posterior probabilities and then applying fuzzy Boolean operators. Expert opinion was used to estimate qualitative likelihood ratios and prior probabilities. The results of using Prospector to map the potential for molybdenum deposits in the Mt. Tolman area of Washington State was also published by Campbell *et al.* (1982) and Bonham-Carter (1994).

Zhou & Civco (1996) also mentioned problems affecting Knowledge-driven data integration models, such as the challenges involved when spatial data present inaccuracy and factor interdependency, assigning scores and selecting the aggregation function. A data-driven procedures require assumptions that are difficult to meet when dealing with geological variables such as linear relationships, variable dependence and normal distributions (Rigol-Sanchez *et al.*, 2003). The use of artificial neural networks (ANN) also offer some advantages over other methods because it does not make any assumptions or restrictions about predictive attributes data. ANN allowed for non-linear and interactive effects among the data (Bishop, 1995; Rigol-Sanchez *et al.*, 2003), this is the reason why ANN is still considered a "black-box" and its modelling involves a more elaborate training process (parameter adjustment) than other methods.

The weight of evidence (WOE) is another form of statistical analysis approach to predicting mineral potential mapping among geo-scientist and statistician, it uses the concept of Bayesian conditional dependence by updating prior to posterior probabilities (Agterberg *et al.*, 1990; Bonham-Carter & Agterberg, 1990). The WOE is determined by the Prospectors using estimate, for example, it uses Bayesian equations in a log-linear form using conditional independence assumption of input predictor patterns. It is probably the most popular technique for

mineral potential mapping due to natural applicability (Bonham-Carter & Agterberg, 1990).

A major work in the area of mineral resource prediction and mapping was also done by Carranza *et al.* (2008); Carranza & Hale (2000, 2001, 2003); Carranza *et al.* (1999). In this work, statistical and spatial analysis of geological data, were employed to predict the mineral potential of a particular area in the Philippines. The analysis include both quantitative and qualitative modelling techniques. However, the work was used to predict the potential of primary gold and copper deposits and involved geochemical analysis, it was a modelling approach that failed to address generalisation or validated using unseen datasets.

Rigol-Sanchez *et al.* (2003) employed ANN by adopting a back-propagation network with three layer variables. They used remote sensing data by applying geological, geochemical and geophysical data to achieve a favourable target of both present and future occurrence.

Porwal (2006) developed some mathematical models for mineral potential mapping and made some comparisons based on predictive accuracy performance regarding how well the model fitted the various mineral attributes. The mathematical model was implemented using an augmented Bayesian classifier, a hybrid model of Fuzzy/WOE and a combination of Neural/Fuzzy and made a comparison of their performance. Even though Porwal recorded some level of success, the performance of Bayesian Network obtained by Porwal was hindered by the paucity of known mineral deposit data and, therefore, suggested the need for more data acquisition for training and testing including data pre-processing before modelling. The work suggests data pre-processing without loss of vital information to avoid random noise in order to augment the training data and ensure more data points are available at an exponential rate to its predictive attributes in order to prevent the problem of the curse of dimensionality. Porwal's mathematical models were however, only applicable to sedimentary exhalative deposits (SEDEX) which are deposits formed through ore-bearing hydrothermal fluids into a water reservoir such as the ocean. The model uses predictive maps for analysis and an area-based approach with few deposits areas as observation data sets.

Just recently Ekosse & Mwitondi (2015) used a data-driven approach to Principal Component Analysis (PCA) to identify distinctive oxides of elements con-

centration that show variation in a lithological arrangement of Kaolin, using data obtained from four distinctive deposit regions of Botswana. The results of the PCA obtained were validated by graphical data visualisation tools of a smaller dimensional matrix of 28x11, with the results of the validation showing sharp differences in the three samples, indicating that discretisation was a significant challenge. Ekosse & Mwitondi (2015) formally used size discretisation of kaolin particles as a tuning parameter measuring kaolin variation among samples used in validating predictive modelling applications. The work, however, suggested a new direction for developing a predictive model based on newly extracted components and discretisation in validating new predictive modelling applications.

As several methods are used for predicting an output or target property, the output can be either a discrete attribute (classification) or a continuous (regression). Many real life problems present a well-structured input or output such as gene function prediction, image processing and geospatial data, etc. Classification in this work is treated as predictive modelling tasks with structured outputs. This work is tailored towards ML classification and spatially distributed data analysis to build a predictive model. The work used secondary mineral deposits represented as points, unlike Porwal's, within a geographical space and used other geospatial elements of the individual points as part of the predictive attributes. The spatial point analysis was used to establish spatial feature using GIS tools, determine point distribution patterns and the correlations among spatial attributes (Bonham-Carter, 1994; Boots & Getis, 1988). The spatial predictive model performance will then be validated to test for its acceptability, using standard ML classification method.

It was evident that the lack of a good predictive mineral deposit model of the PYGR area has led to an inaccurate assessment by mining companies to believe that there was no longer mineral deposits in the PYGR area, while the local miners continue to discover new mineral deposits in the same region. While it is possible that the quality or quantity of the mineral found is not economically viable, this is not the intention of the model proposed. The aim is to attempt to design a predictive model capable of managing both qualitative and quantitative attributes of mineral deposits recognition criteria and generalises well to similar data sets elsewhere, using effective means of model performance validation. Therefore, a

machine learning predictive algorithms was selected and deployed to build a PSM-MPM because the ML classification modelling approach can learn from various knowledge representations and make predictions based on learned knowledge to generalise and quantified the predictive ability of the model based on the predictive dataset.

This research is a pioneering one for the secondary mineral data in the PYGR and the desired data is not readily available. There was, therefore, the need to develop a systematic method of acquiring the appropriate predictive attribute datasets in a manner so as to model the relationship between the target mineral deposits and the recognition criteria in the PYGR experimental datasets effectively.

Some of the fundamental challenges associated with the design of PSM-MPM are first, the acquisition of the desired scientific data and preparing it to be used to develop a predictive model using ML classification. Secondly, is the selection of the classifier and the appropriate predictive model validation techniques to measure the performance accuracy of the PSM-MPM. Since the predictive performance of a good model is mostly data-dependant (Mwitondi *et al.*, 2013), the violation of the attributes of data independence in a spatially distributed cassiterite mineral due to SAC, leads to overfitting and underfitting by the classifier. The element for determining overfitting and underfitting are often realised in an exaggerated high and low predictive accuracy scores by the classifier respectively. The predictive performances of a model are often affected by the distribution pattern of the datasets that account for attributes data independences or SAC. It is, therefore, important to check for the presence or otherwise of SAC in distribution datasets when building a PSM-MPM, to avoid model overfitting and underfitting.

The uniqueness of the work undertaken in this thesis in contrast to the existing methods highlighted, however is firstly, in the type of data (i.e., study area), the techniques deployed in acquiring the data, and the method used to evaluate the performance of the PSM-MPM, through model validations using different data sampling methods for selecting training and testing data. Since mineral deposits exist in different formations and locations, it is very difficult to conclude that a single method of predicting mineral potential mapping can be applicable to all, except through a scientific modelling of a certain existing mineral deposit to predict

the future occurrences and measure generalisation to other similar datasets. The ML classification approach based on the existing mineral deposit data obtained from the PYGR of Nigeria will lead to the development of such a predictive spatial mineral potential model, and give a thorough assessment of the predictive performance based on generalisation through the model validation, measured based on the test or validation set.

A significant general problem identified in terms of classifying spatial data using ML classification is that, a learning classifier tend to overfit or underfit the particular data that has been used for training. In the case of applying ML classification on the spatially distributed secondary mineral data, overfitting may occur due to spatial dependencies or the arrangement among data items in space (Bahn & McGill, 2013; Ibrahim & Bennett, 2014b): recalled that things that are closer in space are more inclined to have the similar attribute values than those that are far apart (Neville *et al.*, 2003). The similarities or dependences among spatially co-located attribute values are often referred to as *spatial autocorrelation* (SAC).

SAC means that an observed value of a variable at one locality is significantly dependent or correlated to the values of that variable (and other related variables) at neighbouring regions (Liebhold & Gurevitch, 2002; Rahbek *et al.*, 2007) thereby, violating the modelling assumption of attribute data independence. Checking for SAC has become a conventional routine for plant ecologists and geographers (Fortin & Dale, 2005; Sokal & Oden, 1978) for the study of variations in the plants to determine the underlying distribution structure and to detect parasitic plants in a given area. SAC analysis and study, are important approaches to predictive modelling of mineral distribution data that include geological and geographical data due to their inherent spatial structure (Bonham-Carter, 1994; Guisan *et al.*, 2006; Kissling & Carl, 2008).

Previously, research linking predictive modelling of spatially distributed data that include the concept of autocorrelation was explicitly conducted by Stojanova *et al.* (2011). The work investigated the idea of spatial and network autocorrelation in predictive modelling and evaluation. The research also involves a predictive clustering framework that deals with both SAC and network autocorrelation, building a spatial predictive system that considered both autocorrelations (spatial

and network) in learning, and developed predictive clustering models using both classification and regression (Stojanova *et al.*, 2010). Although the work investigated several forms of autocorrelations as mentioned earlier, the method employed only deals with predictive modelling by identifying autocorrelation according to the orders of clusters of the similar dataset values associated with each group. The approach of the work combines both regression and classification tasks to form predictive models, for both continuous and discrete response respectively. The major limitation for the work of Stojanova *et al.* (2011) is that it deals with autocorrelation in the datasets at different stages of the SAC phenomenon rather than at a global or general level, where it will be possible to generalise the implementation at both levels (local and global) (Stojanova *et al.*, 2011). The predictive models adapt only to the local properties of the datasets that may hinder efficient execution and transferability. Recalled that the essence of applying ML classification in modelling is to provide an easier, more robust and a generalised predictive model such that, the measure of the generalisation are measured by either accuracy, sensitivity or specificity as the case may be for decision-making. The case may be to either accept, reject or seek to optimise the performance through the predictive accuracy or error rate.

The work of this thesis, however, developed a predictive model performance assessment of the ML standard classifiers to select the best through performance comparison among the classifiers used for building the PSM-MPM. The work also proposed a novel technique of PSM-MPM performance validation that seeks to detect the presence or absence of mineral potential, and to minimise the effect of overfitting and underfitting in an ML classification. SAC in spatial datasets often causes poor validation of training on the test set. The technique considered a systematic and holistic approach to spatial data sampling; that spatially splits the training and test data to improve spatial attribute data independence, thereby, reducing the SAC effect inherent in the dataset. The space cutting technique provides for a more heterogeneous rather than having a very similar or completely different training and test datasets embedded in the clustered arrangement of attributes. The space splitting or sampling enable spread among highly correlating attributes of training and test sets which are often spatial, that supports credible validation of PSM performance. Other forms of PSM validation technique include

random holdout, cross-validation and re-substitution, but are often affected by the adverse effect SAC leading to overfitting and underfitting (Bahn & McGill, 2013). The problem of overfitting and underfitting which has been identified as a big challenge in ML classification, particularly in spatial distribution dataset, as inherent in the secondary mineral distribution data obtained from PYGR, have not been effectively tackled by the traditional methods of model validation in an ML classification (Ibrahim & Bennett, 2014a). Thereby, a new approach was developed to handle overfitting due to SAC within the context of ML predictive model classification.

2.5 Predictive Spatial Model for Mineral Potential Mapping (PSM-MPM)

The research work conducted in this thesis, surveyed the predictive performances by seven selected standard supervised ML algorithms in order to select the best model for the existing mineral distribution pattern of the PYGR region, to produce a predictive mineral potential model based on the spatial association between geological, geographic and geo-spatial features. The most common existing method for mineral potential mapping is the statistical technique of Kriging, which involves a quantitative approach to modelling mineral deposits where quantified evidential weights are taken with respect to areas of known target deposits (Agterberg *et al.*, 1990; Bonham-Carter, 1994). Unlike the method of statistical kriging, the PSM-MPM is a combination of statistics for spatial point analysis (SPA) and ML classification for modelling secondary mineral potential in a given area through the validation of two separate datasets classified as training and test sets. Spatial point analysis is used to determine the distribution pattern to ascertain the non-randomness of the data and also quantify the spatial association between mineral deposits and geological features to ensure that the right attributes are used for model building. The statistical techniques used are a combination of the measure of dispersion, which involves Nearest Neighbour Distance, Quadrat analysis and Kolmogorov-Smirnov tests (K-S Test) to determine the distribution pattern as

well as the correlation between the various features of mineralisation. While statistical analyses are conducted in most cases of modelling, this research is merely deploying statistics to validate the dataset as a distribution of a non-random occurrence as well as verify the predictive attribute data selected to produce the PSM-MPM.

The ML classification modelling is considered to be very effective when the right attributes are selected Breiman *et al.* (1984), but these are often challenged by the problem of overfitting depending on the complexity of the algorithm to the data type used. While the problem of overfitting is a well-documented problem in predictive modelling, it is not widely explored in ML classification involving spatial data, particularly in secondary mineral deposits datasets (Ibrahim & Bennett, 2014b). The common method of addressing overfitting in classification modelling is through model validation or cross-validation (Bradley, 1997; Han *et al.*, 2011), which involves the technique of splitting the training and test datasets, and to validate a trained model with test or unseen data by the classifier. Since secondary mineral data exhibits spatial autocorrelation (SAC), which incidentally is a natural attribute of the dataset and helps in making predictions, the cross-validation or hold-out method of splitting training and testing data is done to ensure generality and to address the problem of overfitting or underfitting in ML classification modelling. Since SAC is a concept of space in a spatial dataset such as the secondary mineral deposits of the PYGR (Tobler, 1970), it is difficult for the ML classifiers to split training and test data along the lines of spatial components but randomly. Hence the need to design a new approach or technique to determine the performance of the spatial predictive model for the mineral deposit potentials of the PYGR that identifies both overfitting and underfitting as against the traditional method of random holdout (RHO) splitting (Porwal, 2006).

2.6 The Implication of SAC to Overfitting and Underfitting

Overfitting occurs when an ML classifier learns specific details of the particular dataset that are irrelevant to the classification problem in the general case. Such

irrelevant details could be noise due to high similarities among variables in the dataset that allows the classifier to learn easily. Underfitting, on the other hand, refers to the situation where a statistical or ML algorithm is not able to capture properly the underlying pattern in the dataset or when the algorithm does not fit data well. Overfitting often occurs usually in a situation where the model is extremely complex (Diebold, 2015; Scheres & Chen, 2012). The major determinant for the occurrence of overfitting and underfitting in classification modelling is the inability of training data to match the performance of test data or vice-versa in a random selection. The typical method of testing for overfitting and underfitting in ML classification is the cross-validation, random holdout (RHO) or the leave one out method of model validation (Bahn & McGill, 2013). This method is usually applied at the point of model performance validation with a test dataset, or when making a prediction using new and unseen dataset by the predictor. While this approach works differently for different learning algorithms depending on the type of data used, there has not been any standard method that has handled spatial predictive models, due to the homogeneity of spatial datasets or due to their similarities of values in space (Bahn & McGill, 2013; Ibrahim & Bennett, 2014a). The similarities of data attribute values in space due to their proximity, leads to high correlation or lack of independence among the predictive attributes even when split into training and test dataset. For instance, several mineral deposits may exist in different locations but at equal elevation height, other reasons for similarity in values could be due to some mineral deposit points having equal proximity or distances values to geological features like the rocks unit.

The equal proximity may be due to chance and not because of they are from same source leading to false correlation. The absence of true independence means that the classification learner is not able to learn effectively on a test or new attribute dataset efficiently what it learnt from the training set, but rather transferred the spurious attributes correlation onto the test set, thereby affecting learning and testing when using traditional RHO selection for model performance validation selection (Porwal, 2006). The similarities in spatial data values are often due to presence of spatial autocorrelation (SAC) among the attribute dataset. The resultant effect of SAC is an over-exaggerated or poor prediction on new data sets (i.e., poor model performance) in PSM-MPM. Both overfitting and underfitting

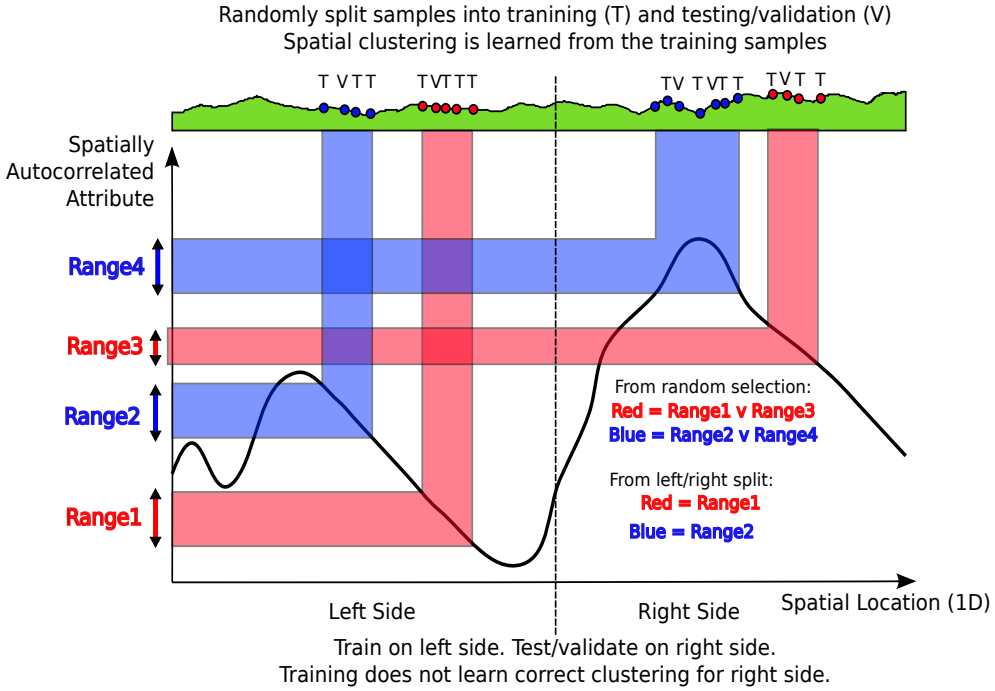


Figure 2.2: A diagrammatic representation of learning in clusters of points pattern on training and validating on test set using synthetic spatially distributed data.

lead to a poor prediction on new or test datasets in ML classification depending on the adaptability of the classifier based on the type of dataset.

While there is no concrete proof from literature to indicate that SAC increases the predictive accuracy of a classifier in a distribution data, the perception is may be based on the regularly clustering arrangement of the dataset in space (Stojanova *et al.*, 2010). One of the causes of high predictive accuracy in clustering arrangement of dataset is often due to ability of the dataset to retain the spatial composition in the dataset during learning and prediction, making the results prone to an excessively high predictive accuracy score. The challenge of learning in spatial data distribution, such as the occurrence of secondary cassiterite mineral distribution in space by an ML classifiers, sometimes varied with the type of classifier presented. The diagram in Figure 2.2 is an example of a particular synthetic spatially distributed attribute point data, drawn to demonstrate the possible cause of failure by the traditional RHO method to handle model validation on test or new dataset (model generalisation), leading to *overfitting*.

The diagram 2.2 shows a representation of synthetic point data, which demon-

strate the inability of a random selection of spatial attribute values split. The random split tends to learn from clusters of data points arranged in space. Depending on the learning style and complexity, some classifiers interact randomly between attributes using majority voting for learning and may fit perfectly well to the data pattern and hence replicate efficiently the clustering arrangement of the data in space (i.e., clustered), leading to overfitting. In some classifiers, the learning process is somewhat linear and required a simple data arrangement to learn better, but unable to replicate well unto unseen data. Such classifiers may perform poorly on the test set due to the absence of real attributes independence that reduces the effect of SAC in the datasets. The Figure 2.2 demonstrated how the training dataset in clusters represented by R1 and R2 and a test or validation set at different clusters range of R3 and R4 might lead to either overfitting or underfitting. Because the learner's ability is restricted to a particular boundary and has no proper information about the true correlation in the testing range and may therefore, predict poorly due to the clustering arrangement learnt in the range between R3 and R4.

2.7 Challenges of SAC in Spatially Distributed Points Data

SAC in predictive variables may not only be a potential violation of a model's assumptions but also lead to inflated model test measures (Segurado *et al.*, 2006). The impact of SAC to a test of model predictive power based on data RHO techniques in machine learning has not received the required attention (Bahn & McGill, 2013). Several studies have identified the testing of models on a presumed independent data as a challenge to model selection (Araujo & Guisan, 2006; Hampe, 2004).

The three different categories of independent testing data typically employed are: data collected independently, temporally independent data and spatially independent data (Townsend Peterson *et al.*, 2007). A focus on spatially true independent testing data is considered here because SAC possibly leads to interdependence among training and test data leading to overly optimistic model

performance results, while natural data fluctuations may have resulted in an excessively pessimistic model assessment result.

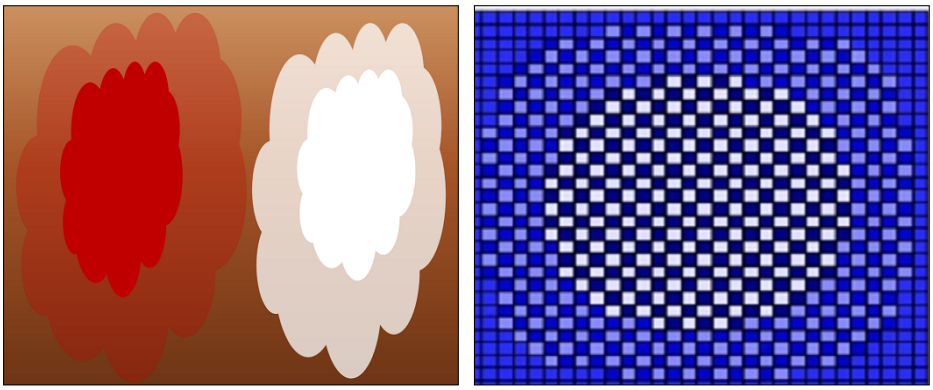
The difficulties in detecting correct model performance in secondary mineral distribution modelling are that only some degree of predictive performance into a new area is known. As such, it is hard to say how well the distribution models predicts for a new field performed well or not. For example, if a model performance measured by area under the curve (AUC) is 65%, it may not be considered a good performance when predicting into a new area. This is because the model may require some other form of evaluations using systematic inference that will allow the identification of those attributes that contribute to increases and decreases in model performance, as well as a systematic performance comparison to other model performances on similar data sets.

Empirical SAC cases involve moderate, active relationship between nearby values on a map. Most socio-economic/demographic data display a moderate positive relationship while remotely sensed satellite images almost always show a strong positive relationship. *Strong positive* SAC may occur in remotely sensed images as against the *moderate positive* SAC displayed in geo-referenced data, this is due in large part to light reflectance spreading of remotely sensed data rather than the neatly contained light reflectance in the pixel boundaries of geo-referenced data.

Moderate positive SAC occurs in maps of the population, where density tends to display average positive SAC, in part due to urbanization at a district, general, zonal or at a resident scale (Griffith, 2013).

A *moderate negative* SAC: in the literature, only a few empirical examples of negative SAC are reported. The SAC phenomenon is conceptually discussed as a term of geographic competition. In other words, if a finite amount of land is available, gains in the earth's size of a given territory is a consequence of the loss of territory size in nearby territories (Griffith, 2013). A war-torn area could be a perfect example of a moderate negative SAC. The image in Figure 2.3a depicts an example of a positive SAC in a map and also indicates how low and high values are clustered together with high values concentrating at the centre but reducing or decaying over distance away from the centre. The maximum correlation is focused at the centre but decline as it disperses. Figure 2.3b on the other hand, displays

a map of point patterns with a uniformly dispersed pattern, indicating negative SAC. Here, the high and low dissimilar variables are interspersed evenly across the map space.



(a) (b)

Figure 2.3: Diagrammatic example of degree of relative correspondence of high and low values of SAC: Figure (a) shows similar values clustered together as positive SAC as arranged while Figure (b) is shows dissimilar values clustered together on a map indicating negative SAC.

Two types of SAC might be distinguished depending on whether the processes generating the spatial structures are causal or occur by chance (Fortin & Dale, 2005; Legendre, 1993; Legendre & Legendre, 1998). In the case of the former, the spatial pattern is caused by factors that are an inherent property of the variable of interest, also known as intrinsic SAC (Fortin & Dale, 2005), for instance distance-related processes such as dispersal or geographical range extensions (Diniz-Filho *et al.*, 2003; Legendre, 1993). On the other hand, there are spatial pattern autocorrelations induced by an external process that is independent of the variable of interest; this is also referred to as induced spatial dependence (Fortin & Dale, 2005). These induced correlations arise on spatially structured environmental attributes such as the wind, topography and climatic constraints, which can cause distribution patterns to be spatially structured. Irrespective of which processes create the spatial structure of the data distribution, the presence of SAC is a significant challenge for classical statistics or model tests (analysis of variance, correlation and classification) because, statistical modelling methods assume independently distributed errors (Legendre, 1993; Legendre & Legendre, 1998). The error problem relates to the inflation of type 1 errors or sensitivity tests in classifi-

cation models (i.e. false attribution of a correlation) which signify that confidence intervals are wrongly estimated when observations are not independent.

A type 1 error is an incorrect rejection of a null hypothesis —i.e., a false positive. Hence, classical tests of significance of correlation or classification coefficients might be biased (Kissling & Carl, 2008; Legendre, 1993; Legendre *et al.*, 2002; Lennon, 2000). SAC may also affect the ability to evaluate the importance of explanatory variables (Lennon, 2000; Lichstein *et al.*, 2002). SAC can also be a major shortcoming for hypothesis testing, predictive accuracy, or in drawing an inference from statistical models (Dormann, 2007; Ibrahim & Bennett, 2014b).

SAC in secondary mineral deposits data from the PYGR is associated with the geo-spatial mineralisation attributes, such as: the lithological components and the occurrence of mineral deposits clustering arrangement, constituting a natural SAC due to their co-existence. The effect of SAC on spatial data distribution follows Tobler's Law, which states that object values close together tend to be more similar to each other than those farther apart.

The closeness of mineralisation attributes in space, as is the case in secondary cassiterite mineral deposits formation, leads to SAC among the predictive attribute values, such as elevation values or values of distances from one mineral location point to a certain geological attributes such as the rocks and other mineral location points. This similarity determines the predictive spatial attributes used to build PSM-MPM (Miller, 2004; Tobler, 1970). The SAC phenomenon is of great concern in this work and the choice of secondary mineral dataset perfectly presents the challenges of SAC on predictive spatial model performances. This is probably the reason why the local miners were prospecting within a certain distance of areas where deposits had already been discovered.

2.7.1 Detecting and Quantifying Spatial Autocorrelation

Before attempting to model distribution data that are affected by SAC, it is reasonable to consider the effect of the planned analysis with respect to SAC. Checking for SAC has become a conventional routine for plant ecologists and geographers (Fortin & Dale, 2005; Sokal & Oden, 1978). The commonly used methods of detecting the presence or absence of SAC include; Moran's $\mathcal{I}$ plots

(also termed Moran's $\mathcal{I}$ correlogram), Geary's C correlograms and semi-variograms (Isaaks & Srivastava, 1989; Legendre & Legendre, 1998; Perry *et al.*, 2002). A measure of similarity of data points (i and j) is plotted as a function of the distance between the points d_{ij}. In all the three methods mentioned (i.e., Moran's $\mathcal{I}$, Geary's C or variance or variogram) distances are usually grouped into bins. Moran's $\mathcal{I}$ based correlograms typically show a decrease from some level of SAC to a value of 0 or below; a value signifying the absence of SAC: The Geary C test is similar to Moran's $\mathcal{I}$ test in terms of conclusions but differs in interpretation. Geary's C has a mean value of 1 when testing the null hypothesis for no SAC and the values range between 0 and 2, i.e., it can never be below zero; values between 0 and 1 indicate *positive* SAC while values greater than 1 indicates *negative* SAC.

Moran's $\mathcal{I}$ is less sensitive to differences in small neighbourhoods, however, is a more global measurement compared to Geary's C that is more sensitive to extreme values. In general, Moran's $\mathcal{I}$ is preferred to Geary's C in most cases and consistently more powerful (Sawada, 2001).

Since the mineral distribution attribute data for this work are represented as points coordinates plotted on a map, their distribution pattern explicitly reveals the particular patterns that indicates the presence of SAC, for example; anisotropy or non-stationary of spatial autocorrelation (Fortin & Dale, 2005; Isaaks & Srivastava, 1989). The technique of Point Pattern Analysis (PPA) using quadrat analysis was used to determine the distribution of the mineral data points obtained from the PYGR. The quadrat method is very suitable for testing point distribution patterns and has been used by plant ecologists to study plant distribution pattern in a farm (Boots & Getis, 1988). The quadrat analysis of points was used to set hypothesis of spatial distribution, under *complete spatial randomness* (CSR) test as the null hypothesis for a random distribution test. Such a distribution is the test of random Poisson distribution. The test includes visualisation and analytical CSR tests to detect the distribution patterns of the secondary mineral distribution obtained from the PYGR area represented as points.

A statistical analysis test was used in Chapter 4 to determine the presence of SAC in the mineral data obtained form the PYGR using a CSR test. The point pattern maps consist of two major components: the points as object and the area in which the points is located. The points may be studied using the concept of

CSR to determine the properties of their distribution in space. The concept of CSR is based on the assumptions of the common conditions of uniformity and independence, which is that, each location of the mineral point has equal chance of receiving a point and the selection of one point does not affect the other (Diggle, 1983).

The first test of the statistical approach is the analysis of *complete spatial randomness* (CSR), where the hypothesis was tested to determines the distribution patterns of mineral data obtained from the PYGR. The CSR technique for point pattern analysis determined the distribution for non-randomness that can test for the presence of certain distribution pattern or visualised data structure using GIS.

2.7.2 Point Pattern Analysis of Mineral Occurrence

A point pattern represents a spatial pattern that constitutes arranged or organised points (Boots & Getis, 1988). The carefully arranged points represented on a map can be referred to as point pattern maps. These maps are commonly used to describe the occurrence of studied phenomena in a map. Although real-world objects are not points, they can be represented as points on a map for the purpose of analysis. The physical sizes of real objects are quite small compared to the distances between them and the area over which they occur. A good understanding of the study of point pattern maps may help in learning about the phenomenon presented as points and the process generating such points. It is equally possible, to build an explanatory model based on sufficient ideas concerning the events. Quite often, hypotheses concerning the location behaviour of the phenomena can be derived from such models. Here the distribution pattern is of great importance in revealing information about the dataset.

Typical examples of studies in PPA include central place theory that suggests that settlements are often regularly distributed over a region. The hypothesis can be supported showing the distribution of central areas. A similar consideration is applied in urban rents theories suggesting that individual occurrences of activities will repel each other, thus dispersing activities (e.g., retailing shopping malls) whereas other activities may attract, thus spatial aggregation, e.g., industrial and office activities.

2. REVIEW OF RELATED LITERATURE

The technique of PPA is a statistical analysis that arose over 50 years ago in plant ecology, as reviewed by (Greig-Smith, 1979) and later extended to animal and as well as plant ecology. The technique of PPA has been used to explore the spatial distribution of individual species and interrelationships of two or more species. The aim was to identify individual and environmental factors that influence such patterns (Connor & Simberloff, 1979; Simberloff & Connor, 1981).

The early 1960's, was the era of the *quantitative revolution* in geography, where the technique of PPA was introduced into geography for the refinement of previous qualitative descriptions, particularly of settlement patterns to predict how the theories of central place could be identified in the real world (Agarwal, 2007; Dacey & Tung, 1962). The central place theory makes efforts to explain the reasons behind the distribution patterns, size, and number of cities and towns around the world. (Agarwal, 2007). Soon afterwards Dacey (1964) developed and extended the models to produce alternative patterns to *central place theory:* particularly, models leading to clustered patterns of the settlement were emphasised (Boots & Getis, 1988; Dacey, 1964). PPA was later extended to the analysis of phenomena other than settlement patterns including retail establishments (Quigley, 1998; Rogers & Brown, 1974). Geographers and scientists have also used PPA to study the spatial characteristics of some physical features of the landscape, including volcanic craters (Kemmerly, 1982; McConnell & Horn, 1972). In this study, PPA is used for detecting distribution patterns of the secondary cassiterite mineral deposits and test for the presence of SAC in the predictive spatial dataset.

The examination of SAC regarding a particular phenomenon is of great concern to geographers for the identification of place in a city, town in a state or location of rock types in a given area. The occurrence of spatial information is displayed in form of a map. Three-dimensional real world data can be viewed as symbols in a two-dimensional plane with the help of cartography. The displayed symbols are usually represented as points, lines or area geometry. From geography, PPA has been adopted by archaeologists and anthropologists to study artefact distributions within a site (Hines *et al.*, 1993; Wilsher *et al.*, 1993).

Over the last decades, more sophisticated and wide-ranging techniques of PPA have been developed for spatial pattern distribution and analysis (Diggle, 1983;

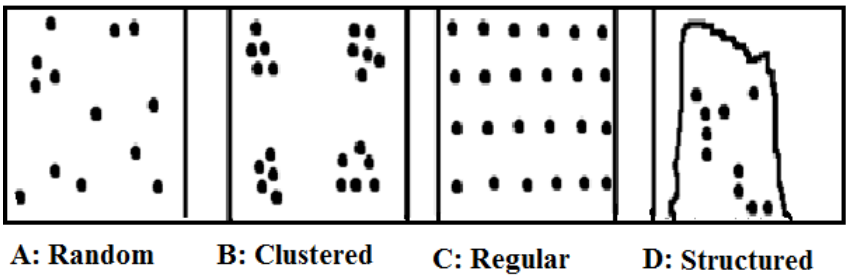


Figure 2.4: Different possible types of spatial distributions point patterns.

Ripley, 1991). Figure 2.4 shows different possible classes of point pattern distribution which cannot be determined by visualisation alone but through some statistical or spatial analysis. The statistical and spatial analysis are used to authenticate the visual analysis of the distribution pattern by providing quantitative or empirical proof to indicate the type of distribution in the dataset.

2.8 Machine Learning Classification of Spatial Data Distribution

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that uses a technique that provides computers with the ability to study a problem without being apparently programmed (Breiman, 2001). An ML classification method was used to build a predictive model that addresses the problem of mineral exploration or discovery (Ibrahim & Bennett, 2014a; Porwal, 2006). The ML classification modelling approach was chosen for this work due to its ability to automates the learning of hidden knowledge about natural spatial phenomena such as secondary mineral occurrence data (Ibrahim & Bennett, 2014b; Porwal, 2006). The ML classification algorithms can model mineral recognition criteria that form the basis for the presence or absence of mineral deposit occurrence, based on the existing or prior mineral deposit information, which are referred to as the predictive attributes. The dataset of already discovered mineral deposits comprises the mining points, some geological and geographical features split into training and test data for modelling and validation, respectively for ML classification.

ML generally deals with the design and the development of algorithms that allow computers to evolve behaviours or learn based on empirical data; it studies

computer programs that automatically progress through acquaintance or learning (Mitchell, 1997). The primary research direction of this work is to use ML classification to automate the extraction of information from secondary mineral distribution data through computational and statistical methods and to make intelligent prediction based on the trained data. In other words, ML classification used the power of computer and statistics to exploit intelligent decision about potential mineral deposits of the PYGR. No ML classification modelling has ever been conducted on secondary cassiterite mineral deposit, and especially in the PYGR; this research is therefore a pioneering research work (Ibrahim *et al.*, 2015a).

There are two possible applicable options in ML depending on the task at hand, of learning: inductive and deductive inference. Inductive machine learning extracts new knowledge from data that describes an experience in a form of learning examples or instances (Bratko, 2001). In contrast, deductive learning explains a given set of rules by using specific information from the data (Langley, 1996). Depending on the feedback received by the learner during the learning process, the learning can be classified as *supervised or unsupervised*. The focus of this work is strictly on supervised inductive ML also referred to as *predictive modelling* using ML technique for learning a function from the mineral dataset. There is a class associated with each example and an answer to a question about the example. It assumes that each learning sample includes some target property, and the goal is to learn and predict this property within certain level of accuracy. Example of supervised ML algorithms include: Naive Bayes (NB), Decision Tree (DT), Tree Bagging (TB), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) etc, (Bishop, 2006). On the other hand, unsupervised inductive ML, also called *descriptive modelling*, assumes no such target property to be predicted. It typically tries to uncover hidden regularities or patterns to detect anomalies in the data. In contrast, examples of machine learning methods for unsupervised ML include clustering, association rule mining etc., (Bishop, 2006).

A brief workings and capabilities of some selected three classification algorithms or classifier, used in details for critical predictive performance evaluation

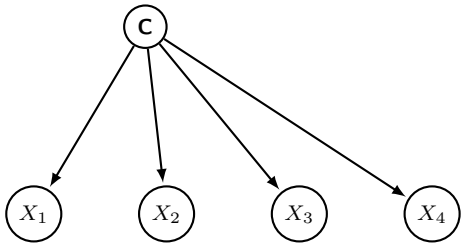


Figure 2.5: The structure of a simple NB diagram showing attributes and class nodes

during the course of modelling mineral potential of the PYGR using the NB, TB and KNN classifiers are as follows:

- Naive Bayes

The advancement of AI led to the development of intelligent Bayesian Network (BN) called Naive Bayes (NB) that is capable of inductive learning and generalization (Cheng & Greiner, 1999; Cooper & Herskovits, 1992; Hepinstall & Sader, 1997; Wang & Cheng, 2008). Figure 2.5 depicts a typical structure of ordinary NB with X_1, X_2, X_3 and X_4 nodes representing the independent predictive attributes while the class is represented by node C, and has shown better performance than most classifiers (Langley *et al.*, 1992). The Naive Bayes (NB) is a simple algorithm that has its parent node as its class and no further links in the structure (Duda *et al.*, 1995). The NB has an advantage over other classifiers because; it is easy to construct with a given priori as the structure so that no structure learning procedure is required. The process of classification using NB is very efficient with the advantage of independent assumption; meaning all predictive attributes are assumed to be independent of each other. The NB has performed better than many classifiers in many datasets, especially where the attribute datasets are less correlated (i.e., independent of features) (Langley *et al.*, 1992). It is also very tractable to statistical computation because the conditional probabilities are a measure of parameters of the inter-variable dependencies. Even though the BN is very effective for knowledge representation and inference under uncertainty, the BN was not regarded as a classifier until the discovery of the NB (Pearl, 1988).

- K-Nearest Neighbour

KNN is an exquisite learning algorithm that is known to perform very well with spatial distribution data (Ibrahim & Bennett, 2014a). The KNN uses a system of voting in classification by ranking the feature vectors according to Euclidean distance and selecting the k-vectors with the smallest distance to each point. The KNN algorithm possesses some unique characteristics that include; being very fast in training and testing models, it is simple and performs well in moderate dimensions, it is also a lazy strategist, devoting most efforts at prediction time and zero effort at training time, empirical error on the training set is always zero and only needs to distinguish between the empirical and predictive error (generalization).

- Tree-Bagging

The TB algorithm is an improved version of the Decision Tree (DT). A DT select attributes or predictors by classifying them according to their values; they describe a feature represented as nodes in observation to be assigned, and each branch determines the value that such node can assume. Observations are classified starting at the root node and sorted based on their feature values. The attributes that best separate the training data would be the root node of the tree (Breiman *et al.*, 1984; Hunt *et al.*, 1966). TB is a better performer to the DT. It uses a method of ensemble where the tree is grown on an independently drawn bootstrap replica. Bagging stands for bootstrap aggregation. TB takes an average of the predictions from individual trees to compute a prediction of an ensemble of trees for unseen data. It has performed very well with spatial data and other data similar to the random forest (Breiman, 2001; Pardos & Heffernan, 2010).

Among the seven sampled standard ML classifiers employed for this experiment, only three which include KNN, TB and NB were considered for further performance evaluation, to investigate the causes of overfitting and underfitting in the dataset as explained above. The remaining four classifiers were not considered since thy were only used for the purpose of performance comparison among the range of sampled classifiers. The work used ML classification to build a PSM-MPM and selected the best predictive model developed by each classifier

to evaluate the predictive performance in order to enhance the predictive accuracies through a systematic approach of predictive model performance evaluation or validation.

2.8.1 Predictive Model Validation

Predictive spatial models with ML classification, need to be validated or tested to ascertain its accuracy or efficiency to generalisation (i.e., performance in respect to independent or new datasets). This validation is necessary because it is a way of getting feedback on the level of model performance that determines its usefulness. The result of testing may lead to possible redesign, optimisation or even outright rejection based on poor performance. It is believed that modelling an elaborate ML algorithm that is complex and produces non-linear class boundaries is a better classifier than straightforward and linear models (Danso, 2006). Duda *et al.* (2012) later stated, however, that complex models over-fit the training data by giving a higher predictive accuracy performance but still performing badly when tested on new datasets despite the higher performance with training data.

Various methods exist for validating models such as cross-validation, stratified RHO selection and also re-substitution (Bradley, 1997; Han *et al.*, 2011). These involve the splitting of predictive data into training and testing sets by selecting the classifier that generalises well, giving the best predictive accuracy or least predictive error rate. The model performance validation also helps to solve the problem of overfitting commonly associated with ML classification on spatial distribution dataset, which may result in an over-exaggerated predictive accuracy. The result of an overfitted algorithm on a data is leads to an overly pessimistic ultimate results. Once a spatial predictive model satisfactorily passed validation stage, it is retained as established and acceptable predictive models for mineral deposit potential mapping for the given area. Otherwise, the modelling process will be repeated or subjected to other forms of evaluation, such as investigating the attribute data type used, consider a change of classifier or using model performance validation technique like re-sampling to optimise model performance. Note that validation can be done both internally and externally (i.e., a similar dataset

from the same area or a similar dataset from different areas for training and testing, respectively). In the case of internal validation, datasets used for training the parameters will be divided into a particular ratio with the smaller ratio being the testing or validation set (Kohavi, 1995). Otherwise, the data can be generated externally through simulation of a synthetic dataset or from an entirely different data source but same attributes.

- Confusion matrix table:

The confusion matrices as presented in Table 2.1 shows how are the result of prediction or model test results using ML algorithms. The ability of each algorithm to accurately predict correctly a specific class or not ia indicated by the confusion matrix . The diagonal values are the correctly classified prediction while others are misclassified. In this experiment, the class order is 0 and 1. Since the research intention is to predict new mineral deposits, there is the need to know places where time, energy and resources needs to be expended when searching for potential mineral deposits. Hence, both classes (0 and 1) are important. The aim of the confusion matrix is to determine the actual number of points that are classified correctly and those not correctly classified by the algorithm based on the test data.

Table 2.1: A typical interpretation of standard ML classification confusion matrix

	predicted (0)	predicted (1)
Actual (0)	True (0 or positive)	False (0 or negative)
Actual (1)	False (1 or positive)	True (1 or negative)

- Predictive performance table:

Predictive accuracy is the measure of the overall predictive ability of the algorithm accurately to detect both positives and negatives (true class and false class). It is usually calculated as the average sum of sensitivity and specificity. Sensitivity measures the rate of true positives against the false negatives rate and represents the ability to predict the zero class correctly based on the confusion matrix presented in Table **??**. Specificity measures the rate of actual negatives against the false positive rate and represents the ability to predict the ones (1) class correctly

based on the confusion matrix in Table **??**. The class order matters here since we are using a class order of 0 and 1, which implies non-mineralised and mineralised points respectively, based on the ground truth.

The formulae for calculating model performance indices such as accuracy, sensitivity and specificity are given as:

$$Accuracy = \frac{\sum True(Positive) + \sum True(Negative)}{\sum Total\ Population}$$

$$Sensitivity = \frac{\sum True(Positive)}{\sum True(Positive + \sum False(Negative)}$$

$$Specificity = \frac{\sum True(Negative)}{\sum True(Negative) + \sum False(Positive)}$$

- ROC and AU-ROC curve plot:

The Area Under the Receive Operating Characteristics (AU–ROC) curve defines the measure of accuracy of a predictive test. The larger the area under the ROC, the more accurate the predictive test is. The AU–ROC curve is measured by the following equation:

$$AUC = \int_0^1 ROC(t)dt \tag{2.1}$$

Where;

$$t = 1 - Specificity \tag{2.2}$$

and;

ROC(t) is Sensitivity

2.8.2 Predictive Model Selection

Various methods exist for selecting "optimal" models in ML classification. These involve the performance validation and evaluation success recorded by a classifier on specific datasets which signifies the generalisation capability of the model as measured through a high predictive accuracy rate (Mwitondi *et al.*, 2013). The

conventional method is to, first of all, view the confusion matrices and then determine the models with the highest predictive accuracy when validated against a test set (Mwitondi & Said, 2013). Accuracy alone, however, may not be enough to conclude explicitly that a model is performing very well, since some models exhibit what is referred to as accuracy paradox i.e., a situation where a predictive model with high accuracy may still lack greater predictive power. This is because greater accuracy only implies an evaluation of the rate of correct prediction to a particular class but does not say much about the robustness of the model — i.e., models that are sensitive to noise by reducing the possibility of fitting noise but generalise well with new and unseen data. This is often evident in a predictive accuracy scores in spatially autocorrelated attribute datasets such as in the secondary mineral distribution data, since the SAC leads to overfitting and underfitting, respectively, when all dependent attributes are trained and tested on almost similar datasets or trained and tested on entirely dissimilar datasets.

Model evaluation includes testing the predictive efficacy, error rate, sensitivity or specificity, and the size of the area under the ROC as part of the assessment when compared to other classifiers. Comparative analysis is further conducted to select the best classifier based on some multiple options from among the criteria listed. A further test of data re-sampling that involve test of true attribute data independence may e conducted to test for overfitting or underfitting due to SAC particularly in spatially distributed dataset.

2.9 Summary

The chapter identified and discussed the various approaches to modelling and predicting potential mineral deposit, using different geostatistical and data mining techniques (Carranza *et al.*, 1999). The statistical approach includes the method of geospatial kriging and the WOE to model mineral potential using the available digitised mineral data sets or maps. Methods for capturing spatial data for useful geospatial analysis of both digital and analogue point data only, using ML classification have hitherto not been very successful. However, attempts to build a predictive model for mineral data and use in an embedded system or expert system have also not been successful, due to the lack of ability of the existing models

to generalise well with unknown or unseen data. The evolution of GIS and ML, however, has significantly improved methods of capturing spatial data, primarily geographic and geological data, from both analogue (cartography) and digital maps (Bonham-Carter, 1994; Ibrahim & Bennett, 2014b) for easier computational and modelling with generalisation.

In this chapter, a method of predicting secondary mineral potential deposits obtained from the PYGR area of Nigeria was proposed using ML classification technique. The proposed model approach seek to overcome the problem of secondary mineral data paucity, that involved a systematic data acquisition technique using statistics, spatial data manipulation analysis with GIS and also, the deployment of ML classification to build a predictive spatial model also referred to as PSM-MPM. The PSM-MPM will be capable of predicting the potential location of the mineral deposit in the PYGR area, and other places, based on the existing mineralisation attributes obtained in the area and elsewhere. The spatial distribution pattern of the mineral datasets represented as points indicated that the attributes data values are mostly dependent on each other due to their closeness (spatially) or may be clustered together leading to lost of attribute data independence. The lack of true attribute data independence leads to high SAC in the dataset leading to either overfitting or underfitting by the ML classifier.

Finally, the chapter discusses the application of ML classification modelling as the state of the art in numerous research areas including using spatial distribution data such as the secondary mineral data. The ML classification offers a unique method of modelling spatial and non-spatial attribute data by automating the statistical process of building models and have the ability to measure generalisation or the predictive performance based on unseen dataset through validation. The current and traditional methods of model validation or cross-validation technique of evaluating the problem of overfitting or underfitting in predictive modelling are through the validation of training performance on the test dataset. This method has not been examined properly in a spatially autocorrelated dataset such as secondary mineral data distribution dataset. A modelling technique for building a PSM-MPM to test and detect the detrimental effect of SAC to overfitting and underfitting by the ML classifier measuring performance has been proposed through model validation. Existing standard methods of evaluating and selecting the ideal

PSM-MPM based on performance, such as the predictive accuracy scores of the classifier. The evaluation will involve effective re-sampling that involves splitting training and test set to generalise or improve the predictive model well.

Chapter 3

Methodology

3.1 Overview of the Chapter

This chapter describes the research methodology adopted for this research by systematically implementing the designed steps that assume the use of geographic information system (GIS), statistics and machine learning (ML) to analyse secondary mineral distribution datasets obtained from plateau younger granite region (PYGR) and use ML classification to build a predictive spatial model for mineral potential mapping (PSM-MPM). Section 3.1 is the general overview of the chapter while 3.2 contain the introduction. Section 3.3 explains the method of data source and the mode of collection with justification. Section 3.4 discusses the method of data collection; which involves the process of physical capture of the mineral distribution occurrence data points to undergo the cleaning, conversion and identifying all predictive attribute data for use by ML classification technique. Section 3.5 highlights the statistical and spatial analysis of secondary mineral datasets, as well as determining the correlation among the selected predictive attributes datasets. It also examines the distribution patterns of secondary mineral data points. Section 3.6 discusses the actual design development method of the PSM-MPM using ML classification algorithms and select some of the classifiers based on performance for further evaluations. Section 3.7 Validates the use of spatial attributes and spatial autocorrelation (SAC) or spatial distribution through data simulation. Section 3.8 shows a method of data pre-processing using PCA that leads to the selection of attribute subset, to optimise the predictive performance scores of the

selected classifiers. Section 3.9 represents the method of performance evaluation of PSM-MPM with a particular focus on overfitting and underfitting. A novel assessment technique that challenges the existing standard method of model performance evaluation but specifically applicable to spatial distribution datasets. Section 3.10 explain the model performance comparative analysis method, between the novel technique of RHO, PCA-RHO and SSS to justify the method of SSS as the ideal method to use in validating PSM-MPM to overfitting and underfitting due to SAC, through their respective predictive accuracy scores. Finally, Section 3.11 summarises the general findings in the chapter.

3.2 Introduction

Table 3.1 depicts the comprehensive methodology adopted for the conduct of this research work. The general method adopted is in four (4) stages. The first two steps are similar to the process adopted by Bonham-Carter (1994) where mineral deposit data maps were obtained from the USGS in a digitised format into GIS and analysed to produce a mineral potential map of the target area. The work of Bonham-Carter (1994) used statistical tools in GIS to classify areas on the map with mineral potential. The method used by Bonham-Carter (1994) involved digital data map collection from individual data repositories. The geoprocessing and modelling of mineral potential based on evidential weight using GIS. The method varies significantly compared to the work in this thesis and is considered an extension of the method employed by Bonham-Carter (1994).

This extension of the work is seen as both major and minor aspects. The minor expansion involves the type of data used, a method of data acquisition and the kind of data analysis done at various stages of mineral data transformation and classification. The major extension is in the third and fourth stages of the methodology as indicated in Table 3.1. Specifically, the application of standard classification in machine learning (ML) to mineral deposit potential based on secondary mineral data that produces PSM-MPM, and the evaluation of its predictive performance through a novel technique of *spatial data separation* that allows data independence as well as improve model validation procedure that can more reliably detect overfitting in the presence of SAC. The steps mentioned also involves

the use of different standard ML classification algorithms to develop a predictive model called PSM-MPM that serves to evaluate the predictive performance affected by model overfitting due to data dimension and spatial autocorrelation (SAC). Other contributions highlighted in the methodology includes the technique of obtaining and converting manual or analogue mineralisation geo-datasets to digital. Recalled that part of the challenges mentioned earlier about this research faces is in the area of data availability and therefore required a systematic approach to a credible but scientific data acquisition that can be used for the type of work intended.

The steps involve in the methodology are summarised as follows:

- Step One:

The first step involves the selection and justification of the study area; it also includes the process of data collection. The procedure includes obtaining information about the mineralisation components of the area. The data collection process in the PYGR area was systematically done to address a problem of data paucity in the area due to the pioneering nature of this research. The systematic approach involves a geological mining data survey of the PYGR area to collect mineralisation geodata such as: the mineral occurrence coordinate positions in latitude and longitude using a Global Positioning System (GPS) tool, the elevation of the mineral occurrence and names of the settlements where the mining areas are found. Other mineralisation data collected outside the survey included a scanned copy of the cartographic geological map of PYGR area showing different rocks types. The entire dataset collected from both the physical and geological surveys in an analogue format including the cartographic (scanned) geological map of PYGR carefully converted to digital to prepare the data in an ML classification acceptable format to build a PSM-MPM for the PYGR area and other places.

The method and type of geospatial data acquisition, mineralisation attribute data extraction and geoprocessing of all attribute map data were prepared as input to the Geographic Information System (GIS) through a standard geo-referencing of various predictive map layers. The standard geo-referencing provides a platform that allowed maps drawn from different projections to be aligned with a single scaling or projection to enable easy analysis and visualisation in GIS or using

Arc-GIS software. Other procedures at this stage involve the combination of the geological map layers consisting of lithological (rocks type), digital elevation data map (DEM map) and another geo-information maps in GIS. The maps of mineral occurrence (presence and absence) data points representing the position of mineral deposits and other components of the map, such as settlements, are spatially joint with other maps of different rock types and the elevation in GIS. The created predictive maps are all combined into single map converted to *shape file* and determine the the nearest distances between each attribute and mineral occurrence points.

- Step Two:

The second stage involves the statistical analysis of the mineral point data and geospatial analysis of all the predictive attributes in the geological map layers in GIS. The activities include creating spatial evidence geological maps and conducting spatial analysis of all the selected predictive components or attributes in the map layers. The selection of mineral attributes or recognition data was carried out based on literature about mineralisation processes through geospatial or statistical data analysis, to establish correlation among the attributes. The procedures involve spatial analysis to extract features that capture the mineralisation distribution pattern of the given area as well as determine the presence of spatial correlation among the geological attributes. The spatial relationship between geological attributes, such as granite rocks and mineral occurrence points, was done by conducting spatial analysis of distance between mineral occurrence points and geological features.

While the quadrat analysis measures the distribution patterns of the mineral occurrence represented by points on the map, the distance distribution analysis test determines the spatial correlation, or relationship, among the mineral indicators, or mineralisation attributes, using the Kolmogorov-Smirnov Test or K-S Test. The K–S test statistic is a concept that determines the level of correlation between two entities to determine if two distribution datasets are similar or differs significantly. The two datasets are the data distribution points for mineral presence and mineral absence points measured in relation to geological features of interest. In this case, the target is the closest of different rock types (represented

by polygons) as contained in the geological map of the PYGR. The spatial predictive attributes of the study area data involved in the statistical and geospatial analysis at this stage include: mineral concurrence points (i.e., mineral presence or absence) locations; the lithology (i.e., favourable host granites or igneous rocks type) and the desirable distance between mineral occurrence points and favourable host rock.

- Step Three:

The third phase is a part of the contribution of this work which is the design and development of a PSM-MPM capable of predicting the mineral potential of the PYGR area, and other places, based on the geological and geospatial attribute data using ML classification algorithms. The procedure involves combining spatial characteristics and evidential predictive parameters using some selected standard supervised ML classifiers to capture the mineral distribution pattern of the PYGR data and make predictions of areas with potential deposits. The PSM-MPM validation based on traditional random hold-out splitting technique was subjected to further evaluation, to observe the effect of either overfitting or underfitting caused by SAC, on the performance of the classifier. The model validation procedure introduced a novel approach to secondary mineral data validation technique called *Spatial Strip Splitting* (SSS) validation, where the splitting of attribute data for validation of the PSM-MPM is conducted spatially, by reducing the attribute data dependence caused by SAC. The spatial sampling of attributes data by splitting into training and test data results provides an answer to the detrimental effect of overfitting and underfitting caused by SAC to provide a more optimistic predictive model performance accuracy by the classifiers. The predictive performance of each classifier was evaluated by comparing the individual predictive accuracy score results obtained using both random hold-out validation selection and the SSS validation methods to identify an ideal approach to validate the PSM-MPM performance.

The importance of spatial components of the dataset that causes SAC were also investigated when building a PSM-MPM; the technique showed a comparison of predictive accuracy scores among methods that deliberately eliminate spatial attributes, one with the spatial attributes only and then the other with both. The

results also highlighted the importance of spatial components and the need to include such components in the model building rather than eliminating them. A further contribution to this work in this section is identification of the importance of *spatial attributes data* in PSM-MPM. The procedure involves simulation of the mineral distribution of the predictive data from the PYGR that deliberately eliminates the spatial components of the data and uses the resultant simulated data as a test set to validate the result of the PSM-MPM produced by the real data. The result of the investigation signified the importance of the spatial component for good predictive outcomes in PSM-MPM. Predictive accuracy scores may be affected by the presence of SAC in the attribute data to determine whether to accept or reject a model, this is because, despite the importance of spatial attributes and SAC in the predictive dataset, they are still prone to overfitting and underfitting which remains a sorce of concern to be address.

- Step Four:

The fourth step involves PSM-MPM performance evaluation: activities in this step include a four-way model performance validation technique of data re-sampling that involves the splitting of training and test dataset. The PSM-MPM four-way performance evaluation involve the test of performance on the testing dataset using re-substitution, traditional RHO, half longitudinal split and the unique quartered longitudinal spatial strip split (SSS). The SSS technique was considered to show clearly the effect of SAC on performance of an ML classifiers through the predictive accuracy change because, it allows for more data independent and increase heterogeneity of attribute data spread among test and training data better in the face of overfitting and underfitting.

A comparative analysis between three ML preprocessing technique was conducted at this stage to determine the best approach that distinguish between accuracy and optimisation that handle the detrimental effect of SAC leading to overfitting and underfitting. The first technique is the standard RHO without any preprocessing, then followed by RHO that involve preprocessing which involves the reduction of attribute data dimension using Principal Component Analysis (PCA) that selects the most important attribute data for model performance optimisation. The result of these three methods evaluations that include standard

RHO, PCA-RHO and the SSS will determines the best method that determines the effect of SAC (i.e., negative and positive effect) in secondary mineral deposits; by comparing the predictive performance of the three different approach of model validation in an ML classification. The selection of an ideal technique was based on the predictive performance that offers a more optimistic predictive accuracy scores rather than a pessimistic or poor predictive performance scores.

A detailed breakdown of how these four (4) steps contained in the methodology Table 3.1 as carried out is provided in the subsequent sections and subsections below:

3. METHODOLOGY

Table 3.1: A tabular structure of the methodology adopted for the thesis.

GENERAL STEPS	ACTIVITIES INVOLVED	PROCEDURES/PRODUCTS
1. **STUDY AREA SELECTION AND JUSTIFICATION** (DATA COLLECTION AND TRANSFORMATION)	SURVEY OF EXISTING MINING LOCATIONS and CARTOGRAPHIC GEOLOGICAL MAP OF PYGR OBTAINED FROM NGS	Using Global Positioning System (GPS) tools to obtain the coordinate points of mining locations within the PYGR area where minerals are present and where they are absent; the coordinate location points include the longitude, latitude and elevation point of each location of mineral occurrence using some population of occurrence density based on the number of points in a particular area and recorded in an Excel format
2. **SPATIAL DATA ANALYSIS TO BUILD SPATIAL PREDICTIVE ATTRIBUTE DATABASE/STATISTICAL ANALYSIS** (CONDUCTED IN and OUTSIDE GIS)	COMBINE SPATIAL EVIDENCE MAPS TO EXTRACT SPATIAL BINARY ATTRIBUTES	LITHOLOGY (Digitized Lithological Map) shape files) / STRUCTURES (Digitized from shaded-relief Map) shape files) / TOPOGRAPHY (Geo-referenced elevation contours DEM) shape files)
	POINT PATTERN ANALYSIS; COMPLETE SPATIAL RANDOM TEST AND K-S TEST	Determine distribution pattern of mineral point data using PPA Quadrat test / Quantify spatial association or correlation of mineral points with known mineral indicators(rocks types) / Determine the presence of SAC in the dataset distribution
3. **PSM-MPM CONDUCTED USING ML CLASSIFICATION WITH MATLAB, R AND WEKA** (RESEARCH CONTRIBUTION)	SELECT SOME STANDARD ML CLASSIFIERS	Develop PSM-MPM from standard supervised ML classifiers/algorithms (using secondary mineral distribution data obtained from PYGR) combined with spatial predictive evidence parameters and select the classifiers with highest and least predictive score to examine presence of overfitting and underfitting respectively
4. **MODEL PERFORMANCE EVALUATION** (CONTRIBUTION)	NOVEL TECHNIQUE FOR TESTING/VALIDATING PSM-MPM	The novel approach to model evaluation and selection through spatial strip splitting of predictive attributes data as against the traditional random hold-out method of model validation, that ensures data independence, often violated due to spatial auto-correlation in secondary mineral distribution datasets.
	PREDICTIVE ACCURACY EVALUATION AND PERFORMANCE OPTIMISATION	Comparative analysis of various predictive accuracy produced using ordinary standard ML random holdout (RHO) validation, attributes data pre-processing of best subset selection using PCA-RHO and the SSS validation approach to determine the ideal and optimised PSM-MPM as a contribution to the research work.

3.3 The Study Area and Justification

The Plateau Younger Granite Region (PYGR) is a component of the larger Nigeria Younger Granite Region that constitutes an igneous province of the best examples of mid-plate magmatism in the world, mainly due to the presence of aluminous biotite granites that are the source of rich alluvial tin and columbite (Pastor & Turaki, 1985). The area lies between Latitude 9 00'00" N to 10 30'00"N and Longitude 8 30'00"E to 9 30'00"E (Ibrahim & Bennett, 2014b; Pastor & Turaki, 1985). The tin deposits, or cassiterite, which formed the basis for the Nigerian tin mining industry are often secondary alluvial deposits of which formation has been explained in chapter two of the thesis.

The justification for selecting the PYGR area is mainly due to the fact that; the Jos Plateau areas constitute the central part of Nigerian Younger Granite Province with the following distinctive features:

- It has the most intensive occurrences of alluvial Tin deposits within the province.
- More than 90% of the Tin mining activities in Nigeria were done in this area. Hence, it has the largest known mineral occurrences of the province, against the production from the nearby younger granite rocks in Bauchi, Nasarawa and Kano States. This made it a suitable choice for secondary mineral distribution data target for this research.
- The geological map of the study area represented by Figure 3.1 with the geological features controlling the formation of the alluvial secondary tin mineral deposit, is available at an appropriate scale for the study. The map was obtained from the Nigeria Geological Survey office in Nigeria. The map was drawn at a scale of 1 mm to 0.5 km of the area of approximately 16,650km^2 as shown in Figures 3.1

3. METHODOLOGY

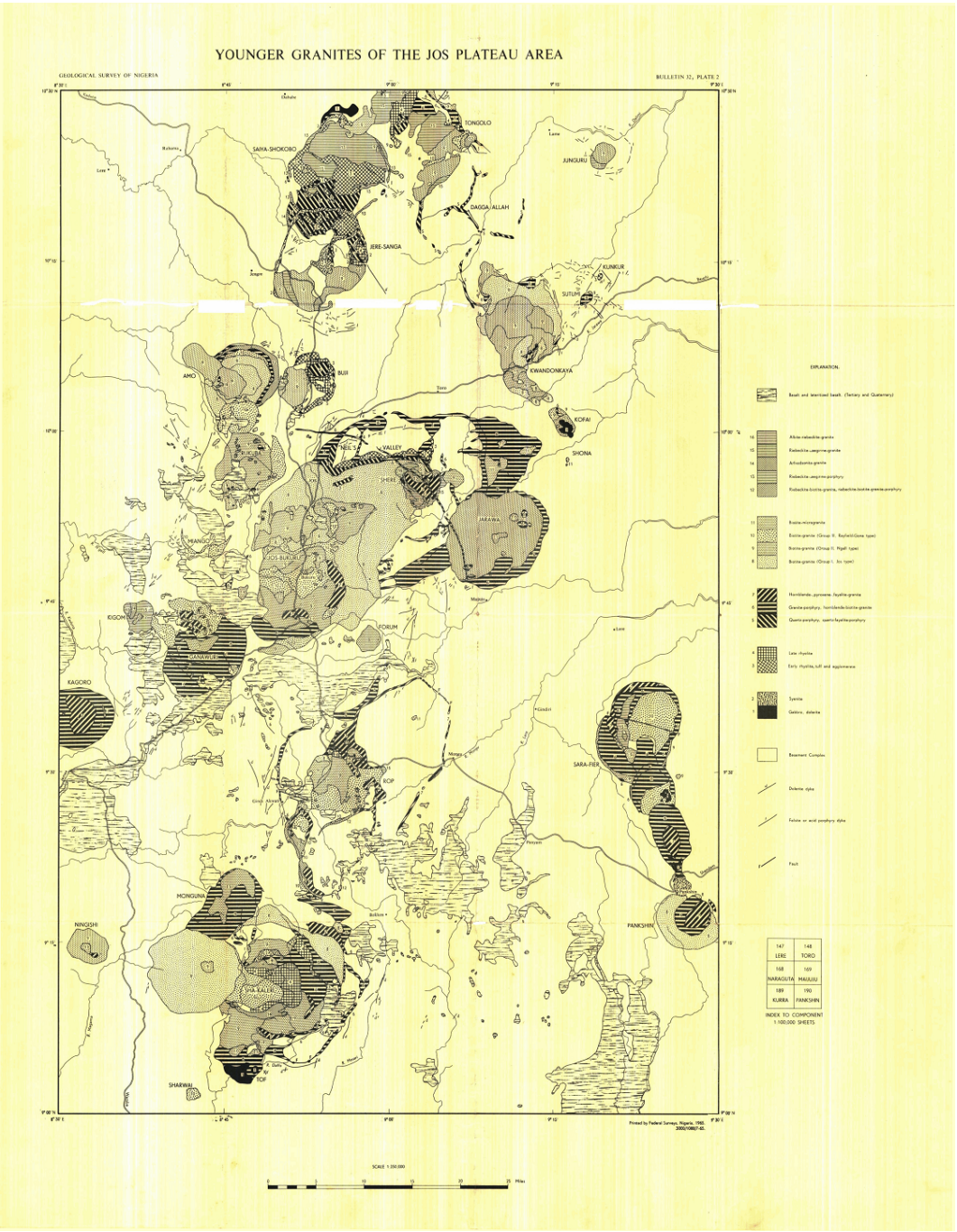


Figure 3.1: Cartographic geological map of Plateau Younger Granite showing only the geological settings obtained from Nigerian Geological Survey.

3.4 Data Collection

The primary datasets were obtained by conducting a field survey of geological data in the PYGR area, to identify areas where mineral deposits had occurred in areas where mining activities are present. The surveyed area corresponds to the area covered by the available geological map of the PYGR area, as shown in Figure 3.1. The survey was conducted in December 2012 covered the PYGR area, to record all existing mining sites by visiting locations where minerals are found and where they are absence. This locations were measured by recording the actual coordinate locations measured (i.e., latitude, longitude and elevation position in metres). The survey team divided the mining district into eight mining areas for ease of identifying of mining sites. Each of these districts contains several dozen mining locations.

SURVEY OF TIN MINING SITES AROUND JOS PLATEAU, NIGERIA
DATA COLLECTION SHEET

S/No	Mine site code	Latitude	Longitude	Elevation above sea level in meters	Status of mine	Remark

Figure 3.2: A geological mining survey template form used for mining point data collection of the PYGR

A meticulous method of data collection was adopted in the survey of all the existing mining sites of the PYGR. The study used a carefully designed template sheet as indicated in Figure 3.2, to record all geo-information of mineralisation datasets, for the purpose of modelling that predict areas with mineral deposit potential. Other benefits of the survey include addressing some knowledge gaps

that exist regarding the lack of comprehensive mineral occurrence data in over 100 years of cassiterite (tin) mining activities around the Jos GYGR area. Although the knowledge gap which is a result of a failure by both the government and geo-scientists to adequately document the locations of the past and current tin mining sites, especially using the accurate technique of GPS, the result of the survey will successfully bridge this knowledge gap and allow for a meaningful use of such data for scientific research such as, the one used for building PSM-MPM of the area using ML classification.

The survey team conducted their work by visiting all past and current mining sites and captured the mining locations as coordinate points in the PYGR. The points were chosen according to the density of mineral occurrence or frequency. Places or points with a larger number of mining sites have a denser sampling interval. The survey team was headed by an expert geologist from the Department of Geology, Ahmadu Bello University Zaria in Nigeria. The team leader was very familiar with the terrain of the PYGR which is a tough terrain. Indigenous miners were employed to assist in carrying out this survey. The local miners served as guides to locate old and new mineral/mining sites in the area. To ensure professionalism in acquiring data uniformity and fewer erroneous records, every member of the team utilised a single data capturing template. The survey was conducted over a period of 21 days in December 2012. A counter check procedure was introduced to double check the collected data using Excel tools to ensure proper compliance with the set objective of collecting the right data as indicated in the data collection sheet. The survey identified and recorded the following geo-dataset:

- Latitude, longitude and elevation of the sites was measured (the location coordinate point representing the location of mining).
- The name of the settlement where the mining activity was performed.
- Ancillary information about the mining sites such as size and status.
- The density values of data sampling depending on the number of mining sites identified in a given district –i.e., areas with a larger number of mining site will have a denser sampling interval.

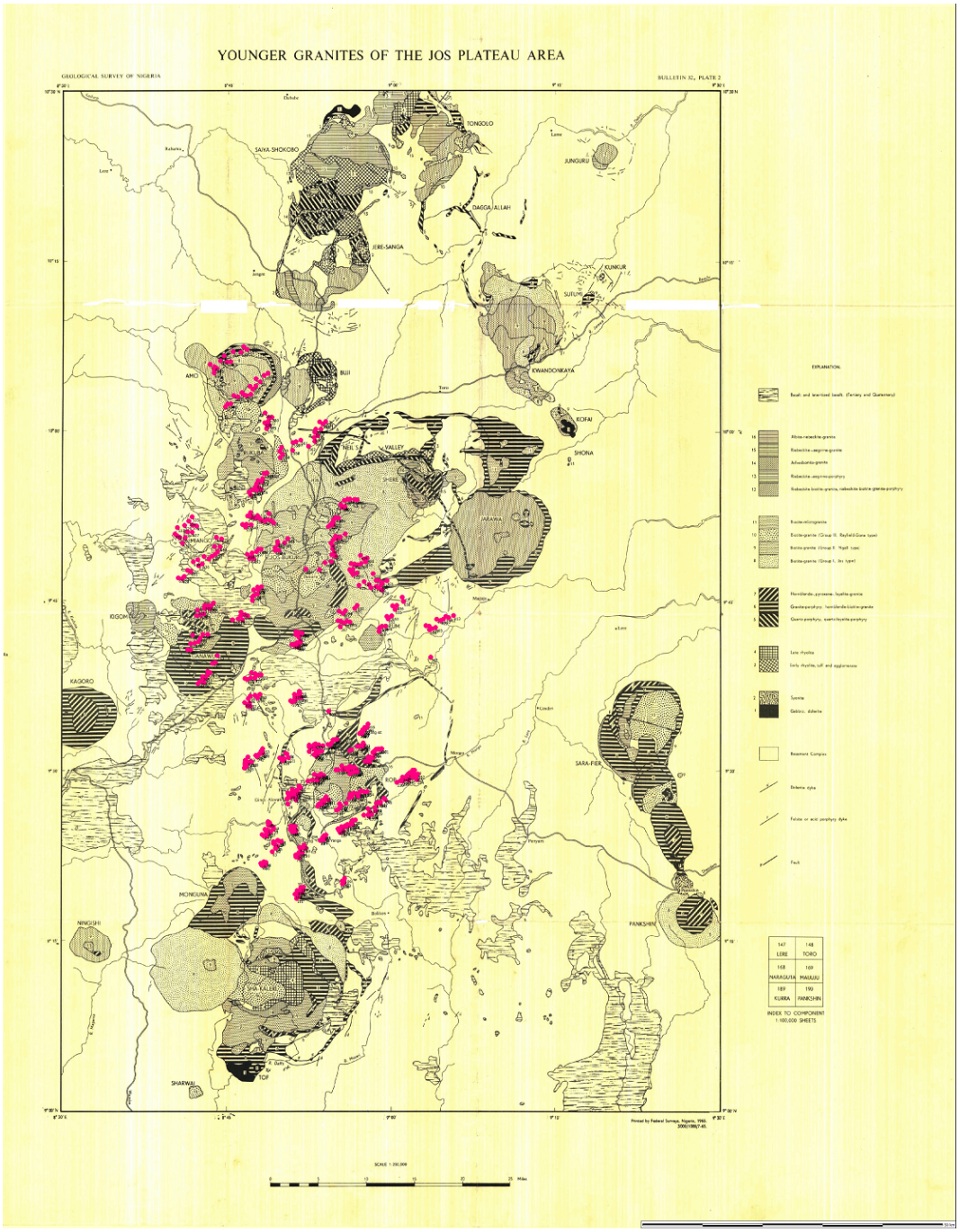


Figure 3.3: Geological map of PYGR showing 749 surveyed mining coordinate points obtained from field survey of the PYGR area.

3.4.1 Mineralisation Attribute Data Selection

There is always a need to identify the intention to which data is being solicited before obtaining such data. Recalling that one of the aims of this research was to develop a PSM-MPM that can predict the mineral deposits based on existing mining activities to show the occurrence of such mineral deposits in a given location. A conceptual model for mineral deposits based on the literature and other physical experience was used to identify and obtain all the mineralisation attributes of the given area based on the geological settings of the area.

A total 749 data points represented the entire observed mineral points; 463 were labelled as mineralised, 286 points were non-mineralised with a total of 21 predictive attributes being recorded consisting of the following: nearest closest distances from each observed points to all the fifteen (15) rock types contained in the geological map of PYGR was presented in the order; DR1, DR2, ..., DR15, elevation, slope values at each marked point, weighted value assigned to the most nearest distance between the rocks and the observed points as NDRPW; the weight is assigned probability values between 0 and 1, with the shortest distances getting higher weight score been most likely the source of deposits. While the farther nearest distances from the rocks get lower probability scores, indicating the most unlikelihood of the rocks type been the source of mineral deposit. For example, a zero distance gets the maximum weight score of 1 value while the highest distance gets a value of 0. Other predictive attributes include measurement of the rock sizes (AreaR) and the rock perimeters (RPerimeter). The mineralisation attributes derived were however, considered to be either spatial or non-spatial geo-data attributes. Detailed mineralisation attribute data selected and their justifications are as follows:

- The geological map of PYGR is a cartographic map originally produced in 1965 by the Nigeria Geological Survey (NGS). The justification for choosing this map for this work was because, although it has no information about the mineral locations and it is in analogue form as indicated in Figure 3.1. It has the structure of all the younger granite rocks and other rock layers, such that when populated through geo-referencing by the surveyed mineral data location coordinate points (i.e., latitude/longitude), it can be joined

through map digitisation easily, using ArcGIS. Besides it is the only available map drawn at an appropriate scale and projection, suitable for proper geo-processing of mineral point's data.

- The Digital elevation model (DEM) constitutes the elevation, which is the height of the earth's surface (i.e., above sea level) of the position of the mineral data point measured in metres. The elevation attribute is essential as mineral deposit movements are from places of higher altitude to a lower altitude. Attributes values derived from DEM includes values such as the slope, hill shade and curvature at each mineralised and non-mineralised point.

- The lithology (rock layers) component is the geological characteristics of rocks found in the area of interest, which include size, circumference and (or) perimeter of the rocks. A total number of fifteen rock layers contained in the map of the PYGR area were collected and recorded. The rock layers are important as they constitute the source of the mineral deposit.

- Longitude and latitude are the actual coordinates (x,y) points of the mineral occurrence on the earth's surface measured in degrees. Note that the coordinate points represent an area of the PYGR where minerals are present or absent, represented as points on the map of the PYGR.

- The nearest distances from the mineral data points to all mineralisation (geographic or geological) components in the area, measured in metres. The nearest distances of the occurrence of a mineral point to each rock types measures the proximity of a possible source of the mineral to the point of the deposit. The distances are necessary to determine which rock type is the likely source of mineral deposits. Since there are 15 rocks type in the PYGR, the spatial distance between each data point to the nearest rock type will be a 15 by 749 attributes data values. The spatial component of distances was obtained using a spatial analysis tool in GIS, and that constitute the major components of the spatial attributes.

In order to represent the mineral occurrence deposits of the PYGR fully for any meaningful scientific research such as this, the following data obtained are summarised based on their type and quantity in Table 3.2 below:

Table 3.2: Datasets used in the experiments

Dataset	Type	Number
Mining Observation Points	Point	749
Number of Rock Layers	Polygon	204
Rock Types	Polygon	15
Predictive attributes	Numeric/Real	22
1965 Map of PYGR	Cartographic Map	1
1975 LandSat Data Map of PYGR	LandSat	1
Map of PYGR	Points and Shape File	1
SRTM–DEM	Elevation Map	1
Class	Binary	2
Study Area	PYGR Nigeria	400km^2
Year of Data Survey	December 2012	N/A

3.4.2 Geo-processing of Mineral Deposit Data Points

The section involves the procedure for the preparation, cleaning and conversion of manually obtained geo-data from the geological survey field and the geological map of the PYGR using GIS. The procedure involves the following steps:

- Spatial data formation

The method consists of a careful uploading and digitization of the scanned copy of the cartographic map of the PYGR, as shown in Figure 3.1, and all the spatial information about the mineralisation features, which include the location of mineralised and non-mineralised points. Plotting the latitude and longitude coordinates along with the topology onto the digitised geological map of the PYGR, as shown in Figure 3.3, contains geo-data (geographic and geological) like shapes or polygons (rocks), and combining the relationship with the mining points obtained from the field survey. The transformation of data was done in two categories: firstly, the format conversion and geometric conversion. Format conversion is very time-consuming and was carried out using GIS to transform all the digitised

data into an acceptable GIS format, which includes mineralisation data points converted from rasters to vectors alongside the topology on the geological map of PYGR. It often used the Universal Transverse Mercator (UTM) system units in metres for such conversion. Vector data always requires topology to be included alongside the coordinate data such as latitude and longitude. On the other hand, the geometric transformation overlays maps of the coordinate data with topology and geological map of PYGR together.

- Attribute data collection

This relates to the selection, verification and classification of attribute datasets in GIS. The selection entailed a search for details of the map items or attributes that are likely to contribute to mineralisation of the area. This information is kept in a tabular form called the attribute table and can be retrieved. The retrieval process for the attribute and spatial dataset in the predictive map is done through a search of selected attributes indicative of mineralisation, obtained either through empirical knowledge or due to spatial proximity, to manipulate for an output. The search operation allows the spatial attributes to be involved because they are stored as values in the database to be assessed directly through the map presented in a computer usable format.

The overlaying process was conducted to adjust and overlay multiple maps including coordinate points, shape files of mineral locations and topology, the geological map of the PYGR shape files and other spatial map layers in the same area, referred to as classification in GIS. The classification was done to group the set of features into groups such that every mineralised and non-mineralised point was classified as such with its respective class, sometimes binary class is assigned a nominal value, e.g., 0 and 1, yes or no. The attribute data in the database was finally checked to verify that the values conformed with the correct attribute values. The predictive attributes collected may be spatial or non-spatial. Spatial attributes are the mineralise attributes that are measured along space and non-spatial are not completely measured by space. Both attributes are extracted and stored for use in PSM-MPM.

- Integrated spatial analysis of spatial attribute data

The analysis in GIS involved a process whereby existing data mining points collected from the field survey and the cartographic (scanned copy) Geological Map of the PYGR were geo-referenced with WGS-1984 and uploaded into GIS. The mineral point data and the geological map of the PYGR were both projected first as image maps and later digitised to form shape file layer maps that show the location and name of the area represented by points as shown in Figures 3.4 and 3.5. The spatial analysis was conducted using the distance distribution method to investigate the spatial mineral distribution patterns, as well as determine the correlation between mineral deposit points and geological features as vectors. A combination of spatial evidence map layers was created using the technique of add and relate map layer data in GIS to represent the structure of the mineralisation indicators or attributes. A simple exploratory analysis was conducted using GIS tools to visualise mineral data points and other geological attributes, as shown in Figures 3.4 and 3.5. Any observed anomaly or sharp deviation from the regular data distribution pattern observed through visualisation in GIS is considered to be possible outliers, and was therefore either removed or an inference was drawn based on the observed patterns displayed.

A combination of Shuttle Radar Topography Mission–Digital Elevation Model (SRTM-DEM) represented in Figure 3.6 and the 1975 Landsat Thematic Mapper (LANDSAT-TM) data map downloaded from the USGS websites were uploaded into GIS in the form of a raster map at a similar projection to the geological map and the mineral location data points (image map). The downloaded maps were added to the existing geological map of PYGR containing mineralized and non-mineralised data points to form several predictive map layers. All predictive map layers are presented in the form of shape files overlay or stacked to form a single predictive map layer in spatial frame and stored in the attribute table format as represented by Figure 3.7. A binary indicators of 0 and 1 were used to denote absence and presence of mineralisation respectively.

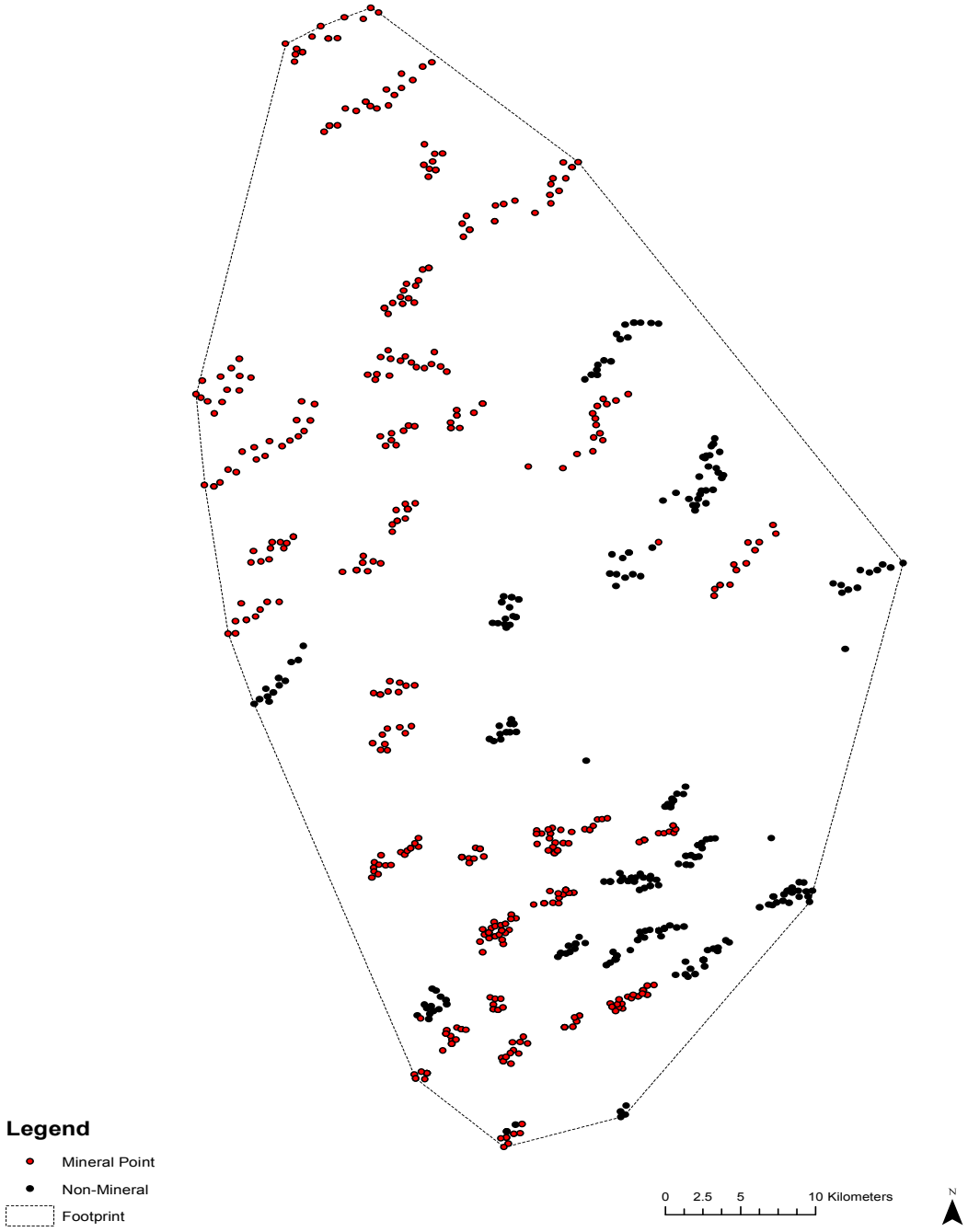


Figure 3.4: A visualised shape file of mineralised and non-mineralised point location on the geological map of PYGR.

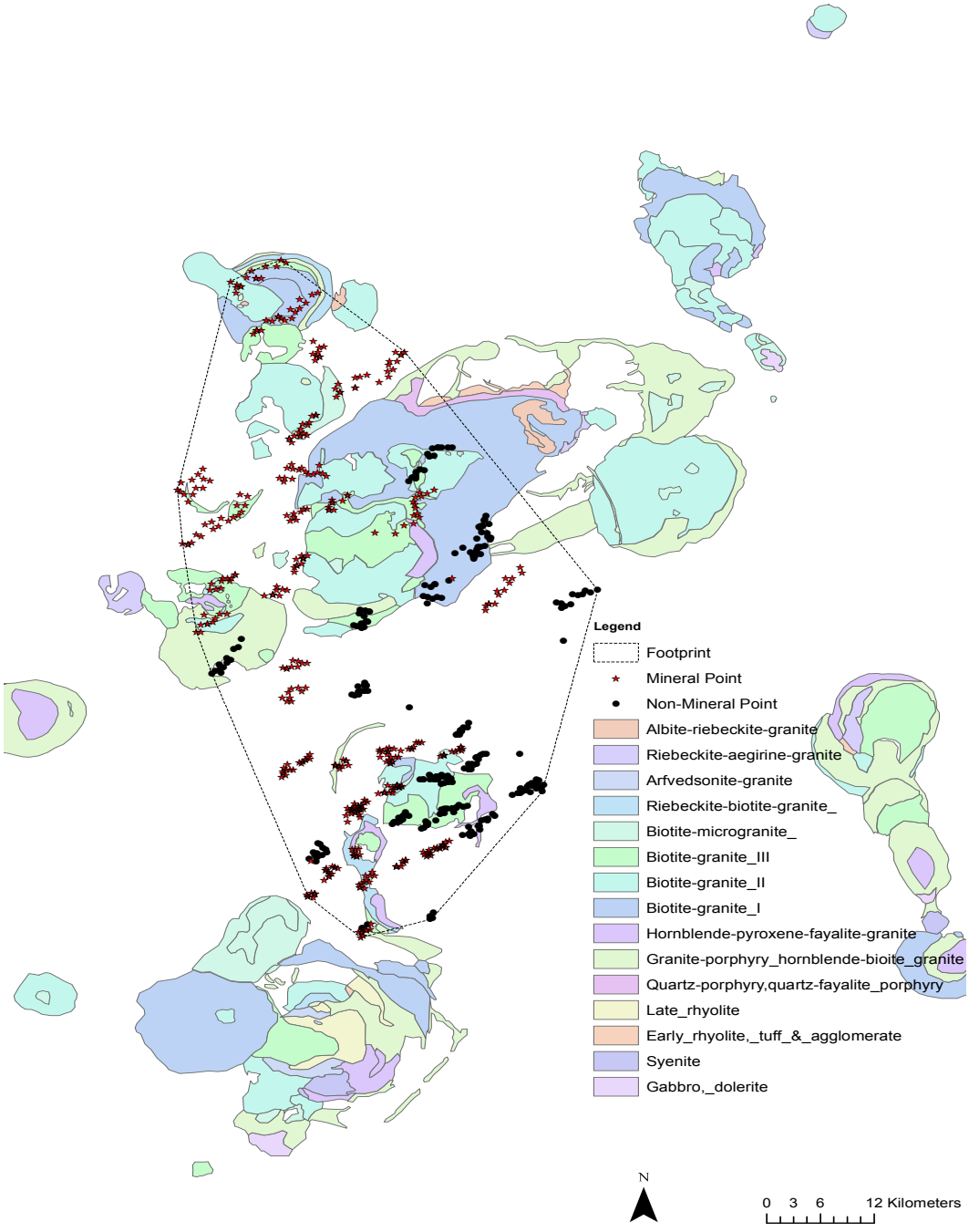


Figure 3.5: Geological shape file map of PYGR showing mineral occurrence points and 204 lithological components within the area of interest.

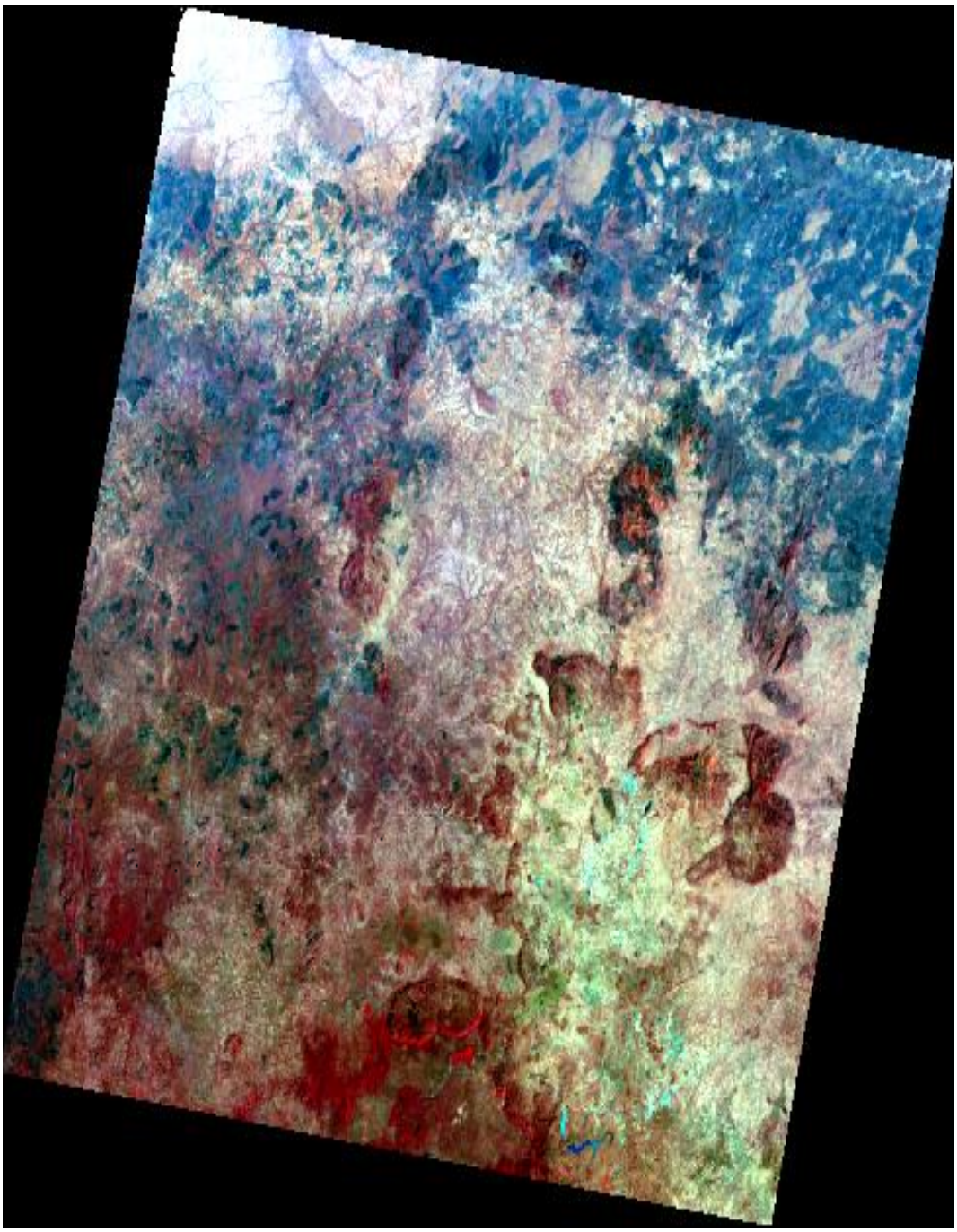


Figure 3.6: A Digital Elevation Model (SRTM-DEM) map obtained from 1965 data map downloaded from the USGS website.

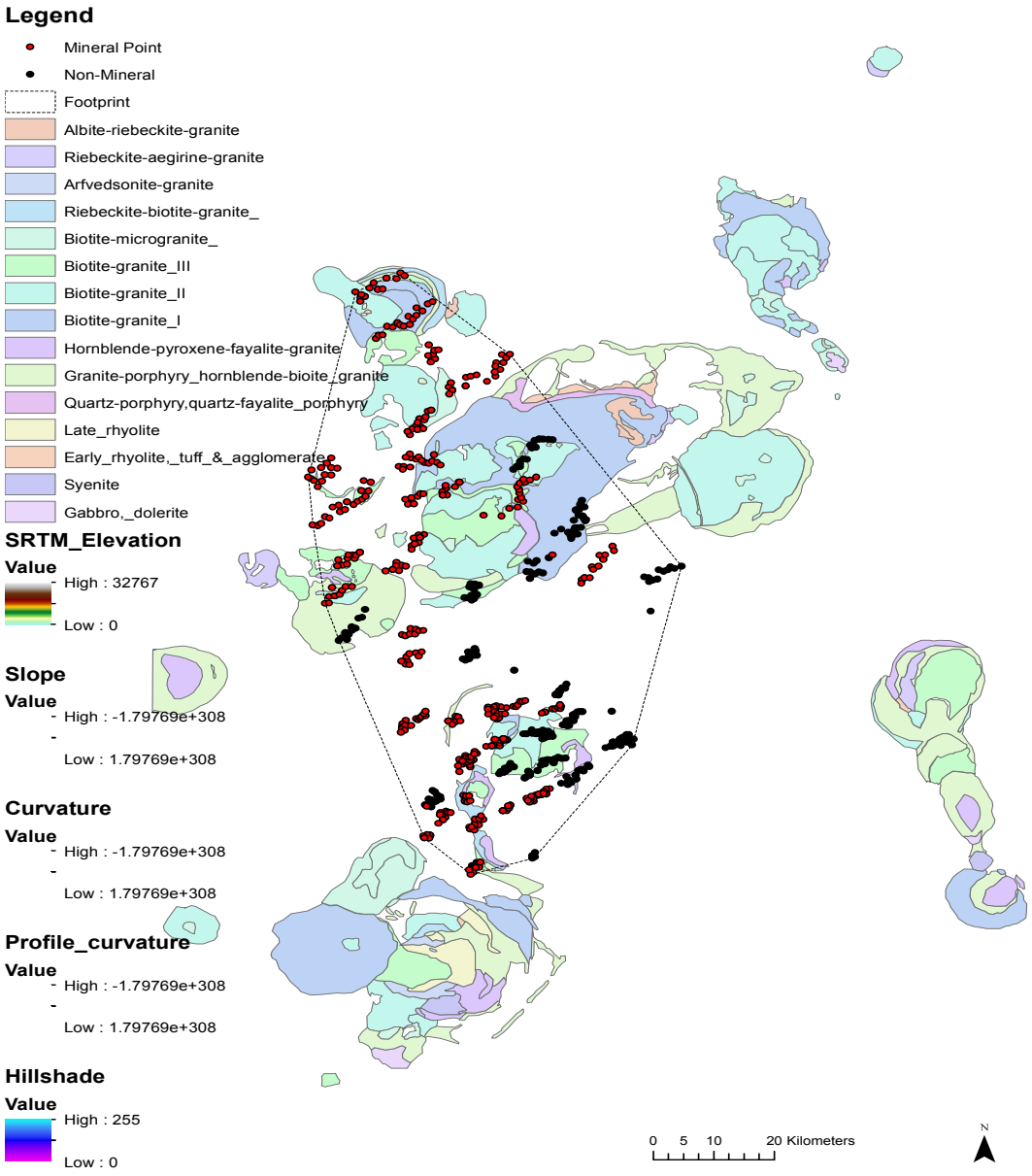


Figure 3.7: A fully digitised predictive geological data map layered according to all the attributes used to build PSM-MPM.

3.5 Statistical Analysis of Mineral Deposit Geo-data

Statistical analysis was performed as a preliminary study ahead of classification; this is a form of exploratory data analysis to determine the distribution pattern of the experimental dataset. A thorough statistical and spatial analysis was conducted to test the distribution pattern of the mineral data points represented in the predictive map. This proved that the distribution pattern is non-random, providing underlying knowledge about the mineral occurrence. Similarly, the spatial point analysis was carried out to test for correlation among the predictive attributes of the mineralisation obtained from empirical evidence such as the rock (mineral source) and the occurrence mineral point. The statistical analysis was conducted sequentially in the order below to determine both the distribution pattern that justifies the use of the selected classification, as well as to establish various spatial correlations among the selected predictive attributes to be used in the PSM. The statistical procedures include:

- Point Pattern Analysis (PPA) for mineral occurrence data points.
- Spatial analysis of points with geological features.

3.5.1 Point Pattern Analysis for Mineral Occurrence Distribution Data

The analysis of point patterns is an analytic method of determining point distribution patterns in a given location. The *complete spatial randomness* (CSR) test was set as a benchmark for testing null hypotheses (H_o) to determine the distribution patterns of a dataset and, in turn, to determine the presence of spatial autocorrelation (CAS). Point Pattern Analysis (PPA) also helped to determine the mineral data point distribution as either random, unknown, cluster or regular as shown in Figure 2.4. There are typically two kinds of PPA: *Measure of Dispersion* (MD) and *Measure of Arrangement* (MA). The MD studies the location of points in the study area (dispersal of points). Whereas, the MA studies point patterns in respect to each other (i.e., the arrangement of points) (Boots & Getis, 1988). The

MA techniques is less rigorous compared to the MD method, because MA does not require the estimation of any value from the observed data to conduct the analysis (Boots & Getis, 1988). However, the MD also has the advantage over the MA because the former is insensitive to some differences in pattern characteristics, such that identical values may be expected for patterns that are different in some ways, and parametric statistics are usually less powerful than the non-parametric equivalent. Additionally, the statistical theory underlying MA is not as developed as that of the MD, hence many subjectivities are involved in the interpretation of analytical results of analysis involving MD (Boots & Getis, 1988).

The technique for *Measure of dispersion* (MD) analysis was hereby used for the purpose of this study because secondary mineral occurrences are a result of the dispersal of solid ore materials by streams or rivers from the source (rock units) to a point of deposits (points). It is, therefore, advantageous to use this method as it is a better representation of the mineral deposit concept obtained from PYGR, than the MA method (Boots & Getis, 1988; Ibrahim & Bennett, 2014a). The MD technique was implemented using the *Quadrat analysis method*:

Using the sample data collected to create a test that answers the type of point pattern distribution, the proportion of variance to mean is determined. The null hypothesis (H_0) test of CSR by Poisson probability distribution is said to be true if the variance of the number of points per quadrat V equals mean, λ (Diggle, 1983; Fowler *et al.*, 1990; Greig-Smith, 1983). Thus in a sample of randomly dispersed objects, $\{x_1, \ldots, x_n\}$ sample variance $V \simeq$ mean λ (Boots & Getis, 1988). The equation 3.1 below from Boots & Getis (1988) was used to calculate the value of variance and mean to determine if the distribution follows CSR as follows:

i.e.,

$$V = \frac{\sum_{i=1}^{n} (x_i - \lambda)^2 fx}{n}. \tag{3.1}$$

and λ is the mean ($\frac{\sum fx}{n}$).

where,

fx = observed frequency of x;

V = Variance of the number of mineral data points per quadrat

x = number of points per quadrat

n = number of quadrats

The ratio of variance to mean is interpreted as follows:

$$\frac{V}{\lambda} \simeq 1 \textit{ the distribution is Random}$$

$$\frac{V}{\lambda} < 1 \textit{ the distribution is Regular}$$

$$\frac{V}{\lambda} > 1 \textit{ the distribution is Clustered}$$

The comparison of V and λ of the quadrat drawn from the secondary mineral data point obtained from PYGR gives a good CSR hypothesis test for a regular pattern where $V < \lambda$. This situation where V to be in excess of λ, therefore, $V > \lambda$., we have a clustered distribution.

3.5.2 Spatial Analysis of Mineral Data Points with Geological Features

Spatial point analysis (SPA) of the geological features of rocks (polygon) was used to quantify the spatial association between mineral deposits and geological features. The Geographic Information System (GIS) was used to build different predictive attribute map layers in a spatial frame of reference within the mineralisation area that described either the presence or absence of minerals in the given area. The analysis uses the techniques of *Kolmogorov-Smirnov statistics or K-S test*. The method was used to determine spatial correlation using distances from points to polygons among various mineralisation attributes.

While some sophisticated spatial analysis was conducted in GIS to join and relate different predictive map layers, as explained earlier, other spatial analysis were performed outside GIS using the spatial statistics of the K-S test. The spatial statistical method was conducted by formulating a test hypothesis that determines the correlation between mineral points data and geological features to draw conclusions based on empirical statistical results. The spatial analysis was also used to investigate and identify the relationship between mineral occurrence, represented as points, and some geological features represented as polygons.

The 2 dimensional K-S test was conducted using the comparison of the cumulative frequency distribution of distance from a set of geological features relative to the mineral deposit location, represented as $D(PM)$ and a cumulative relative frequency distribution of distances from the same set of geological features to a non-mineral deposit locations represented as $D(AM)$ (Berman, 1977; Bonham-Carter, 1985; Carranza & Hale, 2002). A set of opposing hypotheses are proposed to determine the spatial correlation between mineral occurrence points and the geological (lithological) factor, which is considered as the primary source of the minerals as follows:

- H_0: Mineral locations are spatially independent of the set of geological features (rocks).
- H_1: Mineral locations are spatially dependent on the set of geological features (rocks).

$D = D(PM) - D(AM)$. If $D \equiv 0$, it means there is spatial independence while if the value of D is positive (D $>$ 0) it denotes that the graph of $D(PM)$ plots above the graph of $D(AM)$ and, therefore, that there is a positive spatial association between mineral points locations and the geological features. If D $<$ 0, it means the graph of $D(PM)$ plots below $D(AM)$ and, therefore, indicates a negative spatial association between mineral location and geological features.

In determining the correlation between entities of points and polygons (i.e., mineral occurrence points and rocks unit), distance distribution analysis is very key. Recalling Tobler's first law of geography which states: "that everything is related to everything else, but nearby things are more related than distant things" (Miller, 2004; Tobler, 1979). In other words, Tobler's interdependency between spatial data cannot be ignored (Shekhar *et al.*, 2001) in spatial data analysis, as it helps to determine the presence of SAC in the distributed data attributes.

3.6 Design Architecture for PSM-MPM

Figure 3.8 represents a comprehensive and systematic process of *design architecture for PSM-MPM* that involves the collection of data points, data preprocessing,

data analysis, classification model training, cross-validation or testing models, and evaluation and selection based on predictive performance (generalisation). The model design shows how data analysis and the classification of the model design, through machine learning, was used to model mineral deposit point data. The implementation of ML techniques implementation followed data analysis and data export from the predictive attribute geological map in numeric format from GIS. Exploratory data analysis, and other computational experiments, were conducted outside GIS to explore data behaviour in order to select the right classifier to handle particular classification problems. Figure 3.8 also describes the diagrammatic model design architecture from the point of designing the desired attribute datasets in the right dimension, from the predictive attribute data and classes, all the way to implementation and model performance evaluation (i.e., algorithm selection and implementation). It also represents the standard random hold out method of model validation (generalisation) employed in mineral exploration modelling (Porwal, 2006).

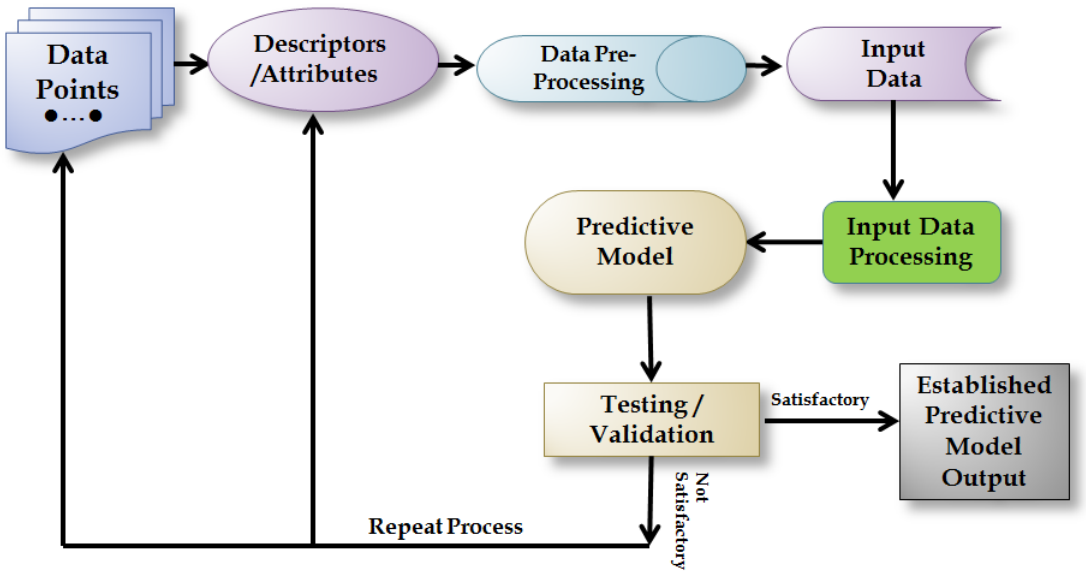


Figure 3.8: Machine learning classification design architecture for PSM-MPM

3.6.1 Selection of Appropriate Classifier for PSM-MPM

Identifying the appropriate classifier is the most difficult aspect of modelling PSM-MPM using ML classification. Since the datasets obtained are spatially distributed in nature, they are therefore likely to correlate in space due to Tobler's law of geography. This correlation often violates attribute independence and causes overfitting in ML classification modelling. The problem of selecting the right algorithm in ML is data dependent, however, since a labelled dataset are often considered to be a supervised ML problem, which is the case with the secondary mineral deposit occurrence obtained from PYGR. Most often, labelled data or instances are identified to be a classification problem —i.e., supervised ML problem. In selecting the appropriate classifier for the PSM-MPM of secondary mineral deposits, seven ML algorithms or classifiers were used at the preliminary stage of this work and was eventually narrowed down to three standard supervised classifiers based on their goodness of fit to the experimental data, and the need to evaluate their performance due to overfitting in the datasets.

The preliminary procedure was a random application of all the available supervised ML classifiers on the dataset in its natural form and then study the performance, instead of starting with the exploratory data analysis. Seven classifiers were used at the preliminary stage of the investigation and modelling, which was subsequently narrowed down to three classifiers as mentioned in Chapter 1 and 2, on the suspicion of overfitting and underfitting. The three selected algorithms used to test model validation techniques that mitigate the effect of overfitting and underfitting in PSM-MPM are KNN, TB and NB.

Since predicting a potential location of mineral deposits is a significant aspect of mineral exploration, the aim of a good ML classifier for PSM-MPM is to build a model that predicts with a degree of certainty, the location of these mineral deposits and generalises well with unseen datasets. The selection of the classifiers was based on their popularity among standard supervised ML, a record of good performance and due to favourable results of basic exploration data analysis (EDA) that examined the mineral data behaviour. Some EDA available for supervised ML classification to select the appropriate classifiers include data

visualisation through scattering plots, re-substitution and principal component analysis, were deployed at different levels in the building of PSM-MPM.

The predictive map layers extracted from the geological map of the PYGR with mineral occurrence points, as shown in Figure 3.1, representing the mineral data sets are a *discrete labelled dataset* categorised as mineralised and non-mineralised representing the class converted to binary values of 0 and 1.

A combination of the joint mineralisation attributes of the PYGR at every mineral location point, jointly referred to as the observation points, consists of the geological, geographic and geo-spatial data contribution to the mineralisation of the area. The mineralisation attributes of the PSM-MPM represented in equation 1.1 of chapter 1 takes a combination of known existing mineralised and non-mineralised points in a spatial frame of reference with the available captured geological and geographic components in the PYGR area, is implemented using the attributes extracted from the maps as shown in the equation:

$$\begin{aligned}\text{PSM-MPM} = \Big\langle & (\text{latitude/longitude}), (\text{Distances to Lithologies}), \\ & (\text{Hillshade}), (\text{Lithological Characteristics}), \\ & (\text{Weight of Nearest Distance to Lithologies}), \\ & (\text{Slope}), (\text{Elevation}), \ldots\ldots \Big\rangle\end{aligned}$$

The concept of PSM-MPM considered each mineral point as an observed data point rather than an area or grid. Each observation point accounts for the position of a mineral occurrence of secondary deposits of cassiterite (representing either presence or absence). Presence here refers to a place where actual mining has taken place and where minerals have been found, and mineral absence means where mining activity has taken place and the minerals sought were not found. It is therefore a representation of the ground truth of the mineral deposit location and the mineralisation attributes of the secondary mineral data of the PYGR.

3.6.2 Predictive Attribute Data and Responses

Data points represent an observation of interest or a location where mineral deposits have been sought and found to be present or absent. Each mining point

was considered to be experimental input data represented in a form of a $2D$ dimensional point. The dimensional data is a representation of latitude/longitude position of mineral occurrence on the earth and is mapped along with the spatial attributes. Points (position) where mining activities have existed, and where minerals, are found are considered to be mineral presence or mineralised points. Places where mining operations have occurred but no mineral ore was found, however, are labelled as mineral absence (non-mineralised points), as represented in Figure 3.3. The data points here represent the total number of data points presented for experimental purposes.

The data attributes and descriptor design is a systematic arrangement of these instances against respective spatial characteristics. The class or response is assigned a binary indicator of 0 or 1 (*yes* or *no*); signifying mineral presence or absence respectively. The data are typically in a form of row and column file format of as in Excel or CSV to create data points, attributes and class. The arrangement is such that every row represents a data point (instance) while the columns represent the attributes, with the last column containing the class or label. The attributes and class values are in the form of either nominal or numeric (real number). The arrangement of data used in a classifier depends on the type of software deployed or the type of data being analysed. For instance, MATLAB and WEKA are excellent examples of traditional ML algorithms and are the desired software used for this research, since they handle both nominal and numeric data efficiently. The entire data attributes and descriptors are pulled out from the attribute table in Arc-GIS software which serves as the database of the predictive map layers, but assigned classes or labelled outside GIS environment.

3.6.3 PSM-MPM Input Data

Predictive modelling uses the application of ML classification techniques to model data in order to make a prediction of potential areas of mineral deposits. The mineral potential mapping of an area is considered the predictive classification of the individual spatial units, using the combination of unique conditions or patterns, as being mineralised or non-mineralised. The classification task consists of a binary class, a predictor pattern that characterises the type, and a particular

condition to consider such as a feature vector containing instances of attributes values. The first step is to identify the type of ML problem at hand and to select the right algorithms capable of handling such a problem. In the case of the secondary mineral deposits data in the PYGR, all the data points are labelled as either mineralised or non-mineralised.

Recalled from the literature review in chapter two, part of the advantage and achievement of this research is, the direct implementation of the *point-based* approach to the modelling PSM-MPM using ML classification (Ibrahim & Bennett, 2014a). The observation data (mineral occurrence position) was considered as points on the map, as against the common use of area or grid approach to conducting mineral potential prediction (Carranza *et al.*, 2009). The *area or grid-based* approach involved considering the mineral location area as polygon or grid lines drawn on a map to represent the site of the mineral position as well as the target of prediction. The polygonal or grid representation of the mineral occurrence location procedure involves generalising the mineral site as an area rather than a point, unlike the point-based approach that allows for the direct implementation of the ground truth using coordinate location (longitude and latitude) without much generalisations assumptions; such as converting an entire points to a polygon. The grid or the area-based approach will not give the precise location of the prediction but just an overview of the mineral position.

About 463 mineralized and 286 non-mineralized data points were split in the ratio of 60% for training and 40% for testing using RHO selection splitting method in MATLAB. The 60:40 ratio split was arrived at after it was found to be the most consistent results after several other ratios failed to perform better. A total of 21 features formed the predictive attributes used for this experiment. All the attributes used for this experiment were spatial attributes except two that were non-spatial (i.e., the rock type and the probability weight of the closest rock to the point). Attributes such as latitude and longitude were not included as input, in the ML classification modelling because, they represent an explicit x, y coordinate points that could be spurious predictors which may mislead the predictor, as they are non-transferable and using them will mean learning exclusively the clusters and not the real attribute. Simple models with fewer attributes sometimes performs well with reduced computational run-time to make for faster execution of the task.

All the selected seven standard classifiers presented in this work were used to test their ability on secondary mineral distribution data represented as points.

3.6.4 Validation and Testing of PSM-MPM

Various methods exist for validating models such as cross-validation, random hold-out also referred to as the stratified RHO selection and, sometimes re-substitution methods (Bradley, 1997; Han *et al.*, 2011). Except in the case of re-substitution, validation of model in ML classification involves an act of splitting predictive data into training and test sets and selecting the model that generalises well and gives the best predictive accuracy or lowest predictive error rate. Validation also helps to solve the problem of overfitting that is commonly associated with spatial data (Ibrahim & Bennett, 2014b), which may result in an exaggerated predictive accuracy score; the result often presents an overly pessimistic predictive accuracy score results. The re-substitution measures the goodness of fit that determines how well the classifier fits the datasets. Where the classifier fits perfectly well, may presents a typical case of overfitting but where the fitness is poor is attributed to underfitting; which imply that the classifier did is not fit the datasets well. However, this problem are often surmountable through data re-sampling, or holdout validation technique.

Once a model passed the validation stage successful, it is retained as an acceptable predictive model for the target dataset, which is the mineral deposit distribution of the of the given area. Otherwise, the modelling process will be re-evaluated and subject to other forms of performance assessment such as investigating the attribute data type used, considering a change of classifier, or the validation technique used to optimise model performance. Validation was done internally, where the datasets used for training the parameters were split into the 60:40 ratio, with the smaller part being used for testing or validation and the larger part used for training (Kohavi, 1995).

3.6.5 PSM-MPM Performance and Selection

The predictive performance evaluations of the PSM-MPM was conducted by looking at the predictive accuracy scores, since the performance of the classifier is

mostly data dependent (Mwitondi *et al.*, 2013). Some predictive models or classifiers that perform very well with certain data may do poorly when tested on different sets of data (Mwitondi *et al.*, 2013). A good PSM-MPM classifier performance on a dataset is determined by how well it fits the test data, or how well it predicts the unknown from the known. A higher predictive accuracy score on test set signifies a good model generalisation to new or unseen datasets, even though the accuracy score may still require further evaluation such as the value of the ROC score, test for overfitting or underfitting etc.

Various methods exist for selecting the "optimal" predictive model. These involve the evaluation of success recorded by a classifier on specific datasets that signifies the generalisation ability of the classifier measured through a high predictive accuracy rate (Mwitondi *et al.*, 2013). The conventional method is to, first of all, view the confusion matrices and then determine the models with the highest predictive accuracy to be selected over one with low predictive accuracy scores (Mwitondi & Said, 2013) when compared. In most cases, choosing a model is often a trade-off between predictive accuracy in terms of finding mineral deposits with the least error or false positives. Lower predictive error scores give higher accuracy, which is necessary because mineral exploration is very expensive and hence requires a high level of assurance in the predictive results before embarking on it. Care must also be taken, however, when implementing a model based on high predictive accuracy alone, but however, a good model in this context to be selected is one with the high predictive accuracy devoid of overfitting or underfitting.

The established method of identifying overfitting in classification is to compare the performance of the classifier based on the test data. If the performance on the test data is significantly worse off, then it suggest overfitting has occurred. The conventional method of identifying overfitting in ML is by splitting dataset into training and test set to perform either RHO or "leave one out" method of sampling. These methods are usually applied at the point of validating the test dataset or when predicting on an unseen dataset. While this approach has worked differently for distinct learning algorithms depending on the type of data used, there has not been any standard method that has truly handled spatial predictive models and takes account of the similarity of the dataset values and their closeness in space.

The similarities of data attributes in an area due to their proximity have rendered them too dependent and hence may result in the classifier either not learning the real independent attributes efficiently, or learning from wrong correlation even when using the conventional data splitting into training and test set highlighted. The similarities in values of spatial data due to their proximity to each other is referred to as spatial autocorrelation (SAC). The exaggerated predictive accuracy score in PSM is often caused by SAC making models to overfit.

Although correlation among attributes is necessary to make a prediction, it is critical to ensure that training and test data are entirely independent in order to establish true correlation. Predictive models are robust enough only when they perform well on a completely new and unseen dataset . Spatial datasets such as the secondary casiterite mineral distribution are often correlated with each other in space (Lary, 2010) and validating such predictive model relating to such phenomena will require a new datasets that is truly spatially independent from the training dataset. Otherwise, there will be little or no significant difference between the training and test data, even if the two datasets are obtained independently (Bahn & McGill, 2013). The PSM-MPM evaluation include testing the efficacy of model performance parameters such as: accuracy, error rate, sensitivity, specificity and the area occupied by the ROC (AU-ROC), for selecting the right classifier for PSM-MPM.

3.7 Effect of Spatial Distribution and Spatial Attributes on PSM-MPM

It is very imperative to assess the importance or the effect of spatial attributes to PSM-MPM performance. The space attribute impact evaluation is meant to answer the question of whether spatial attribute data are needed to build a in PSM-MPM, despite been the reason for overfitting or underfitting in ML classifiers due to SAC in spatial distribution datasets. The assessment of this effect includes testing the efficacy of model performance by deliberately excluding spatial attributes and comparing the performance of the model with the one containing spatial attributes. The predictive attributes dataset are categorised as spatial,

non-spatial and exclusively spatial attributes. The exclusively spatial attributes are the longitude and latitude coordinate points, which have been eliminated in all the learning and predicting process, as it is unchangeable, and non-transferable, i.e., it tends to be bias due to the constant clustering formation at each coordinate point location of the mineral, and this equally affects generalisation of the model. The test for the effect of spatial attributes in PSM-MPM will require a simple test of building a model by first using spatial attributes only, then without the spatial attributes and finally, using both and compare their performance. All the three categories of data set were assigned individually and then together to identify the importance of presence or absence of the spatial attributes data to learning and predictive accuracies of the PSM-MPM produced.

The second evaluation is to test the effect of spatial distribution or spatial composition of the dataset to the modelling performance of PSM-MPM; this method involves the simulation of real datasets that removes spatial correlation in the data distribution completely and use it as a new test sets for the model validation. The resultant performances using these evaluations will show the efficacy of PSM-MPM based on an entirely new dataset (i.e., simulated data), but without autocorrelation or SAC. The modelling and simulation of real-world phenomena often require the generation of random numbers. To simulate, the command *rand* and *randn* functions in available MATLAB were used to generate random numbers. The random numbers are drawn from a random distribution within some given ranges of $[0, 1]$ and $[-\infty, \infty]$ respectively. The simulation tool in MATLAB provides a perfect means of executing simulations involving the distribution of random inputs. The MATLAB Statistics Toolbox offers functions that generate a sequence of random data based on many collective uni-variates or different variable distributions. This set of features includes a few functions to generate random data from multivariate distributions since there is no integrated technique so far for creating multivariate distributions for all marginal distributions, or where the distinct variables are from different distributions (MathWorks Documetation, 2015). The Statistical Toolbox has an extended syntax function for *distribution fit* data and can generate parameters from data for the simulation of real data.

The PSM use correlation among attributes in space to make efficient prediction of spatial phenomena within a given area, and SAC is a concept of this relationship in space. To further demonstrate the need for the presence of SAC in PSM-MPM, the mineral attribute data was simulated to deliberately eliminate the spatial component in the dataset to observe the resultant performance of the model by comparing the performance of the model with SAC.

The systematic steps for simulating real secondary mineral deposit distribution dataset to justify the consideration and importance of SAC on the performance of PSM-MPM is highlighted as follows:

- Calculate the mean and standard deviation of each attribute data values represented in the columns of the dataset.
- Randomly generate new predictive attribute data values based on the parameters of each attribute obtained. The generation of the simulated data values should be done individually, and not collectively to achieve spatial independence in the new datasets.
- Test for the presence or absence of SAC in the datasets. Plot a correlation heatmap to test for autocorrelation in the new datasets and compare with the correlation heatmap of the original dataset.
- Consider the entire simulated synthetic datasets as a new validated or test set.
- Apply ML classifiers or algorithms to learn from original data as to determine performance by validating on the simulated synthetic data as the test set.
- Determine the predictive accuracy scores and the AU-ROC values of each classifier used.
- Evaluate the performance of the classifiers by comparing the predictive performance obtained using real and simulated datasets as test and validated set respectively to assess the importance or otherwise of SAC based on performance of the PSM-MPM.

The traditional model validation evaluation involves the splitting of dataset by learning on trained set of data and validate on another called test set. However, the performance of the PSM-MPM is measured based on the result of the test set also referred to as the validated set. Therefore, the test set has to be genuinely new and unseen data set from the training set. Simulating the mineral distribution data produces some sets of synthetic new datasets with similar attributes as the original or real datasets, but having certain limitations, such as the removal of autocorrelation which is basically the SAC inherent in the distributive dataset. The new synthetic dataset generated by the simulation helps to evaluate model performance on another set of unseen datasets. Simulation of synthetic mineral dataset has made PSM-MPM performance evaluation even more easier particularly for mineral distribution data, since it is often hard to obtain mineral deposit data due to strict mining company policies regulating how such data is given out. Sometimes mineral dataset can be very expensive to acquire. Simulating mineral data distribution that eliminates the presence of SAC from the primary PYGR dataset was used to evaluate mineral deposit distribution dataset as well as determines the predictive accuracy of the PSM-MPM when tested on entirely different or imperfect datasets.

3.8 Data Preprocessing for Predictive Attribute Feature Subsets Selection using PCA

Feature selection method was used to obtain a suitable subset of the most important attributes and to construct the classifiers with the selective attribute (Kohavi & John, 1997; Langley *et al.*, 1992). The Principal Component Analysis (PCA) was deployed using R statistical software, so as to reduce the data dimension and select the best attribute set based on predictive attribute accuracy estimates and determine the subsets of attributes to include in the classification (Kohavi & John, 1997). The predictive characteristics selected uses PCA method of feature dimension reduction that serves to select fewer but more efficient attributes data. Feature subset selection technique has been used in ML to improve model performance in ML classification (Kohavi & John, 1997). Using the correlation and

variance of the attribute values, determined by the principal component eigenvalues, the analysis algorithm will choose attributes with the fewest errors during the model fitting in order to improve the performance of PSM-MPM.

Although classifiers are not normally assessed based on only reducing the error rate during training alone, it is possible that considering SAC through data preprocessing on the training and test set, makes the classifier use knowledge of attributes that are most important in the underlying data distribution. The objective of conducting PCA here is to serve as a data pre-processing technique. It is done to remove redundant data attributes that may be distorting the ability of the classifier in selecting the most important attributes thereby, avoiding noise or redundant attributes in the datasets. Since the attributes are highly correlated in space, conducting a data attributes dimensional reduction attempts to select only the best subsets of the attributes and test if this may help to handle the effect of SAC in the distribution dataset.

The dimensional data reduction of the mineralisation attributes was achieved through the creation of the principle components (PCs) that are the primary explanatory variables. Each principal component is independent and quadratic or four-sided; it is determined by the relative contribution of each of the original variables to each of the principle analysis. The procedure includes standardisation of the mineralisation variables values obtained to the same relative scale and prevents some variables from becoming prevailing due to overriding large measurement units. A combination of correlation coefficient using heat map plot, eigenvalues and the contribution of PCs variable factors will be used as a yardstick to select attributes most important subsets using PCA. Since the problem of overfitting and underfitting in spatial attributes dataset is attributed to high correlation among attributes data in space (SAC), PCA is employed to reduce dimension in the natural datasets and select only attributes that are most important and not those that are highly dependence or noise that may influence classification.

3.9 Evaluation of PSM-MPM Predictive Performance

The evaluation process of PSM-MPM involves the assessment of model performance to select an ideal model based on empirical evidence. Most often, predictive model evaluation are for improvement in performance, to predict mineral location potential and non-mineral location accurately with minimal error rate. The accuracy can be achieved through the adjustment of some attribute data also known as preprocessing (Ibrahim & Bennett, 2014a). The experimental data from the PYGR acquired for this study clearly indicated the mineral distribution points spread as, north towards the east, i.e., the distribution is in the form of the clusters of points as visualised using GIS represented in Figure 3.9.

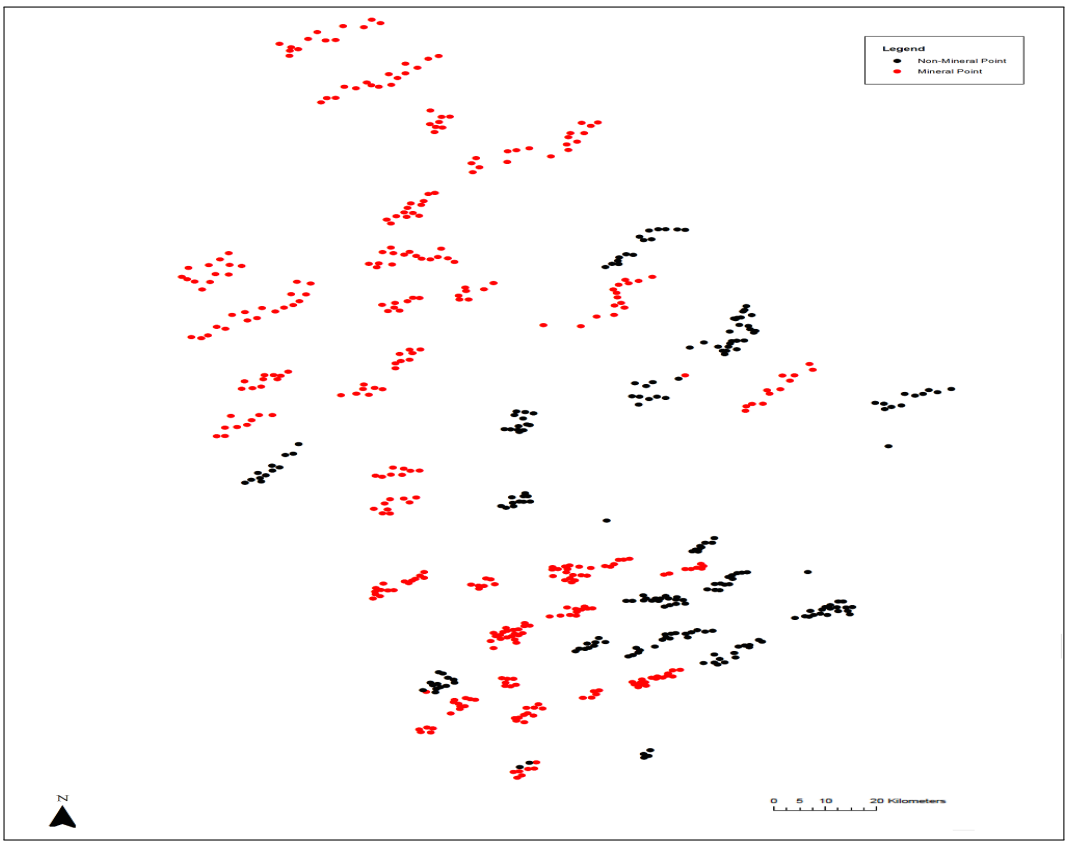


Figure 3.9: Mineral occurrence in the PYGR showing mining points distribution of clusters in 2D space

3.9.1 Four-way Sampling Technique for PSM-MPM Performance Evaluation

Predictive model performance scores result comparison is the conventional method used in most ML classification problems to evaluate model performance and make selection (Bahn & McGill, 2013; Ibrahim & Bennett, 2014b). The method of evaluation deployed here compares the predictive accuracy scores and the AU-ROC scores for the four different sampling methods used to evaluate the performance of PSM-MPM.

A complete four-way sampling approach to model validation of the PSM-MPM performance evaluation technique is proposed to compare the best approach to modelling spatial geodata affected by SAC. A new and unusual spatial strip splitting method of separating spatially autocorrelated training and test datasets is also introduced to select the best classifier by considering predictive attribute independence of the dataset during model validation. By enforcing the spatial disjoint between attributes, spatial strip splitting allows for a space splitting technique which will enable all the spatial characteristics data to reasonably acquire some spatial independence, thereby, reducing the adverse effect of SAC in the dataset (Bahn & McGill, 2013).

The four-way validation of PSM-MPM including testing for *re-substitution*, *random holdout*, *half longitudinal spatial strip* and *quarter longitudinal spatial strip* techniques, was adopted from the work of Bahn & McGill (2013) in the evaluation to test for performance of distributive model performance. The method of the four-way evaluation showed a gradual advancement from the total separation of training and test data (no split or re-substitution) to a severe spatial separation (half strips) and then to a systematic spatially separated split between training and test datasets (strip split) (Bahn & McGill, 2013; Ibrahim & Bennett, 2014a). The approach employed here will not only highlight the way to investigate and identify the detrimental effect of overfitting in the datasets that affects model performance as carried out by Bahn & McGill (2013), but will also test for *underfitting* too. The indication of underfitting is seen in the poor predictive accuracy score by a certain classifier due to SAC in contrast to the high predictive accuracy scores indicating overfitting by other classifiers.

The four way assessment of PSM-MPM performance is considered, to evaluate the effectiveness of the popular methods of data sampling of training and test dataset, by comparing the performances of individual classifiers with respect to overfitting and underfitting, under the following four methods:

- Re-substitution

The re-substitution method simply used the data as a whole without splitting it, as shown in Figure 3.10. Algorithm 1 described the procedure to implement the half split of training and test data sets to be used in an ML classification. In this case, the same data sets used for training a model is also used as test sets. The method is often used to test the goodness of fit of the classifier or algorithm (Fielding & Bell, 1997), in order to select a good classifier for a given dataset. A good classifier should predict perfectly well on test data when trained on the same dataset while an indigent classifier will not (Bahn & McGill, 2013; Ibrahim & Bennett, 2014a). The idea of splitting as validation is to check for generalisation ability of the model first, by investigating the classifier's ability to adapt well with unseen data sets. Generalisation of PSM-MPM is the ability to predict mineral deposit potential in similar but new or different locations that it had not been exposed to, and this is the idea behind validation or cross-validation to identify overfitting or underfitting.

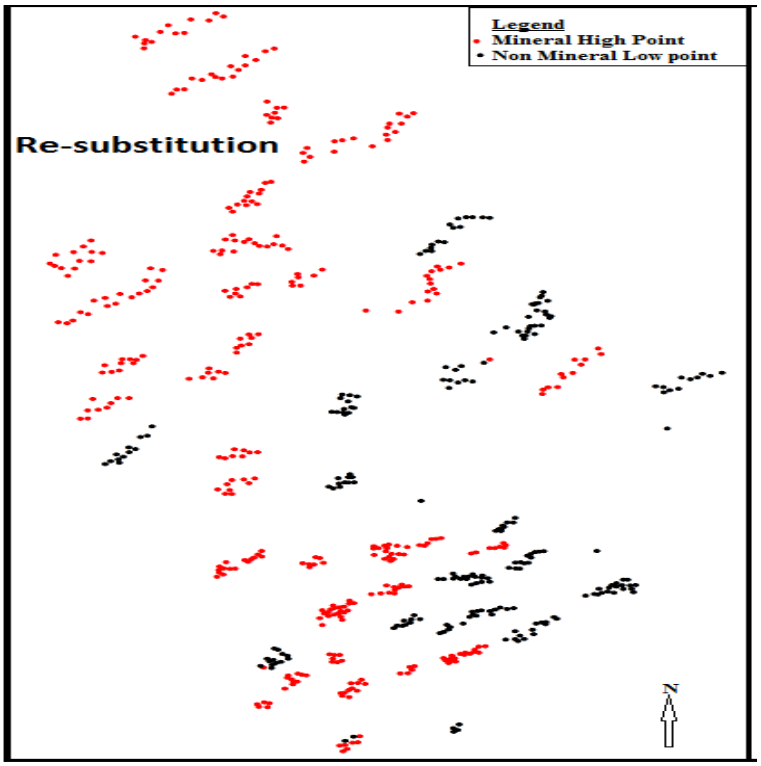


Figure 3.10: Diagrammatic representation of the Re-substitution method of splitting or validation using the same data as the training and test set.

Algorithm 1 A Supervised ML classification using Re-substitution sampling technique for PSM-MPM performance

procedure (Re-substitution Validation Technique of PSM-MPM Performance Evaluation)
    Given S to be mineral occurrence data sets containing N samples; for examples $\{(X_1, Y_1), ...(X_M, Y_M)\}$ ▷ X_M is the feature vector of the M^{th} example and Y_M is the class.
    Consider S as the entire dataset.
    Set S to be the training and also the test set. ▷ S is the entire mineralisation attribute with class.
    Train classifier on S.
    Validate the classifier on test set S and record the correctly and incorrectly classified in confusion matrix
    Report the percentage of correctly classified as the predictive accuracy score for the model.
    Evaluate the predictive accuracy.

- Random holdout split (RHO)

A RHO split is the most widely used method of validating potential mineral deposit and indeed using ML classification. The RHO selection is the method that divides data sets into significant classes randomly until all data are selected (Bahn & McGill, 2013; Ibrahim & Bennett, 2014a). The random split is the

most regularly used method of validation in modelling mineral potential mapping (Porwal, 2006). RHO testing splits a portion of the data into a smaller fraction that is used for validating the model while the larger part of the dataset is used for model building or training. The RHO data selection validation is a pretty straightforward method and is very easy to implement as indicated in Figure 3.11. The split is done in clusters and is thereby unable to separate truly independent datasets for training and testing, as shown in the diagram where the training and test sets are selected by + and **x** symbols respectively. The splitting is done by ensuring that the ratio of training is much higher than testing (i.e ratio of say of 60:40 for training and testing sets respectively), and this is applicable to all the splitting techniques.

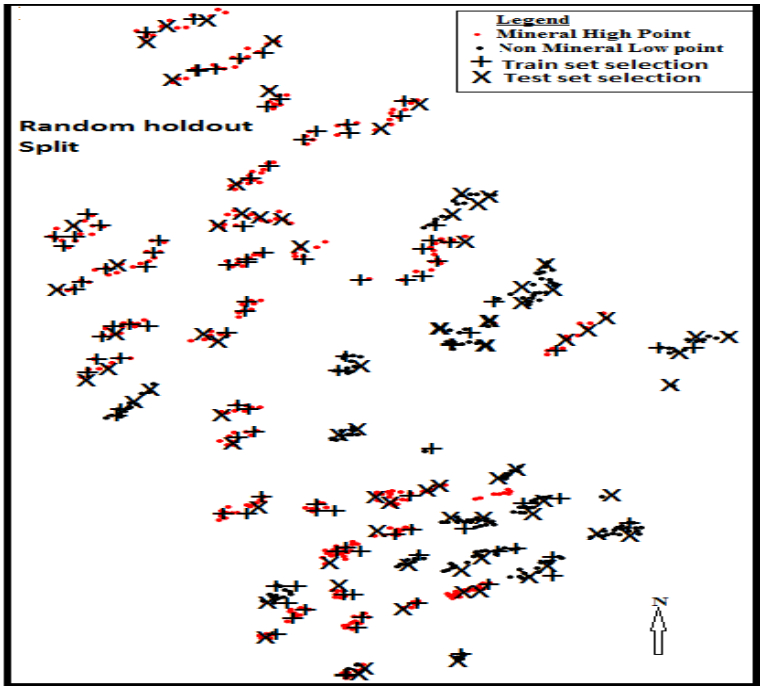


Figure 3.11: Diagrammatic representation of RHO splitting, the + and x symbols represents the splits into training and test sets respectively.

Algorithm 2 (used in this Ph.D. research) describes step by step procedure for executing standard supervised ML classification based on the RHO technique of validation. The procedure details how a PSM-MPM was developed and the performance evaluated, using the traditional RHO technique of sampling to validate

training sets on the test set.

Algorithm 2 Standard supervised ML algorithm using RHO validation

procedure (Random Hold-out Validation Technique Algorithm of PSM-MPM Classification)
Given W to be training samples containing M samples; for examples $\{(X_1, Y_1), ..., (X_M, Y_M)\}$ ▷ X_M is the feature vector of the M^{th} example and Y_M is the class
Partition W into N subsets of $W_i(i = 1 \rightarrow N)$ each having (M/N) samples.
Leave W_1 out and pool the remaining $(N1)$ subsets to generate a new set $\widehat{W_1}$ as training sets ▷ consider training datasets $\widehat{W_1} > W_1$ test set.
Train the classifier on $\widehat{W_1}$
Validate the classifier on test set W_1 and record the number of correctly classified in a confusion matrix table.
Report the percentage of correctly classified for all subsets as the predictive accuracy.

- Half longitudinal spatial strip split

Algorithm 3 (which resulted from this PhD research) represents the third evaluation employed is the space splitting of the dataset into half along a longitudinal line as indicated in Figure 3.12. To split the range in half, the central longitude of all mineral data containing areas splits the dataset along this longitude. Half of the data was used to build the model as a training set and the other to validate it as test sets. The splitting along the longitudinal approach was done to allow each resulting part to contain attributes in locations covering the study area. The methods showed a drastic advancement in the separation of training and test data, causing a severe spatial separation among predictive attributes. There is a complete absence of correlation among the predictive attributes, as shown in the diagram in Figure 3.12. The classifier is only able to learn from one set of data but can not replicate the learning in the other set because the distribution pattern from the left-hand side due to the complete absence of SAC between the training and test datasets as represented in the diagram. For an efficient validation of training on a test set, the data must learn from the real correlation between the dataset on both the right and left-hand side of the datasets. Otherwise, the result of the performance may be affected due to the far drastic separation of attributes' SAC in the dataset.

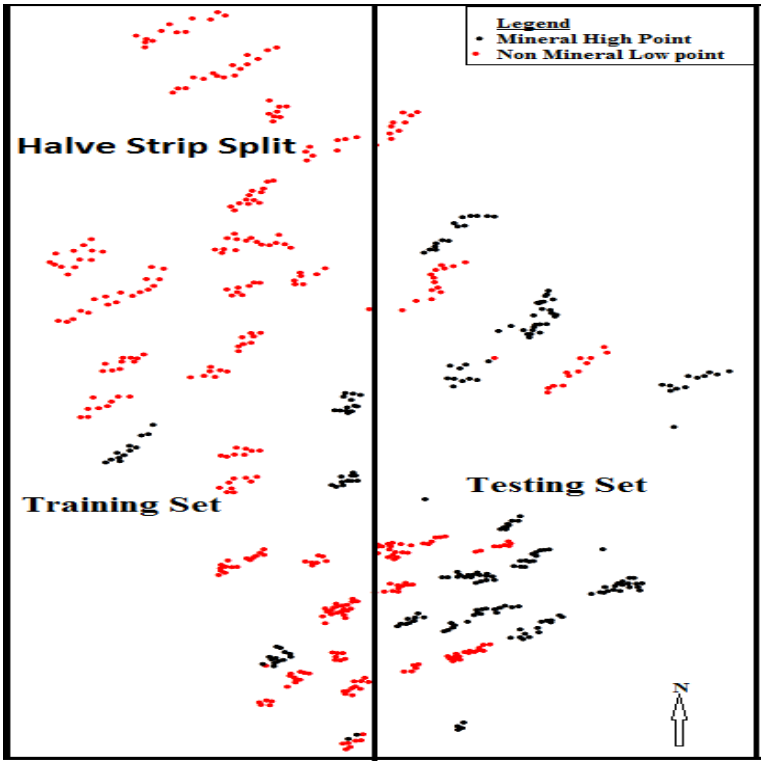


Figure 3.12: Diagrammatic representation of Half Way spatial strips split for real data, the vertical lines represent the splitting of data along the longitude into training on one side and test set on the other respectively.

Algorithm 3 A Supervised ML classification using Half longitudinal spatial split sampling technique for PSM-MPM validation

1: **procedure** (Half Longitudinal Spatial Split Validation Technique of PSM-MPM Performance Evaluation)
2: Given S to be training samples containing N samples; for examples $\{(X_1, Y_1), ...(X_M, Y_M)\}$ ▷ X_M is the feature vector of the M^{th} example and Y_M is the class.
3: Partition S vertically by half along longitudinal spacing into S_1 and S_2 sets.
4: Set S_1 to be the test set and S_2 as training sets. ▷ consider the sample size of $S_2 > S_1$.
5: Train classifier on S_2.
6: Validate the classifier on test set S_1 and record the correctly and incorrectly classified in confusion matrix
7: Report the percentage of correctly classified as the predictive accuracy score for the model.
8: Evaluate the predictive accuracy.

- Quarter longitudinal spatial strip split

The final approach is a dataset spatial strip that divides the dataset into quarters along three longitudinal lines for training and testing, as shown in Figure 3.13. The quarter spatial strips, or the longitudinal spatial strips, worked differently. The technique uses a systematic spatial separation by spatial strip splitting the

training and test datasets into strata spatially along the map longitude (strips). The aim is to reduce the effect of spatial autocorrelation, but not drastically or entirely, due to remaining inter-dependency or SAC in the datasets due to more separation but less in the half. The four-way longitudinal spatial strip split is deployed here for the first time to validate a classifier that models secondary casiterite mineral potential mapping that is spatially distributed.

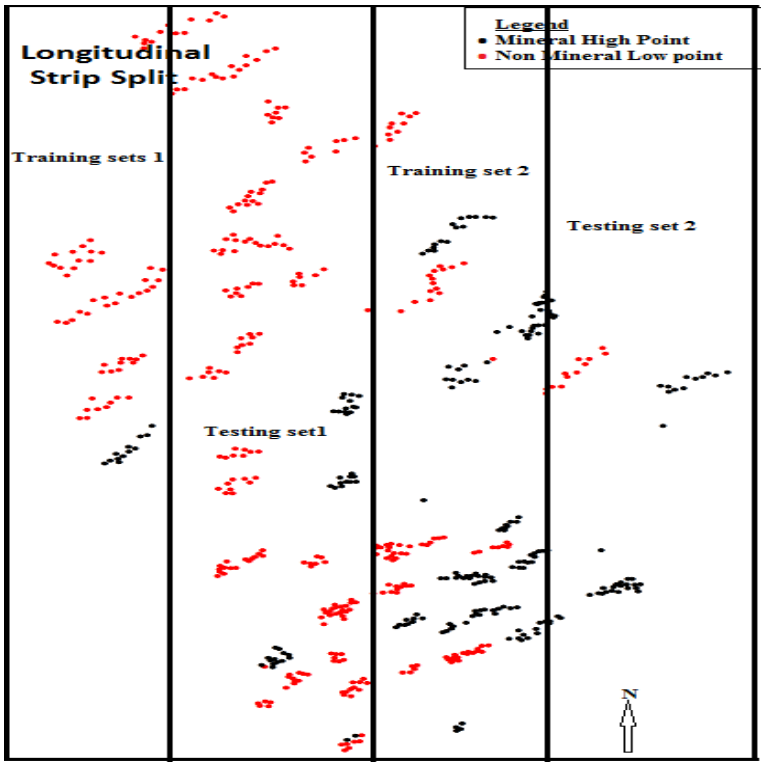


Figure 3.13: Diagrammatic representation of the longitudinal spatial strips method of splitting data, adapted from Bahn & McGill (2013). The vertical lines represent the splitting of data into training and test set using the spatial strips method.

This assessment technique was developed and used to explore the PSM-MPM performance evaluation approach that based on existing work by Bahn & McGill (2013) on the model performance assessment of species distribution. The method was adapted to mineral distribution data since these are both spatial data but of different distribution type. The secondary mineral distribution data obtained for this research are discrete and non randomly distributed whereas the specie distribution used by Bahn & McGill (2013) which is more of a continuous distribution

dataset. The common idea in the two approach is to reduce the effect of high homogeneity (dependence) or SAC among attributes sets (i.e., between testing and training data). SAC is principally a concept of distance related similarity in spatial datasets. The method of dealing with the problem arising from SAC in the data obtained therefore, must be spatially handled. In other words, an SSS divides the data into training and test datasets along space, thereby, allowing for a more independent dataset with higher covariance which may be impossible when using the stratified RHO selection in secondary mineral deposit spatial data. It is difficult to eliminate SAC in spatial data modelling while retaining high predictive accuracy. But it is possible to evaluate the model performance of a predictive spatial model using carefully pre-processed splitting of data sets into training and testing models to assess the effect of SAC in spatial data.

Algorithm 4 A Supervised ML classification using novel longitudinal SSS sampling technique for PSM-MPM validation

1: **procedure** (Longitudinal Spatial Strip Split Validation Technique of PSM-MPM Performance Evaluation)
2: Given S to be training samples containing N samples; for examples $\{(X_1, Y_1), ...(X_M, Y_M)\}$ ▷ X_M is the feature vector of the M^{th} example and Y_M is the class.
3: Partition S quarterly along longitudinal spacing into S_1 and $\widehat{S_2}$ subsets of even and odd combination respectively.
4: Combine subsets of all the quarter split into even and odd partition such that $S_1 = (S_1 \cup S_3)$ and $\widehat{S_2} = (S_2 \cup S_4)$
5: Set S_1 to be the test set and $\widehat{S_2}$ as training sets. ▷ consider the sample size of $\widehat{S_2} > S_1$.
6: Train classifier on $\widehat{S_2}$.
7: Validate the classifier on test set S_1 and record the correctly and incorrectly classified in confusion matrix
8: Report the percentage of correctly classified as the predictive accuracy score for the model.
9: Evaluate the predictive accuracy.

Algorithm 4 (which resulted from this PhD research) describes a step by step procedure taken to execute the supervised ML classification based on SSS technique of validation for PSM-MPM, as a way of mitigating the effect of overfitting and underfitting. The procedure details how a spatial strip splitting method of validation for training sets with the test sets, that eliminates the overdependent of attribute data values caused by SAC or caused by spatial proximity. Since the traditional RHO expressed in algorithm 1 have failed to address the problem of SAC, leading to bad predictive performance (i.e., overfitting and underfitting).

3.10 The Comparative Analysis of PSM-MPM Performance Evaluations

Optimisation techniques of predictive model performance vary depending on the task the model intends to accomplish. In classification modelling, the trade-off may be between predictive accuracy, task execution speed, the simplicity of the algorithm and generalisation. The purpose of optimising PSM-MPM here is to evaluate the performance of the classification algorithms ability critically, to predictive accuracy effect on the overfitting and underfitting. The objective of the work is to examine the predictive performance of the different classifiers to determine the effect of SAC in the dataset that leads to overfitting and underfitting through experiment. The procedure was to conduct a comparative analysis of the predictive performances of TB, NB and KNN classifiers, using the three data preprocessing model performance validation evaluation already highlighted. The techniques which include RHO, PCA-RHO and SSS, were selected for analysis because the other two methods of assessment which include re-substitution and half split do not represent a proper method of data splitting of training and test data. The re-substitution was used to verify the goodness of fit of the classifier, while half split is meant to ascertain the effectiveness of the dataset to ML classification learning.

The comparative analysis offers an empirical approach to evaluating the model performance using different methods and selecting the best method that best describe or detect the presence of overfitting and underfitting through the predictive performance accuracy of the classifiers adopted. The analysis compares the predictive performance of the standard ML classification that uses the RHO approach of validation first without preprocessing and then, with preprocessing using PCA that selects feature best subset (i.e., PCA-RHO), compared with the technique of spatial strip split (SSS) method of validation evaluation. The three methods aimed to shape the PSM-MPM by determining which classifier performed best or improved the best, based on predictive accuracy score that generalises well on the designed unseen dataset; and capable of showing the effect of both overfitting and underfitting due to SAC inherent in the given datasets.

3.11 Summary

The Figure 3.1 represents a logical and sequential standard procedure for that analysis and prediction of secondary mineral distribution data using statistics, GIS and ML classification algorithms respectively. The plan gives a comprehensive explanation of the research methodology adopted in this work. First, it shows how the study area and secondary data from the PYGR were collected and how predictive attributes were extracted using Geographic Information Systems (GIS) and analysed for predictive model building or PSM-MPM. The statistics will identify point data patterns and determine spatial autocorrelation (SAC) in secondary mineral occurrence data obtained from the PYGR. The chapter gives further explanation as to how geographic and geological data are processed, and numeric attribute values are extracted, using GIS or Arc-GIS software to generate a well labelled datasets into standard supervised ML algorithms or classifiers. The overall concept of ML design was systematically explained with the approaches to conventional methods of the model building being set out, from the point of data collection, data assembly, model building, model validation or evaluation and finally model selection. Other procedures for the selection of an optimal model based on predictive performance were also explained. These methods formed the foundation of the process of developing PSM-MPM using a standard ML approach.

The chapter also discussed how MATLAB, R and WEKA software are deployed to implement the design architecture of the building, validation or evaluation and the selection of an appropriate algorithm for PSM-MPM. The aim was to show the predictive performance evaluations to developing the best approach to PSM-MPM sampling for the validation that addresses the problem of overfitting and underfitting due to the presence of SAC in the occurrence mineral dataset. The evaluation was conducted data-wise using three different approaches. Firstly by building and validating the PSM-MPM using the traditional RHO splitting technique without any data preprocessing as is the practice in most ML classification. The second is similar to the first but involving attribute data preprocessing using PCA technique of attribute best subset selection before implementing the RHO technique (i.e.,RHO-PCA). The third approach was considered to be the novel approach which shows how the SSS method was selected after all the conventional

methods proved less effective in mitigating the detrimental effect of overfitting and underfitting.

In the concluding part of the chapter, a method of comparative model performance analysis of the three classifiers that include TB, KNN and NB, was highlighted to determine the ideal performance. The assessment is base on the results of their performances between the proposed novel SSS, PCA-RHO and the traditional standard RHO of data attribute selection and validation method preprocessing. The result of which determined the best approach. This methodology highlighted in this chapter was carefully implemented in chapter 4.

Chapter 4

Analysis and Implementation of PSM-MPM

4.1 Overview of the Chapter

This chapter presents an analysis and implementation of Predictive Spatial Modelling for Mineral Potential Mapping (PSM-MPM) using a technique of ML classification. Parts of the work presented in this chapter have already been published. The three papers published are: *The assessment of machine learning model performance for predicting alluvial deposits distribution* (Ibrahim & Bennett, 2014a); *Point-Based Model for Predicting Mineral Deposit Using GIS and Machine Learning*) (Ibrahim & Bennett, 2014b); *The Optimisation of Bayesian Classifier in Predictive Spatial Modelling for Secondary Mineral Deposits* (Ibrahim *et al.*, 2015b).

Section 4.1 highlights the general overview of the chapter. Section 4.2 introduces the chapter. Section 4.3 discusses the process of exploratory data analysis by implementing a pre-modelling statistical and geospatial mineral point data analysis to determine the mineral distribution points pattern and the spatial autocorrelation (SAC) among the predictive attributes of mineralisation. Section 4.4 discusses the actual implementation of the mineral points data approach to PSM-MPM using some standard supervised ML classifications that measure and compare the predictive performance of three standard classifiers and highlight the importance of the spatial attributes in the PSM-MPM that cause overfitting. The verification of the effect of SAC on the predictive accuracy of PSM-MPM was also

conducted by validating the PSM-MPM with a simulated mineral distribution datasets obtained from the PYGR without the SAC, indicates the importance of spatial distribution to model generalisation. Section 4.5 is the selection of best attribute subset in PSM-MPM using Principal Component Analysis (PCA), the method deployed a data preprocessing attempt to reduce redundant and highly correlated attributes data that may tend to influence predictive ability of the classifiers either negatively or positively. The method tried to reduce the number of attributes to only the most important few subset to be included in the modelling. The consequent of this approach saw an increase in the predictive accuracy of PSM-MPM using fewer attribute datasets that form a more simplified model in some of the classifiers while experiencing poor predictive accuracy in others . Section 4.6 details the implementation of the four-way PSM-MPM performance assessment techniques of re-substitution, random hold-out (RHO), half longitudinal split and longitudinal quartered *spatial strip split* (SSS) sampling of splitting training and test sets for PSM-MPM performance validation compared to the traditional random holdout (RHO) splitting methods of PSM-MPM validation in an ML classification. The SSS technique tries to improves spatial attribute data independence, through a gradual segregation of attribute values in space and thereby eliminating the effect of SAC that causes overfitting by the classifier.

A careful implementation of the SSS technique of generating train and test dataset influences the predictive accuracy of the classifiers to a more realistic accuracy scores. Section 4.7 discusses the results of the comparative analysis of the three major techniques employed in this work, which include RHO, PCA and SSS used to address the key fundamental issue in this research work. The critical problem is determining the effect of overfitting and underfitting caused by SAC in spatially distributed mineral data, that tends to influence the predictive performances of classifier when building PSM-MPM. Section 4.8 highlights all the major contributions achieved by this thesis in line with the set objective.

Finally, section 4.8 summarises the work carried out in this chapter which include: data analysis, pre-modelling exploratory spatial data analysis, PSM-MPM implementation and evaluations, results discussion and knowledge contribution of the thesis.

4.2 Introduction

The implementation of experimental data analysis and modelling approach to the mineral deposit potential of PYGR area was conducted using statistical, GIS and ML classification. The statistical method uses some established statistical techniques such as the Point Pattern Analysis (PPA), Complete Spatial Randomness (CSR) test and Kolmogorov- Smirnov Test or K-S test to determine distribution patterns and the spatial correlation among the attributes selected for modelling. The GIS was used for data pre-processing, that combines all geodata map attributes into a spatial predictive attribute map layer to describe the dataset in a spatial frame of reference using Arc-GIS software. The ML classification algorithms, meanwhile, were deployed to produce a PSM-MPM using MATLAB, WEKA and R software following a systematic adoption of the design approach to using ML classification as highlighted in Chapter 3.

There are various methods of predicting mineral occurrence through predictive modelling but the implementation of the PSM-MPM is specific to the type of mineral deposits distribution data used. The secondary cassiterite mineral deposit formation stages is multi-facet and therefore requires a systematic and scientific approach to the building of a predictive spatial model for mineral potential mapping (PSM-MPM). The adoption of ML technique of Artificial Intelligence considered as the state of the art in predictive modelling is been deployed for the first time on secondary cassiterite deposits distribution dataset. The secondary cassiterite deposits distribution dataset is a very unique mineral datasets due to its formation process, which involves point of formation, medium of dispersal and the deposit point (target of interest). These three processes has made the modelling and prediction to be spatial in nature since it is required to interconnect the three stages of formation to predict their future occurrence. So far, based on existing knowledge, it is the first time an attempt is made to use a *point based approach to create a PSM-MPM using ML classification on secondary cassiterite deposits*, particularly from the PYGR of Nigeria.

This chapter constitutes the main contribution to the work of this thesis, and attempts to answer the fundamental research questions and objectives of the work. This involve the implementation of the research methodology mentioned

in Chapter 3. The major contribution to this work is in the use of point-based approach to building PSM-MPM using ML classification and conducting various predictive model performance evaluation techniques which include developing a spatial strip splitting of the secondary cassiterite mineral deposit data to mitigates the negative effect of SAC that causes both overfitting and underfitting. Other things done leading to success achievement in this work include the geostatistical and geospatial analysis of points as a way of providing analytical analysis to the approach of data collection, model implementation and performance evaluation.

4.3 Implementation of Statistical and Geo-spatial Data Analysis

The outcome of the geostatistical data analysis results in the transformation of all analogue mineral datasets to digital; beginning with the cartographic geological map of the PYGR and the coordinate points of mineralised and non-mineralised areas as points, combining the spatial mineralisation attributes from different map layers to generate a single predictive map output that represents the predictive attribute dataset of the entire PYGR, to be used outside GIS for ML classification. Figure 4.1 represents the sequential transformation of the map data layers. The procedure a method of how data for the experiments are been extracted, that includes map digitization, map conversion into shape files, joining and relating of all predictive geo-spatial attributes data analysis for the establishment of single predictive attribute data map and the extraction of point data values from the attribute table. The first stage of the analysis is the projection or geoprocessing of the analogue geological map of the PYGR, the coordinate mineral occurrence data points and the combination of all attribute data maps in layers. The second stage is the digitisation of all collected analogue geodata converted into shape files, with all the attribute values being stored in the attribute table in GIS. The third stage involved the process of spatial analysis of mineral occurrence data points for both mineralised and non-mineralised areas, along with other geological attributes associated with polygons or as spatial attributes of distances between points and other geologically attributed polygons. The fourth is the combination of all the

predictive spatial map layers into a single map layer and the extraction of the attribute values from ArcGIS attribute table to export into an ML algorithms acceptable format, in the form of attribute data point values and classes.

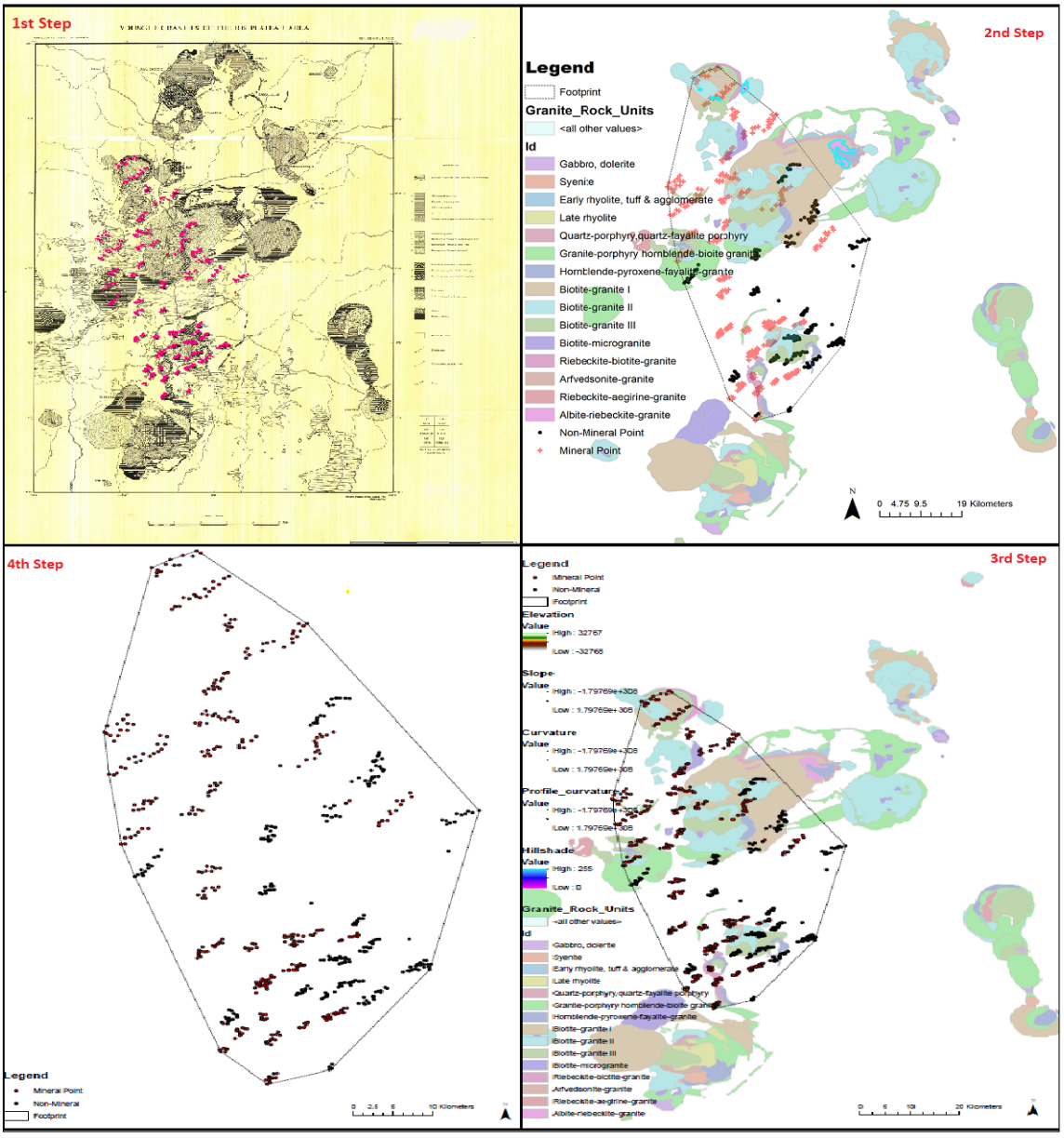


Figure 4.1: The Four (4) stages of map conversion, maps geo-referencing, manipulation and mineralisation attribute point data value representation and extraction.

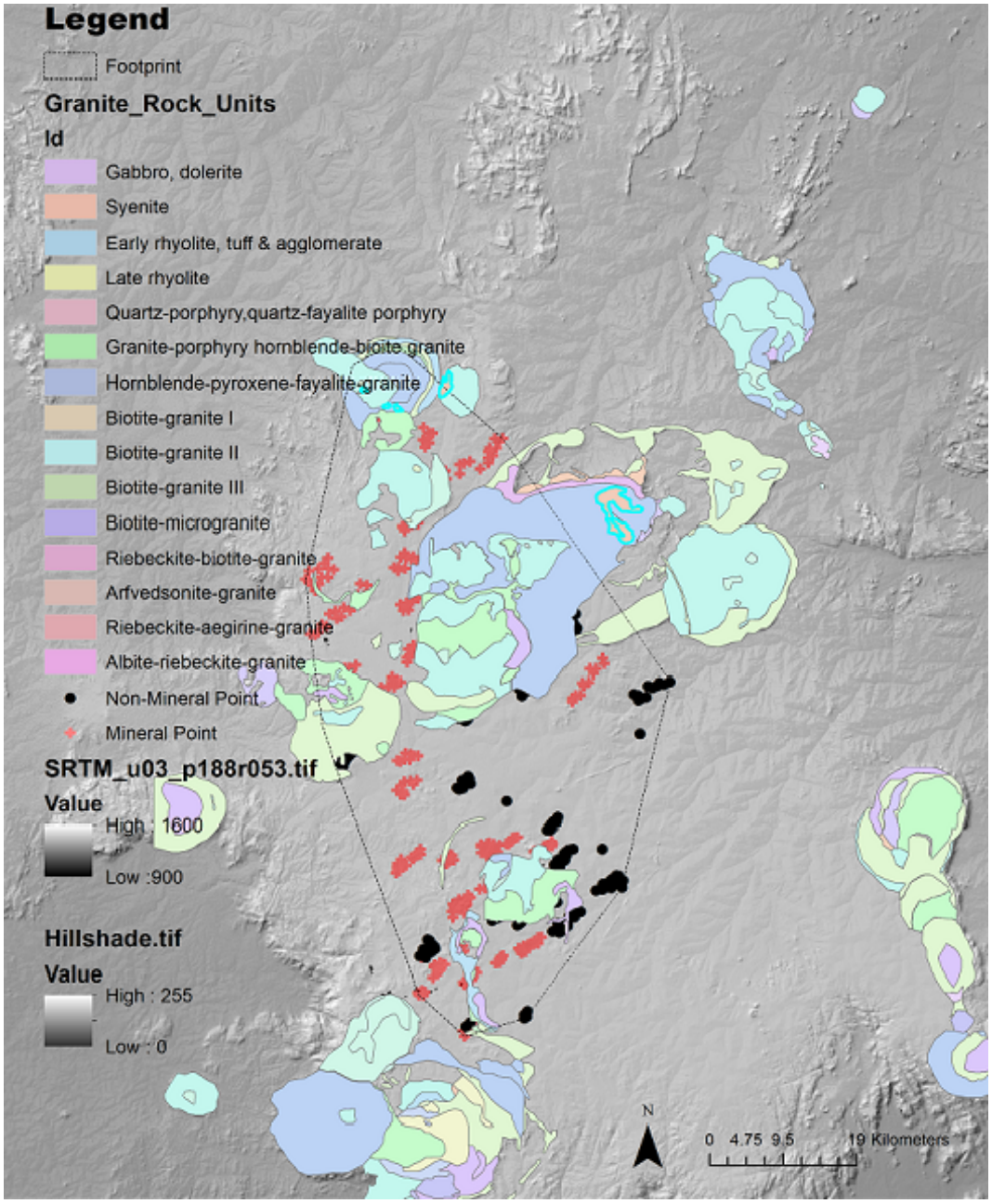


Figure 4.2: Geo-referencing and combination of the raster and vector map layers for all mineralisation attributes data represented as predictive map of PYGR.

This stage generated a total number of 749 labelled mineral occurrence data points, with 463 mineralised and 286 non-mineralised points. The mineralised and non-mineralised points are assigned binary indicators of 1 and 0 respectively.

The analysis conducted included measurement of the nearest neighbourhood (NN) distance from each mineral occurrence point to all geological and lithological points (rock units). The attribute values were extracted from the GIS attribute database table in the appropriate format for use in an ML applicable software such as WEKA, R and MATLAB for computational predictive modelling.

4.3.1 Quadrat Analysis Results

Figure 4.3 illustrates a quadrat formed from the point distribution data map represented in Figure 4.4. A portion of the point data area fitting the squared size of a quadrat was captured to create a quadrat that checks for the distribution pattern. If mineral occurrences are equally likely to occur at every point on a map, the numbers of occurrences in each cell of a uniform grid should follow Poisson distribution. Given a Poisson distribution test as a random distribution test, two opposing hypotheses were tested for Poisson distribution as follows:

- H_o: the distribution is random, (i.e., a Poisson distribution).
- H_1: is that it is not a random distribution (not a Poisson distribution).

The value of the mean (λ) is expected to equal variance (V). From the sample data, the value of V and λ were calculated as 7.7 and 73.3 respectively. The ratio of variance to mean is 9.5 which is greater than one so we reject the null hypothesis (H_o) and conclude that the mineral distribution in the PYGR map had a non-random pattern and based on that, quadrat sampling, it is a more likely to be clustered.

The results indicated that certain geological or geophysical process controls the occurrence of the point mineral distribution in the PYGR area and therefore has a non-random distribution pattern. The failure to accept the null hypothesis validates the point distribution data as an accurate representation of the mineral occurrence points data and can, therefore, is used to build a predictive model for the mineral potential mapping of the type represented. A random distribution pattern cannot be predicted, while a non-random pattern signifies the validity of the dataset analysed here to be the accurate representation of a natural process or phenomenon and not a product of chance.

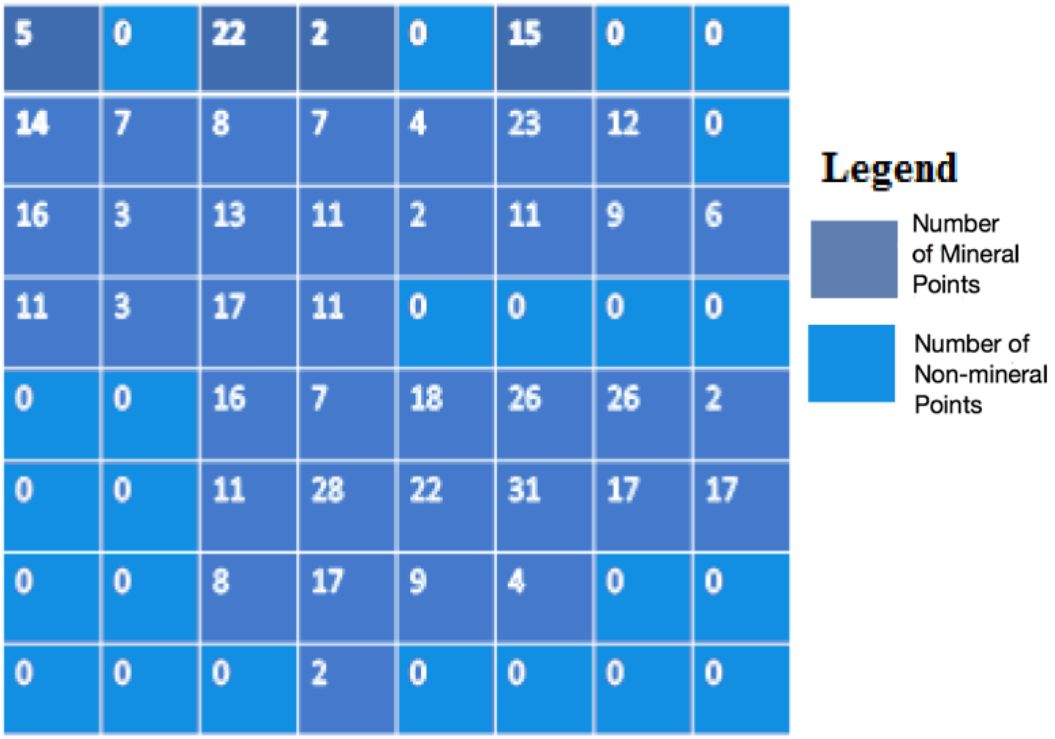


Figure 4.3: A quadrat representation of secondary mineral points distribution of the PYGR

The result obtained shows a high variation (V) in the quadrat, more than the mean λ. Since $V > \lambda$, it is said to have a clustered distribution pattern which also indicates the possible presence of SAC in the data sets. Because the distribution pattern is a cluster, some points appear to be close together and showed correlations among the attribute values represented by the points. Care must be taking, therefore, when modelling this type of spatial distribution data, or when making a prediction, since the predictive attributes of nearby locations may not be independent of the other closer to it during modelling, and may affect its performance. A carefully organised model performance evaluation that involves the preprocessing of attribute dataset will therefore be needed, to ensure a truly independent and correlated dataset are used in order not to violate true attribute data independence due to clustering arrangement in space, which makes model to be bias.

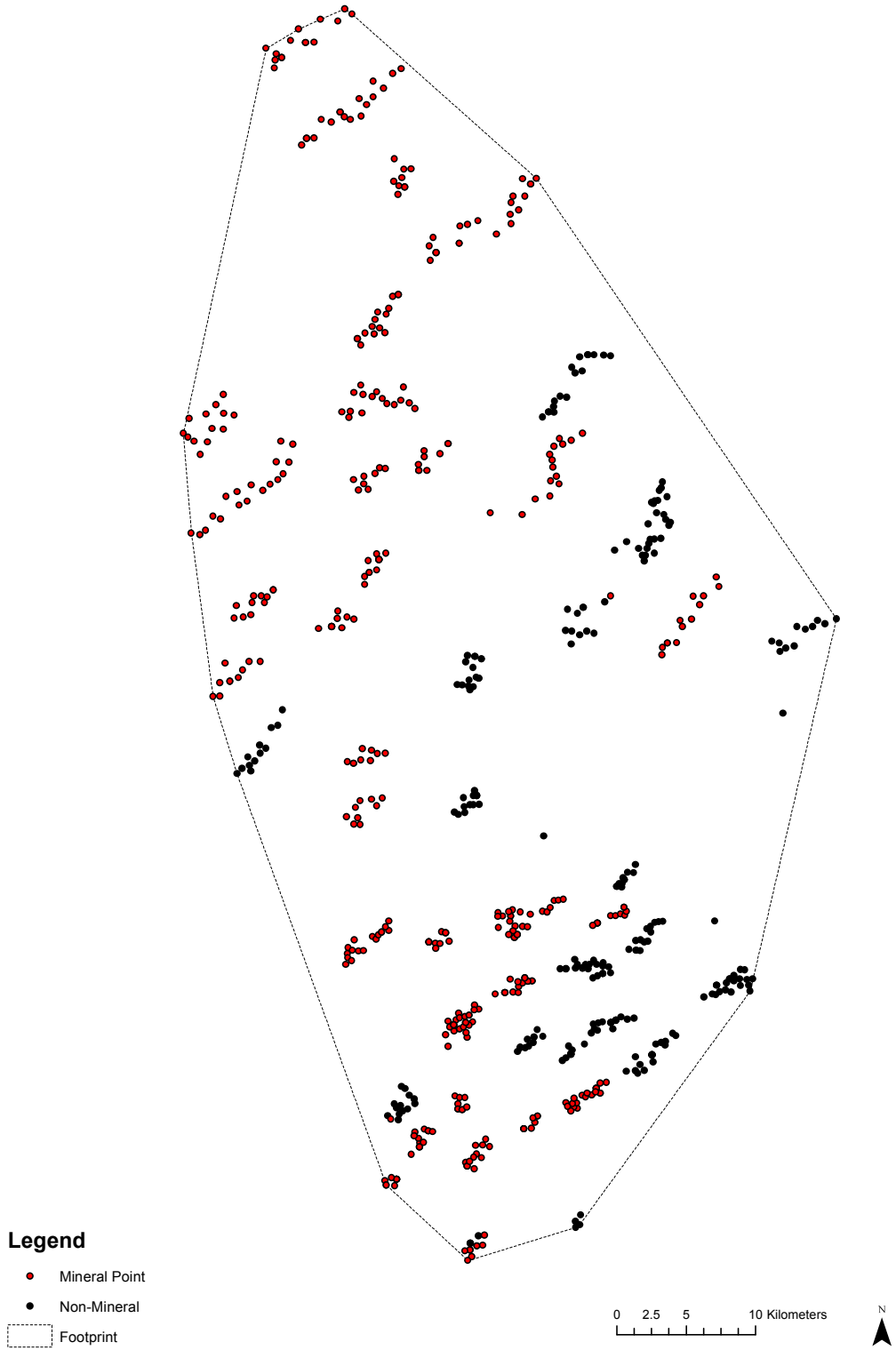


Figure 4.4: Geological mineral occurrence points data map layer of the PYGR in a 2-D space.

4.3.2 The Spatial Analysis of Points with Geological Rock Features

The Kolmogorov-Smirnov or K-S test was conducted using MATLAB and the result is as shown in Table 4.1 where the parameters 'D', 'p', 'h', $D(AM)$ and $D(PM)$ are calculated as give below:

$D = D(PM) - D(AM)$

Table 4.1: K-S test and cumulative distribution functions of the attribute values result.

Variable	Value
h	1
p	$1.4929 \exp -87$
D or ks2stat	0.2284
$D(AM)$	0.3380
$D(PM)$	0.5664

From the result obtained in Table 4.1, the differences between the cumulative nearest distances of mineralised points to a geological feature $D(PM)$ and the cumulative relative nearest distances between non-mineralised points $D(AM)$ to a geological feature (rock) value "D" is given as:

D $= D(PM) - D(AM) = 0.2284$ where,

- D = the non-negative scalar value determined by the maximum differences in cumulative distribution function (CDF) plots, as indicated in Figure 4.5.
- h = the hypothesis test result; logical value of 1|0 where 1 signifies rejection of H_0 and 0 failure to reject H_0.
- p = the probability of observing a test statistics, it is returned as a scalar in the range between 0 and 1.

If the value of D = 1 (D > 0) we reject the null H_0 but if D < 0 we failed to reject H_0. Based on the result obtained, the value of D is greater than 0 (i.e., 0.23) we therefore, reject the null hypothesis H_0 and conclude that mineral locations points are spatially dependent on the set of geological features (rocks), meaning there is a spatial correlation between the source of mineral and the final point of deposits. We also observed that since the graph of $D(PM)$ plots above the graph

of $D(AM)$ in Figure 4.5, this shows that there is a positive spatial correlation or spatial association between mineral deposits (mining points) and geological features (rock units) represented as polygons. Similarly, the second alternative hypothesis also holds that both points samples are not from the same distribution. According to Carranza & Hale (2002), a positive spatial correlation between the mineral deposit points and the geological features is crucial in mineral potential mapping as it helps to validate the importance and selection of the right geological attributes needed to compose a PSM-MPM and of the quantity of mineral found at a given location, although this is not the main objective of this work . The value D or 'kstest2' determined by the two data distribution; is the maximum difference between the two curves as shown by the arrow in Figure 4.5 corresponding to the value of "D'.

Figure 4.5 represents an empirical CDF of distances from mineral occurrence points (presence and absence) to the nearest geological rock features that measure the relationship between the location of mineral occurrence and the rocks within the area. The empirical CDF indicated that at a shorter distance, both the $D(MP)$ and $D(MA)$ converge at a distance of 0.018 representing 68% of the total CDF. The Figure shows that the $D(MP)$ plots over $D(MA)$ as it converges, separated before it finally dispersed. The Figure 4.5 showed consistent with the theory that the secondary mineral deposits tend to be more common at a certain distance from certain rocks (from which they originated). The distances of points to the rocks (geological attributes) are presumably related to the presence of alluvial flow that creates the secondary deposits. In other words, there is indeed a spatial correlation between the mineral occurrence position and the existing rocks closer to it, as indicated by the rejection of the null hypothesis (H_0). The CDF also helps to justify the inclusion of nearest to rocks distances as predictive attributes that also shows the spatial relationship between mineralised points is higher than CDF for non-mineralised points.

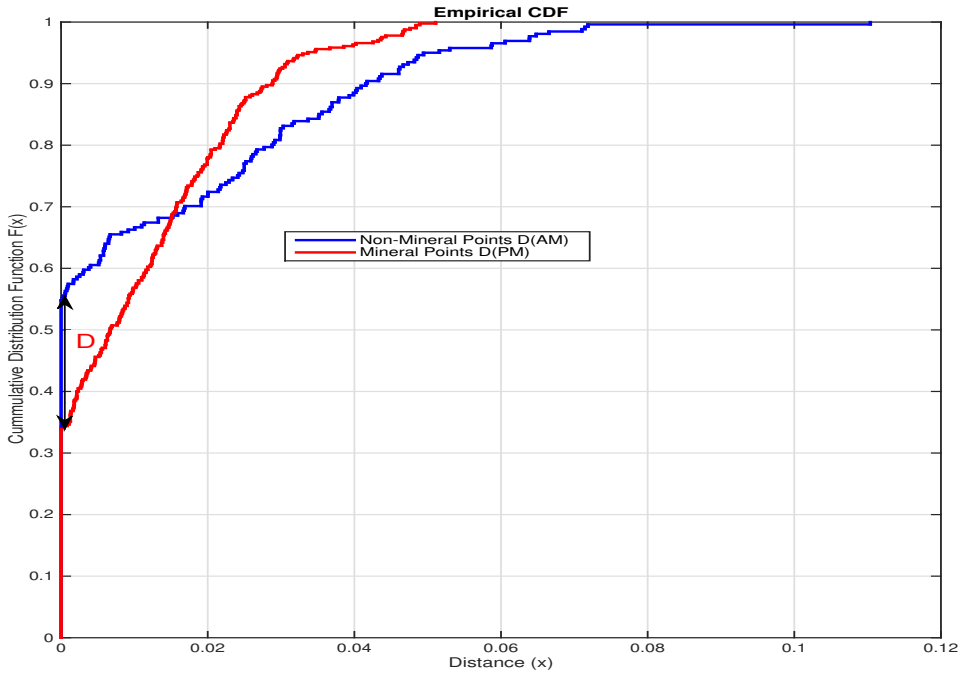


Figure 4.5: K-S test for empirical CDF plots showing the cumulative distribution functions of nearest distance from mineralised points to a geological feature $D(PM)$ and the cumulative distribution functions of relative nearest distances from non-mineralised points to geological features $D(AM)$, with value "D" representing the maximum differences of the two plots.

4.4 Implementation of Supervised ML Classification for PSM-MPM

The implementation of the PSM-MPM using a standard supervised ML or classification approach was conducted using the discrete secondary mineral distribution data obtained from the PYGR. Since the major concern is to implement the PSM-MPM using the available dataset, only the performance through evaluation is the focus and not the details working of the ML classifiers. The traditional random holdout (RHO) splitting method of model validation was performed on the data in order to split the predictive dataset consisting of 749 observed mineral points data, of which 463 are mineralised and labelled 1 or yes, while 286 non-mineralised and labelled 0 or "no" with a total of 22 predictive attributes. The result of the model performance using Naive Bayes (NB), Support Vector Machine (SVM),

K-Nearest Neighbour (KNN), Decision Tree Bagging (TB), Decision Tree (DT), Logistic Regression (LGR) and Discriminant Analysis (DA) is given. The PSM-MPM produced were analysed regarding their individual classification ability by comparing the predictive accuracy scores, the area under the ROC score as well as the misclassification rate of each classifier used, in order to determine the best performing classifier for the PSM-MPM.

4.4.1 Predictive Performance Result for PSM-MPM

The results of the PSM-MPM performance is as displayed in Tables 4.2 & 4.3 and Figures 4.6 & 4.7 below, showing the confusion matrix, predictive performance table, misclassification histogram and ROC scores, respectively. The predictive performance of the PSM-MPM produced by the seven classifiers related to the 40% validated or test datasets using the RHO method for model validation. Figure 4.6 is a diagrammatic representation of the confusion matrix represented in Table 4.2 that indicates the ability of each classifier to predict from the actual test dataset and to show the rate of misclassification by each classifier. The result showed that the KNN and TB predicts the presence and absence of mineral points in the area best from the true test sets while the NB was the least performing classifier due to its high level of misclassification error when predicting possible mineral presence compared to the other classifiers used in the modelling (a predictive error of 33%, i.e., having least predictive accuracy score of 67%). The results of the confusion matrix performance was further justified by the misclassification histogram shown in Figure 4.6. The predictive accuracy performance of the PSM-MPM produced was presented in Table 4.3 and the *Receiver Operating Characteristic* (ROC) plot in Figure 4.7, showing that the KNN and TB has predictive accuracy scores of 98% and 97%, and the AU-ROC plots showing higher plot value for KNN and TB respectively; close to a perfect AU-ROC area of 1. The seven classifiers used for implementing the PSM-MPM were assessed by comparing the performance of the individual classifier and selected the best performing model based on either lower misclassification and high predictive accuracy scores or low predictive error rate.

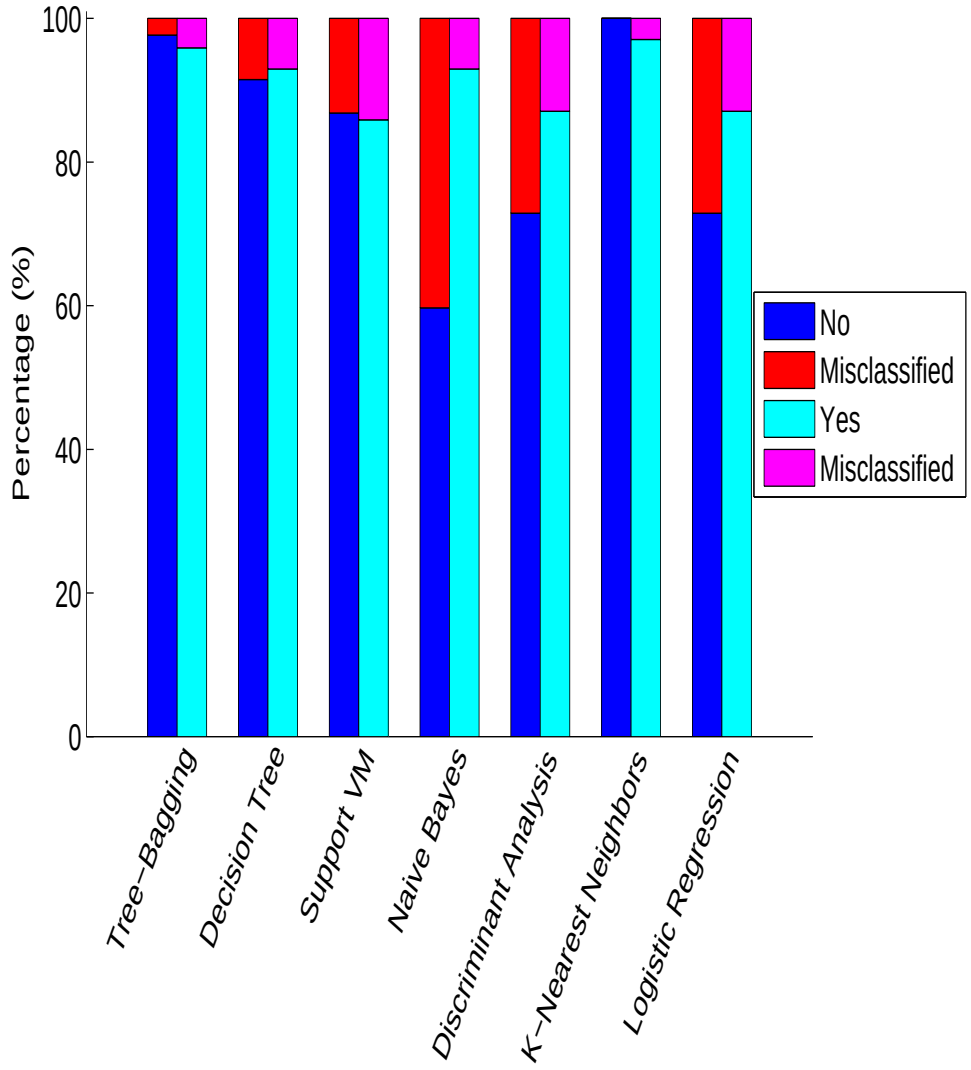


Figure 4.6: Misclassification of PSM-MPM performance for all the classifiers using standard ML classifiers evaluated by RHO.

Table 4.2: Confusion matrices labelled (a–g) for TB, DT, SVM, NB, DA, KNN and LGR algorithms respectively, showing performance evaluation of PSM-MPM using standard RHO data selection.

(a) TB confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	126	3
True MA	7	163

(b) DT confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	118	11
True MA	12	158

(c) SVM confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	112	17
True MA	24	146

(d) NB confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	64	65
True MA	33	137

(e) DA confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	94	35
True MA	22	148

(f) KNN confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	129	0
True MA	5	165

(g) LGR confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	94	35
True MA	22	148

Table 4.3: The model performance scores for all the classifiers in percentage (%)

Classifiers	Accuracy	Error	Sensitivity	Specificity
TB	97	3	98	95
DT	92	8	91	93
SVM	86	14	87	86
NB	67	33	50	81
DA	80	20	73	87
KNN	98	2	100	97
LGR	80	20	73	87

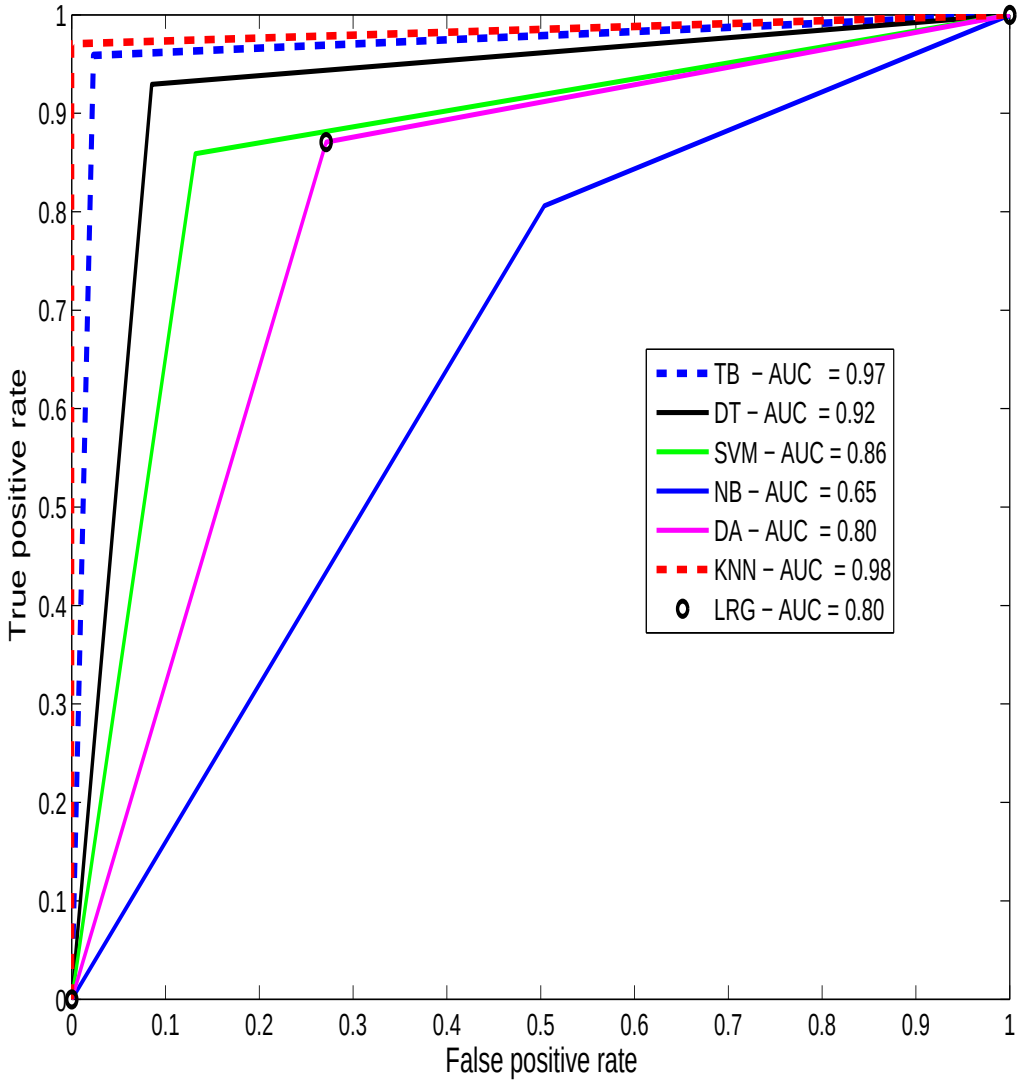


Figure 4.7: The ROC-AUC curve plot showing the performance of seven ML classifiers based on test dataset using standard RHO selection.

Based on the results of the model performance analysed above, the generalisation challenges are addressed through the validation method of RHO, which shows an exaggerated predictive accuracy score to a near perfection by the KNN and TB classifiers, but a low predictive accuracy in the NB classifier. These performances are indicators of model overfitting and underfitting by the classifiers

caused by SAC, inherent among the spatial attribute dataset such as: geographical, geophysical and geological attributes of the PYGR area. A model validation evaluation is, therefore, necessary to ascertain first the significance of the spatial attributes causing SAC in developing a PSM-MPM, and to develop a technique of model validation that addresses the problem of overfitting and underfitting due to SAC in spatial attribute data modelling, especially in PSM-MPM.

The PSM-MPM performance evaluation presented a novel approach that challenges the existing RHO validation technique that splits attributes data randomly. A multiple model validation evaluation approach was conducted to allow for a comparison between the various methods of validation in ML classification in order to select the best model validation approach that addresses the concept of SAC as a primary cause of overfitting or underfitting in spatial data modelling such as the PSM-MPM.

4.4.2 Justification for Spatial Attributes Selection in PSM-MPM

The implication for the inclusion of spatial predictive attributes as components of SAC used in building PSM-MPM is demonstrated by the results expressed in Table 4.4 and Figure 4.8 that shows the predictive performance of PSM, with first the use of spatial attributes only, then, the non-spatial attributes only and finally the combination of both spatial and non-spatial attributes in building PSM-MPM. The spatial attributes are indeed very important attributes in predictive modelling of spatially distributed data such as the secondary mineral distribution data used in this experiment.

The deliberate exclusion of spatial attributes during PSM-MPM building using the seven standard classifiers seek to implement the test method as previously mentioned in Chapter 3, and to demonstrate the importance of spatial components in spatial data modelling. The results showed an insignificant difference in the predictive accuracy scores when non-spatial attributes were removed and when only the spatial attributes were selected for modelling. A sharp drop or differences in all the predictive accuracy scores was noticed when all the spatial

attributes were excluded in the modelling. The entire seven classifiers or algorithms used indicated that despite the exclusion of the explicit spatial attributes of x and y coordinates represented by latitude and longitude, to prevent the classifier from learning only the clusters to avoid overfitting, the classifiers still overfits the datasets due to SAC. The predictive accuracy scores shown in the bar chart Figure 4.8 shows that the deliberate exclusion or otherwise of spatial attributes, highlights the role of spatial attributes in building PSM-MPM.

The test for the importance of spatial attributes inclusion in PSM justified the need to include the spatial attributes in building PSM-MPM. The result of the performance test in Figure 4.8 clearly shows that non-exclusively spatial attributes of distances, such as the various nearest distances of observed points to each of the fifteen rocks, and the elevation culminating to form SAC, have similar predictive influence as the exclusively spatial attributes of longitude and latitude coordinate points on PSM-MPM performance. Thereby, allowing classifiers to learn from clusters formed based on location and leading to overfitting. The result of the PSM-MPM performance using latitude/longitude coordinates leads to an exaggerated predictive accuracy score (perfect fit) or underfitting (poor fit) when learning randomly by the ML algorithm, using the test set. However, PSM-MPM that was constructed using the non-spatial attributes, have equally showed some potentials to contribute to learning, an instance is the accuracy of 71% recorded by KNN as shown in Figure 4.8 and Tables 4.4 respectively.

Table 4.4: The effect of spatial attributes and SAC on PSM-MPM accuracy performance result in percentage (%).

Classifiers	Spatial + Non-spatial	Non-spatial only	Spatial Only
TB	97	67	97
DT	92	67	92
SVM	86	60	90
NB	67	57	67
DA	81	60	83
KNN	98	71	99
LGR	81	60	81

The second evaluation is of the effect of spatial attributes causing SAC on the PSM-MPM predictive performance for mineral potential. The results show the

importance of spatial components that induce SAC in predicting mineral deposit potential. The predictive performance results for all the algorithms used in the evaluation process remain consistent and the predictive accuracy is not affected when non-spatial attributes are isolated from the datasets, but saw a drastic drop in the accuracy scores when only the non-spatial attributes were used alone, regardless of the learning algorithm used, as seen in the results presented in Figure 4.8. The two test or assessments results presented in Figure 4.8 and Table 4.4, shows the importance of spatial attributes indicating high predictive accuracy scores when present and poor predictive performance when absence.

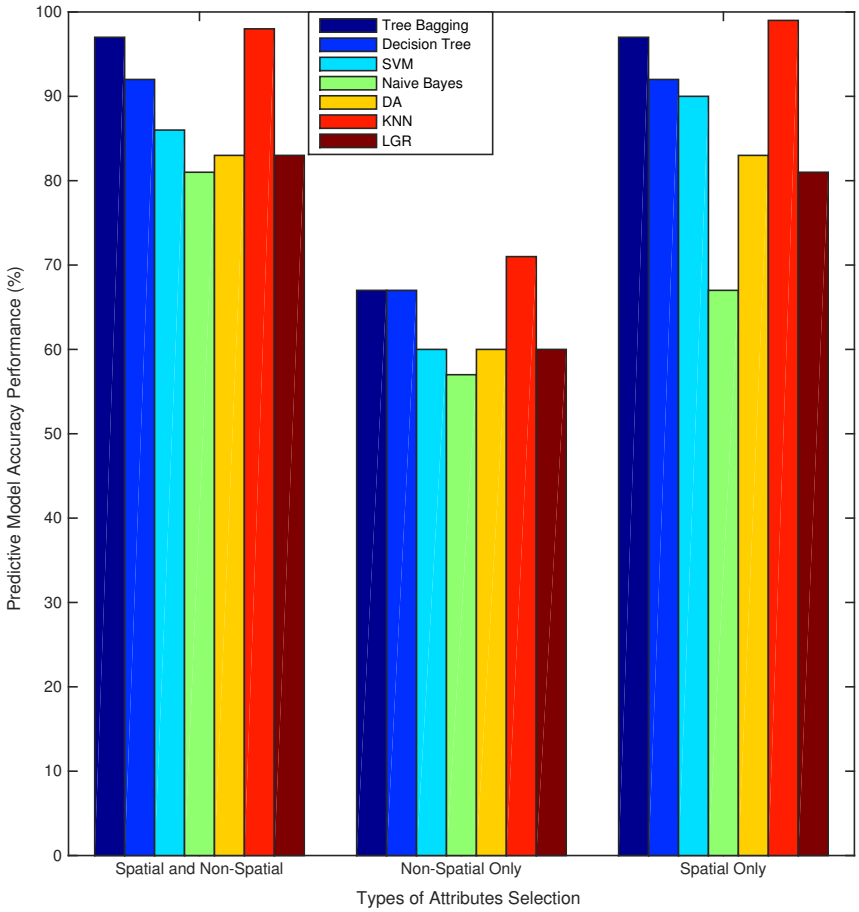


Figure 4.8: The result of PSM-MPM performance evaluation based on presence and absence of spatial attributes.

4.4.3 Verification of the Effect of SAC in PSM-MPM Using Simulated Mineral Distribution Point Data

The creation of synthetic dataset through simulation of mineral distribution datasets obtained from the PYGR, with multiple attributes of X_{ij} was conducted using random number generation syntax command in MATLAB, the parameters of i represents the number of point observation row ($i = 1, 2....., 749$) and j accounts for the attributes value number in column ($j = 1, 2, ..., 21$). The mean and standard deviation as (μ, σ^2) was used to generate random mineral data point values. The mean and standard deviation of the dataset matrix X_{ij} was used to produce the desired amount of data patterns also in MATLAB. The simulated dataset generated using the new parameters of mean and standard deviation, however, eliminates the presence of spatial autocorrelation (SAC) in the datasets and is used for the validation of the PSM-MPM. The datasets were obtained under the following assumptions that: the mineralisation attributes of the original datasets are entirely independent and, as such, each data point was randomly sampled from a random geological survey because, the mean and standard deviation of the *real data* for each attribute captures the distribution of the real dataset.

The result of PSM-MPM validation using the simulated data shows poor performance in all the algorithms that have hitherto performed very well with the real data. The reason is simply that the simulated data completely removed correlation (SAC isolation) in the datasets, as indicated in Figure 4.10. The correlation heat map shows a complete absence of any significant correlation which are mostly spatial, among attributes compared to the correlation between attributes in the real datasets with SAC, as shown in Figure 4.9. Simulated attributes are assumed to be independent as each attribute is generated independently, free from interaction with another attributes within the datasets, thereby isolating SAC in the data to be used for model evaluation. The Predictive accuracy performance was slightly above 50% among the high performing algorithms such as KNN but less than 50% in most of the algorithms used; however the highest predictive accuracy score of over 55% was recorded by the DT as shown in Tables 4.6.

The predictive accuracy of the simulated data was similar to the results of random guess classification, however, meaning that the performance is very poor,

hence predictive attributes in spatial data should include spatial elements that exhibit a reasonable amount of SAC. Simulating data distribution using random parameters that assume attribute independence will generate a replica of the data but without the autocorrelation, as demonstrated in Figures 4.10 and 4.9. The simulated data removes all spatial correlation among simulated attributes for validation of the model. Testing the generalisation of PSM-MPM using simulated data derived from real data will display the extent of the predictive model performance both with and without attribute data correlation or the presence of SAC.

Table 4.5: Confusion matrices labelled (a-g) for TB, DT, SVM, NB, DA, KNN and LGR algorithms respectively, showing the performance of PSM-MPM validated using a simulated secondary mineral distribution dataset without SAC component.

(a) TB confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	329	299
True MA	57	64

(b) DT confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	363	265
True MA	69	52

(c) SVM confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	341	287
True MA	62	59

(d) NB confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	274	354
True MA	47	74

(e) DA confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	293	335
True MA	55	66

(f) KNN confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	288	340
True MA	54	67

(g) LGR confusion matrix based on test set

	Predicted MP	**Predicted MA**
True MP	309	319
True MA	55	66

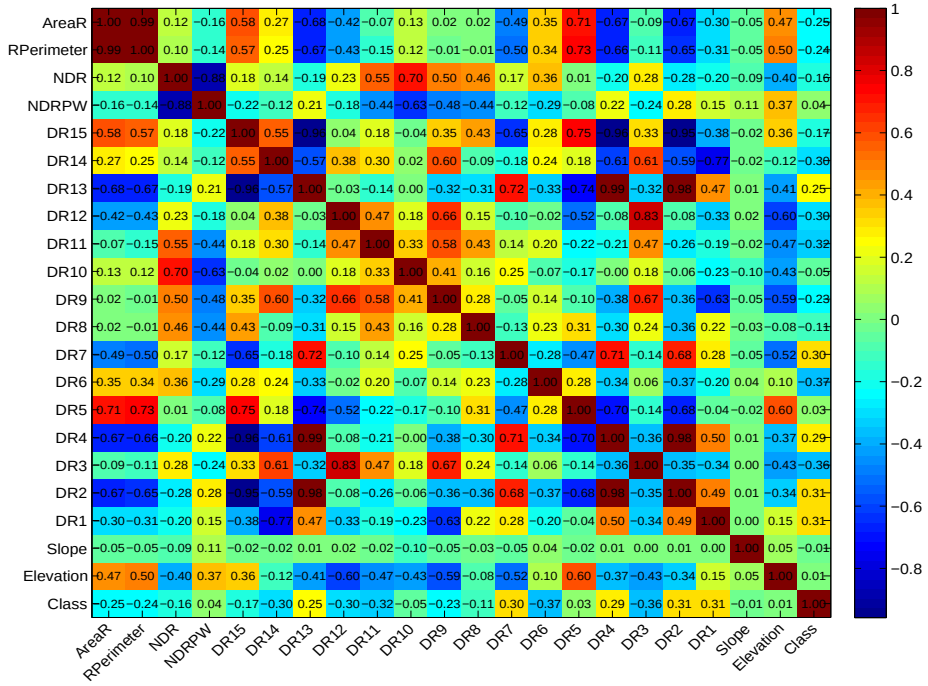


Figure 4.9: Correlation heatmap for real secondary mineral attribute data showing correlations and SAC among attributes data.

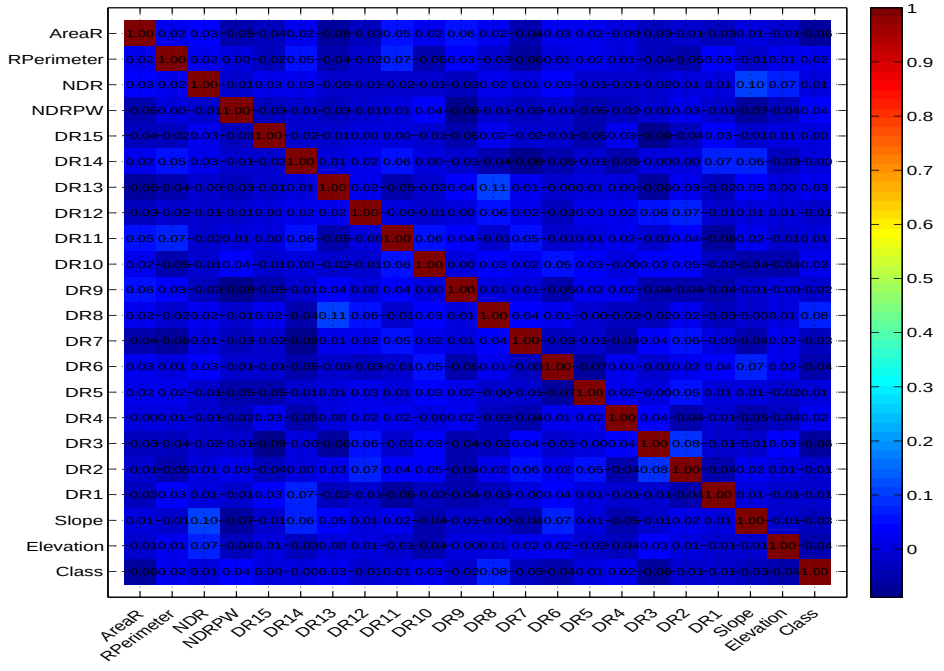


Figure 4.10: Correlation heatmap for simulation secondary mineral distribution data showing absence of correlations and SAC among attributes data.

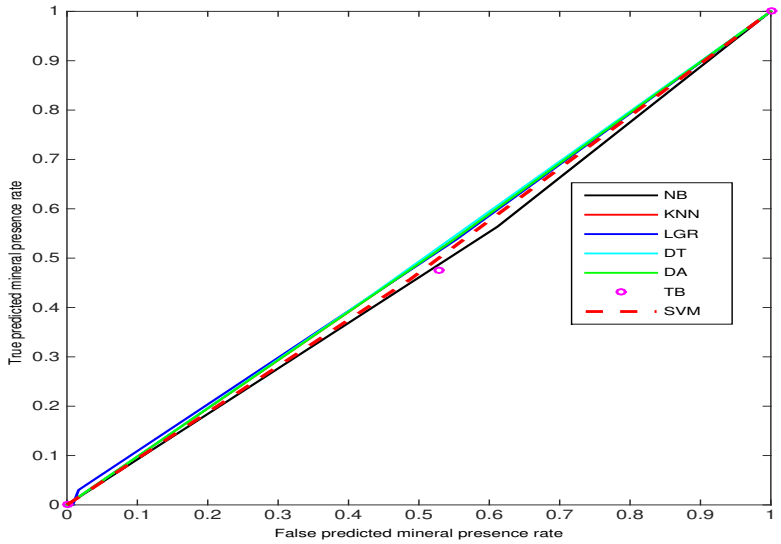


Figure 4.11: The ROC of simulated secondary mineral distribution data without SAC showing performance of the classifiers.

Table 4.6: PSM-MPM performance validation evaluation test using the simulated data in percentage (%).

Classifiers	Accuracy	Error	Sensitivity	Specificity
TB	52	48	52	53
DT	55	45	58	43
SVM	53	47	54	49
NB	46	54	44	61
DA	47	53	47	55
KNN	47	53	46	55
LGR	50	50	49	55

The simulation of real data conducted offers alternative datasets when conducting external cross-validation of data in ML. The simulation of mineral distribution data obtained from the real data obtained from the PYGR was used to validate the model built based on the real data did not perform well. The poor performance obtained using simulated data as shown in Tables 4.5 to 4.6 and also in ROC Figure 4.16, is attributed to complete absence or removal of SAC in the simulated data distribution. The simulated data attributes attained more independence as compared to the real data that shows high interdependency among predictive attributes as shown by the correlation heat map in Figures 4.10 and 4.9.

4.4.4 Discussion of PSM-MPM Performance Results using Standard ML Classification

Based on the results obtained from these experiments, seven standard ML classifiers were employed to develop a PSM using secondary mineral data of cassiterite represented by points on the map of PYGR. The ML algorithms used were: Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Decision Tree Bagging (TB), Decision Tree (DT), Logistic Regression (LGR) and Discriminant Analysis (DA). Each mineral location was considered to be a point of observation on a map, and the prediction of potential mineral deposit locations was conducted using *point-based analysis approach* as against the regular use of area plot and knowledge-based approach. The seven supervised classifiers performed differently based on their ability to model discrete types of spatial data from secondary mineral deposits, as shown in Table 4.3. The KNN and TB have predictive accuracies of over 98% (i.e., the accuracies of close to perfection), while SVM, DT, LGR, DA and PSM predictive performances were between 81% and 92%, whereas the NB has the least predictive accuracy score of about 67%.

The result of the area under the ROC curve and the misclassification rate, as shown in Figures 4.7 and 4.6, also justified the performances of the models regarding the predictive scores. The KNN and TB classifiers were preferred to others because they yielded the highest predictive accuracies, bigger AU-ROC and less misclassification compared to other classifiers. The high predictive accuracy scores recorded in the KNN and TB and least in NB suggests, a further evaluation of the results, as overfitting or underfitting often associated with exaggerated have either high or low predictive accuracy values, especially, in spatial datasets.

As earlier discussed in the previous chapters, the cause of overfitting and underfitting is the SAC inherent in secondary mineral distribution data, leads to the predictive performance results obtained. Specifically, the predictive accuracies of the KNN, TB and NB were further subjected to more evaluation to investigate the presence of either higher predictive accuracy performance of KNN and TB or poor predictive accuracy performance in the NB.

The PSM-MPM performance based on RHO selection using standard ML classification shows possible overfitting of the test data by the classifiers as presented

in the predictive accuracies scores in Table 4.3 using the standard ML classifiers. An approach to the verification and justification for the use of spatial attribute dataset or SAC in the model building through a method that deliberately isolates spatial attributes during modelling is the result of model performance shown in Table 4.4 and Figure 4.8. A summary of the analysis that supports the earlier assertion that SAC and the spatial components are very useful in modelling spatially distributed data such as the one used for this experiment was confirmed by the poor results of the predictive accuracy, despite the absence of the explicit x and y coordinates of latitude and longitude. The spatial attributes of distances indicated that it is strong enough for the ML algorithms to make good predictions using these spatial characteristics in a dataset. Consequently, simulated secondary mineral distribution datasets with complete absence of spatial attribute correlation and independent performed very poorly as indicated in Table 4.6 and ROC Figure 4.16, where the predictive accuracy scores for all the classifiers are less than 50% which are results not too different from the product of a random guess.

4.5 Implementation of Predictive Attributes Important Subsets Selection using PCA

The aim of this section is to deploy methods of attribute selection as a preprocessing technique using principal component analysis (PCA), by selecting the best predictive attributes subsets to optimise the predictive performance of the PSM-MPM in respect to the secondary mineral deposits obtained from the PYGR, and generalise well. The technique for optimisation based on feature subset selection using PCA was employed. The predictive performance optimisation procedures involve a comparison of the three classifiers through preprocessing of the predictive attribute dataset that selects the best feature subsets of the attribute dataset and reducing the size or dimension of the attribute dataset in developing new PSM-MPM.

The predictive accuracy scores of KNN, TB and NB here were determine using the subsets of selected attributes that are most important and less in dimension.

The PCA aim to investigate and reduce the effect of redundant and high auto-correlation attribute dataset on the learning and prediction. The method used RHO selection of preprocessed of reduced attribute data into training and test set to train and validate the PSM-MPM performance. The aim is to investigate how attribute data were preprocessed using PCA, can help to mitigate the unreliable predictive accuracies of the classifiers due to SAC leading to overfitting and attribute underfitting. The result of this preprocessing aim to compare the predictive performance evaluation technique of RHO on the classifiers using original data attribute without any preprocessing as well as with preprocessing. Figure 4.13 shows the correlation heat map indicating association among the natural attributes of mineral distribution that implies there is a collective influence of some mineralisation attributes to variance in the data distribution. The application of PCA will, therefore, aid in reducing the effect of correlation by capturing the natural structures that may either be similar to the original predictive attribute data.

Tables 4.7 and 4.8 shows the weight of principal components (PC) rotation values or the Eigenvalues and the percentage of variance (POV) used to determine correlation or multi-linearity between the natural mineralisation attributes that validates the selection of attributes with higher POV. A total of twenty-one components obtained from PCA corresponds to the total number of predictive attributes of the natural mineral distribution data, are identified using two class labels of 0 & 1 (mineral presence and absence). Figures 4.13 and 4.14 represents the predictive attributes PCA factor map and their variance components plot respectively. The PCA factor map extracts dominant patterns of the most important variables that explains the variations in the original predictive mineral attributes and the variance plot indicates both high and low variations among the principal mineralisation data components used to select new attribute subsets, for implementation of the optimised PSM-MPM.

The subsets of relevant attributes selected based on the highest predictive attribute contributor as shown in variable factor map using dimensions one and two of percentage values 35.6% and 23.6% totalling 58.9% as presented in figure 4.15. The indication that only five (5) attributes with highest contributive predictive

values and a strong joint correlation with the components, are selected to optimise the PSM-MPM predictive accuracy. To achieve generalisation, a PSM-MPM overfits the training set using test set if it has low bias but high variance. The aim here is to avoid overfitting of the dataset by the ML classifiers used; it is expected that only the most relevant predictors or attributes will be selected. The attributes selected include the area or size occupied by the rocks type; the various distances between mineral points (observation point) to the nearest selected four different type of rocks labelled totalling five attributes as; AreaR, DR15, DR13, DR4 and DR2. An improvement in the PSM-MPM accuracy performance when using the five sub-selected attributes is evident in the result of model performance given in Table 4.9.

Some of the correlated features are, however, retained in the selected attributes based on the PC output selection. The correlation heatmap and factor important plot represented in Figures 4.12 and 4.15 respectively, selects the few subset used for modelling. The aim of the selection is to use only the newly selected best attributes to reduce the effect of SAC in the dataset and test for the overfitting and underfitting effect in the model performance.

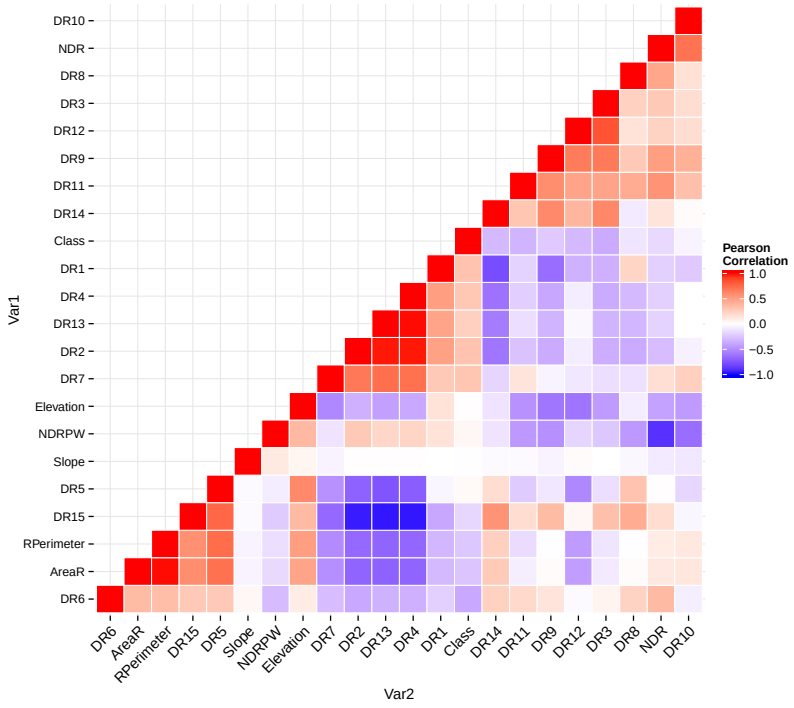


Figure 4.12: Correlation heat map of secondary mineral predictive attributes for PCA.

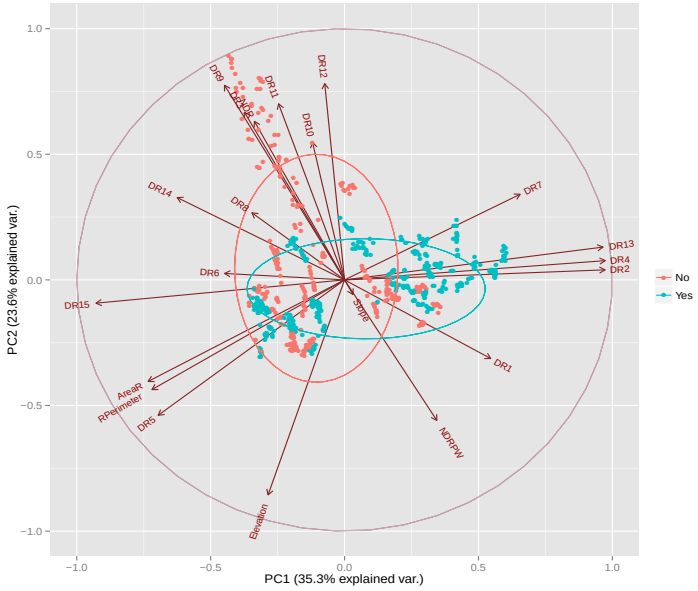


Figure 4.13: The mineralisation attribute factor map for PCA showing the best attributes selection.

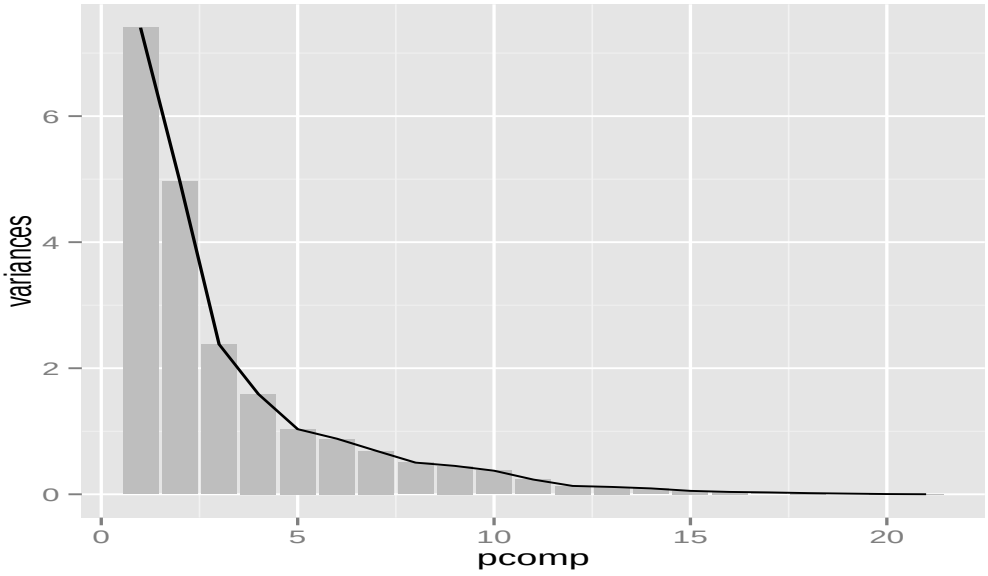


Figure 4.14: Variation among attribute components plot.

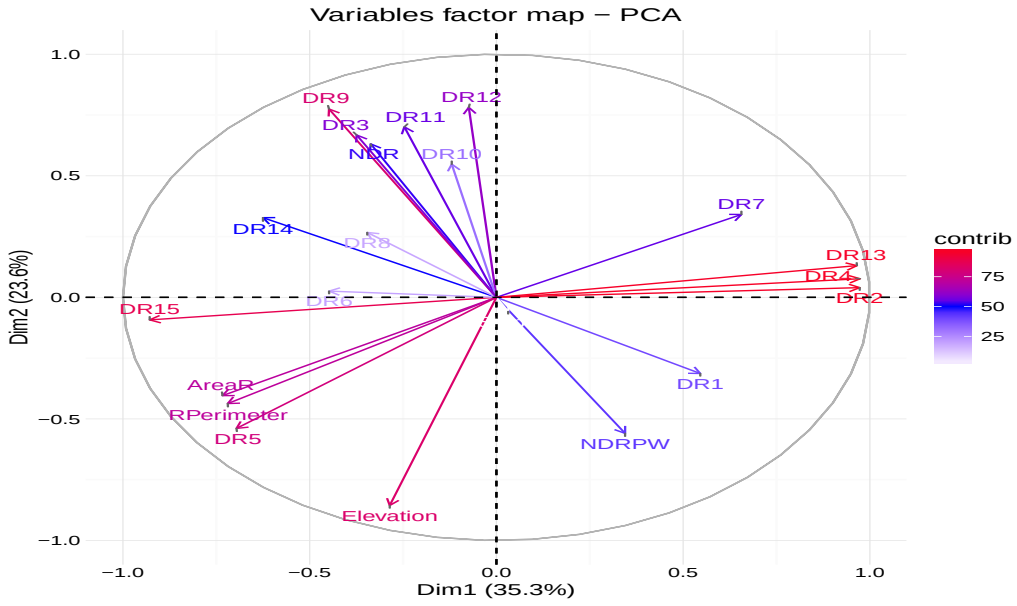


Figure 4.15: The factor importance plot based on the PCA showing level of contribution among attributes component.

Table 4.7: The eigenvalues with percentage of variance and the cumulative percentage of variance for all attributes principal components CP

PCs	Eigenvalue	% of variance (POV)	Cumulative % of variance (CPV)
comp 1	7.4034989963	35.254757125	35.25476
comp 2	4.9659675100	23.647464333	58.90222
comp 3	2.3791870941	11.329462353	70.23168
comp 4	1.5897132503	7.570063097	77.80175
comp 5	1.0331147638	4.919594113	82.72134
comp 6	0.8816541533	4.198353111	86.91969
comp 7	0.6899077425	3.285274964	90.20497
comp 8	0.5024497182	2.392617706	92.59759
comp 9	0.4504138472	2.144827844	94.74241
comp 10	0.3739398757	1.780666075	96.52308
comp 11	0.2333933303	1.111396811	97.63448
comp 12	0.1324428230	0.630680110	98.26516
comp 13	0.1165870270	0.555176319	98.82033
comp 14	0.0931832273	0.443729654	99.26406
comp 15	0.0538523175	0.256439607	99.52050
comp 16	0.0373157981	0.177694277	99.69820
comp 17	0.0297229992	0.141538092	99.83974
comp 18	0.0185060203	0.088123906	99.92786
comp 19	0.0107919475	0.051390226	99.97925
comp 20	0.0036118771	0.017199415	99.99645
comp 21	0.0007456812	0.003550863	100.00000

Table 4.8: Correlation of predictive variables against selected principle components.

Attributes	PC1	PC2	PC3	PC4	PC5
AreaR	0.73354982	-0.40435759	0.251825084	-0.34610072	0.029255598
RPerimeter	0.71943518	-0.43694790	0.243794924	-0.36189503	0.025616847
NDR	0.33658624	0.63104474	0.618092564	-0.04966323	0.106914233
NDRPW	-0.34454167	-0.55935100	-0.624741359	0.04397388	-0.033560783
DR15	0.92840850	-0.09268335	-0.078579969	0.22075352	- 0.124246323
DR14	0.62504269	0.32778403	-0.478213139	-0.28711014	0.030382437
DR13	-0.96527875	0.13054631	0.093798784	-0.09328861	0.067641709
DR12	0.07400604	0.78124719	-0.477360678	0.19492788	-0.023937659
DR11	0.24650745	0.70098574	0.123055730	0.21129109	0.075876451
DR10	0.11962029	0.54748629	0.513717275	-0.39046999	-0.116022965
DR9	0.44820869	0.77462731	-0.152644559	-0.10410722	-0.034936730
DR8	0.34588514	0.26706338	0.419372924	0.70655915	-0.086948185
DR7	-0.65590178	0.34092028	0.293432431	-0.21403409	-0.018990845
DR6	0.44811287	0.02560183	0.212681380	0.06847590	0.543703155
DR5	0.69601942	-0.53910788	0.230725348	0.11795953	-0.081580438
DR4	-0.97358485	0.07655319	0.122083365	-0.07728795	0.057029965
DR3	0.37420690	0.66759678	-0.408993398	0.17262301	-0.083662360
DR2	-0.97286200	0.03995076	0.044381605	-0.09543704	0.031217943
DR1	-0.54585414	-0.31317177	0.375542992	0.55196117	-0.108260353
Slope	-0.03142353	-0.05805935	-0.156260763	0.14868681	0.802466259
Elevation	0.28565021	-0.85581806	-0.008311395	0.11600546	0.003159617

4.5.1 Result Discussion on Attribute Subset Selection using PCA

The results of the three classifiers performance based on attribute subset selection are as shown in Table 4.9 and ROC curve plot Figure 4.16. The results which indicate that the preprocessing that led to attribute sub-selection when modelling PSM-MPM optimised the predictive accuracy of performance of only TB and KNN with accuracy scores of 96% and 99% respectively but failed to improve the predictive performance of NB which has accuracy score of 59%. The optimisation was measured and verified based on the initial performance by the simple RHO using same standard ML classifiers using natural datasets without any preprocessing as obtained in Section 4.3.

The classification technique of subset attribute selects used fewer and best-contributing attributes that performed best for KNN and TB except for NB. Because the approach provided a learning procedure similar to the RHO with the selected attributes values showing a high correlation as indicated in both table 4.9 and figure 4.12. The results also showed that using attributes important selection in a spatially distributed dataset only further simplified the predictive data complexity and thereby ease learning ability of the complex algorithms such as the KNN and TB because, it uses lesser predictive attributes and still less independence, therefore, but failed to address the problem SAC inherent in the datasets. The predictive performances of the three classifiers clearly indicated that KNN and TB are still overfitting while the NB seem to be underfitting or under-performing as shown in the performance table 4.9. The PCA technique of attribute subset selection showed that it is only meant to improve the predictive accuracy scores of the classifiers and not to optimise the overall performance to acceptability and reliability. The NB predictive accuracy rating is however, lowered —i.e., and present yet another pessimistic score too despite the data preprocessing aimed to improve its performance, the reason is clearly due to the existence of high correlation (SAC) among the selected attributes as shown in the correlation heatmap Figure 4.12. The NB assumption of attributes independence is violated using PCA and hence, failed to improve the predictive performance of the NB classifier. Only a method that eliminates or reduce spatial correlation

among attributes can significantly influence the predictive performance of the NB classifier in a spatially distributed dataset such as the secondary mineral data. It is, therefore, necessary to consider the effect of SAC when building PSM. The predictive accuracy optimisation showed bias towards the spatial arrangement of the dataset as seen in the performance Table 4.9 and Figure 4.16.

Table 4.9: PSM-MPM performance table for TB, NB and KNN based on selected attribute subset in percentage (%).

Classification Method	Accuracy	Error	Sensitivity	Specificity
Tree Bagging	96	4	98	94
Naive Bayes	59	41	64	55
K-Nearest Neighbour	99	1	100	98

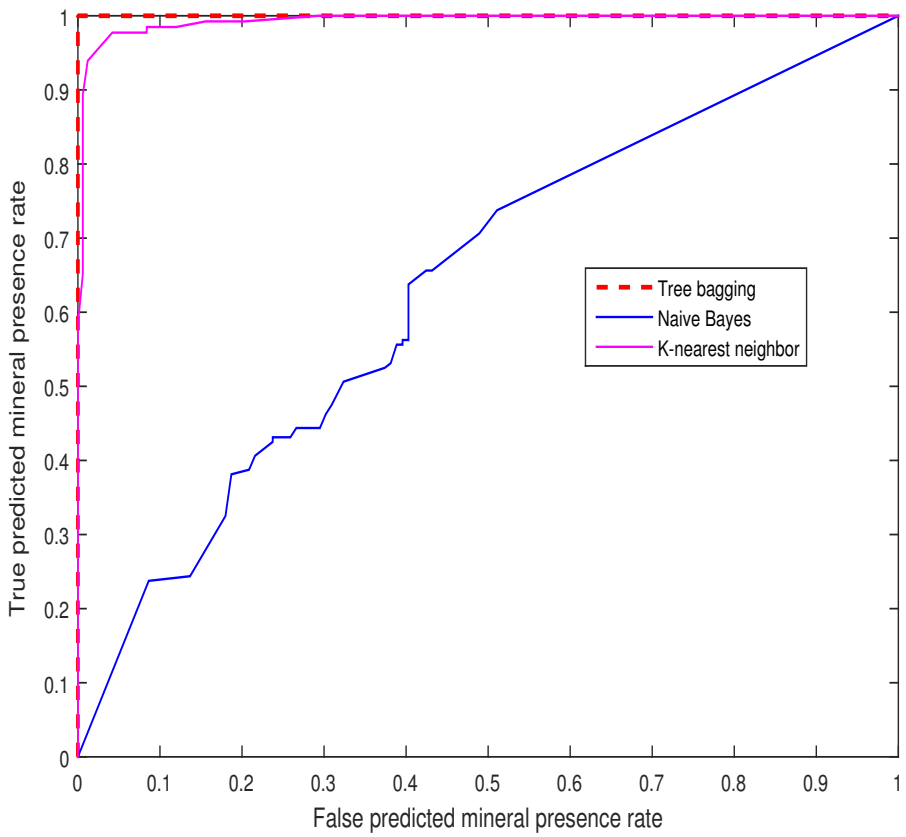


Figure 4.16: The PCA ROC plot showing the PSM-MPM performances of KNN, NB and TB classifiers based on.

4.6 Implementation of Novel Approach for PSM-MPM Performance Evaluation

A four-way data sampling for PSM-MPM validation assessment was conducted to address overfitting and underfitting as summarised in Figure 4.17.

Firstly, a method of re-substitution uses the entire dataset for training is then repeated as test data. The performance shown in Tables 4.10 to 4.12 indicated a 100% predictive accuracy for both KNN and TB but a predictive accuracy score of about 71% for the NB. The re-substitution method is considered the traditional approach for measuring the goodness of fit for a classifier; and it is apparent from the result displayed in Tables 4.10 to 4.12 that TB and KNN perfectly fit the datasets used which indicated that the classifier is overfitting datasets while the NB classifier does not fit the dataset well and as such considered to be underfitting.

The second evaluation results involve the use of a conventional RHO data sampling method with predictive accuracy scores of 98% and 97% for both KNN and TB while NB recorded about 67%. The scores are suspiciously high and are probably influenced by SAC in spatial data modelling in respect to KNN and TB. The results for the NB remain unsatisfactory, which is clearly due to the violation of attribute independence, also influenced by SAC. It is indeed, necessary to develop a method that better validates the PSM-MPM through the spatial split of attributes selection, and that identifies and addresses the concept of overfitting in PSM-MPM.

In addition to the two sampling evaluation methods explained above, a strip spatial approach divides into third and fourth paths that are similar but implemented differently, with the third method of splitting the entire dataset into half longitudinally while the fourth, strip-split the dataset into quarters longitudinally for the spatial selection of training and testing. The strip splitting along the longitudinal approach is done to allow each resulting part to contain the same type of attributes in locations within the range of the study area. The results of the investigation followed the adoption of the methodology explained in Chapter 3 and summarised in Figure 4.17. The misclassification result represented in Figure 4.18 shows the diagrammatic performance of the confusion matrices of the quarter spatial strip split validation of PSM-MPM using the three classifiers. The

predictive performance of the four ways approach to PSM-MPM validation using KNN, TB and NB is as shown in the Tables 4.10,4.11 and 4.12 respectively. The ROC performance was also represented diagrammatically in Figure 4.19.

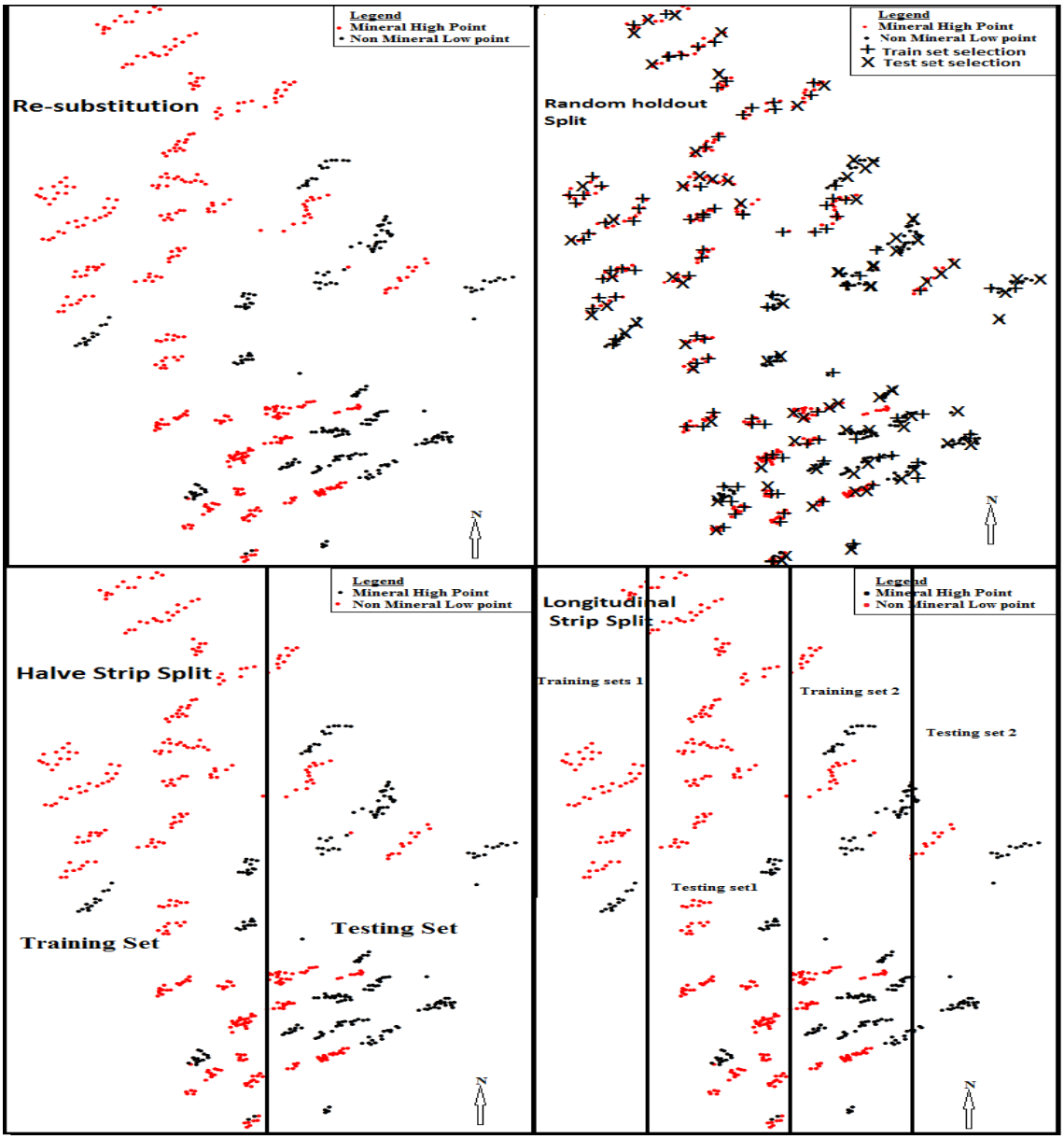


Figure 4.17: A diagrammatic representation of the four-way validation technique of Re-substitution, Random holdout, Halve strip and Longitudinal quarter strip methods of splitting secondary mineral data from the PYGR. The vertical lines represent the strip splitting of data method while the **x** and + signs symbols in the RHO split represents the split into training and test set respectively.

The result shows that the longitudinal strip splitting approach presents a better data splitting method for model validation to adopt in building PSM-MPM. The method of SSS reduces the effect of SAC due to clustering and increases spatial variation (covariance) in the spatial data attribute values along the spatial split, thereby reducing the effect of SAC in the data when selecting training and test data to assess the performance of the PSM-MPM. As against the random selection, which chooses attributes in a RHO selection and learns spatial datasets from clusters of points that are spatially autocorrelated.

Figures 4.19 and 4.18 shows that the KNN and TB algorithms still performs well despite the spatial splitting while the predictive accuracy of the NB classifier improved slightly. This result agrees with the work of Bahn & McGill (2013) but with some improvement that handles underfitting, by improving the NB predictive performance from the performance obtained using RHO and re-substitution (Ibrahim & Bennett, 2014a). The small drop in the predictive accuracy of KNN and TB was due to the slight reduction in the level of SAC among the attribute values forced apart. The slight improvement in the predictive accuracy of the NB classifier is based on the presence of more independent attribute data selection, and higher covariation among the predictive characteristics in the datasets. Attribute data independence values supports the underlying assumption of NB algorithm. The attribute spatial freedom is considered a major plus to the work which was not recorded or discovered by Bahn & McGill (2013), which was limited to investigated overfitting only. The SSS data sampling technique for PSM-MPM performance validation addressed both overfitting and underfitting performance by the classifiers used, as reflected in the results of the PSM-MPM predictive accuracy scores shown in Tables 4.10, 4.11 and 4.12. The quartered strip-splitting technique, or longitudinal strip splitting, offers a more reasonable predictive model accuracy score (performance) when evaluated against the other three methods of model evaluation techniques presented. Overall, therefore, the SSS is a more promising approach to validating spatially auto-correlated data distribution models since it detects and handles both overfitting as well as underfitting in secondary mineral data distribution better than the conventional RHO and PCA-RHO techniques of data sampling for model optimisation and generalisation.

Table 4.10: Four-way PSM-MPM performance validation comparison scores table for KNN in percentage (%).

Split Method	Accuracy	Error	Sensitivity	Specificity
Re-substitution	100	0	100	100
Random holdout	98	2	100	97
Half spatial strip	43	57	30	70
Quarter spatial strips	93	7	80	99

Table 4.11: Four-way PSM-MPM performance validation comparison scores table for TB in percentage (%).

Split Method	Accuracy	Error	Sensitivity	Specificity
Re-substitution	100	0	100	100
Random	97	3	98	95
Half strip	40	60	11	100
Quarter Strips	85	15	74	91

Table 4.12: Four-way PSM-MPM performance validation comparison scores table for NB in percentage (%).

Split Method	Accuracy	Error	Sensitivity	Specificity
Re-substitution	71	29	62	78
Random	67	33	50	81
Half strip	47	53	25	92
Quarter Strips	74	26	73	74

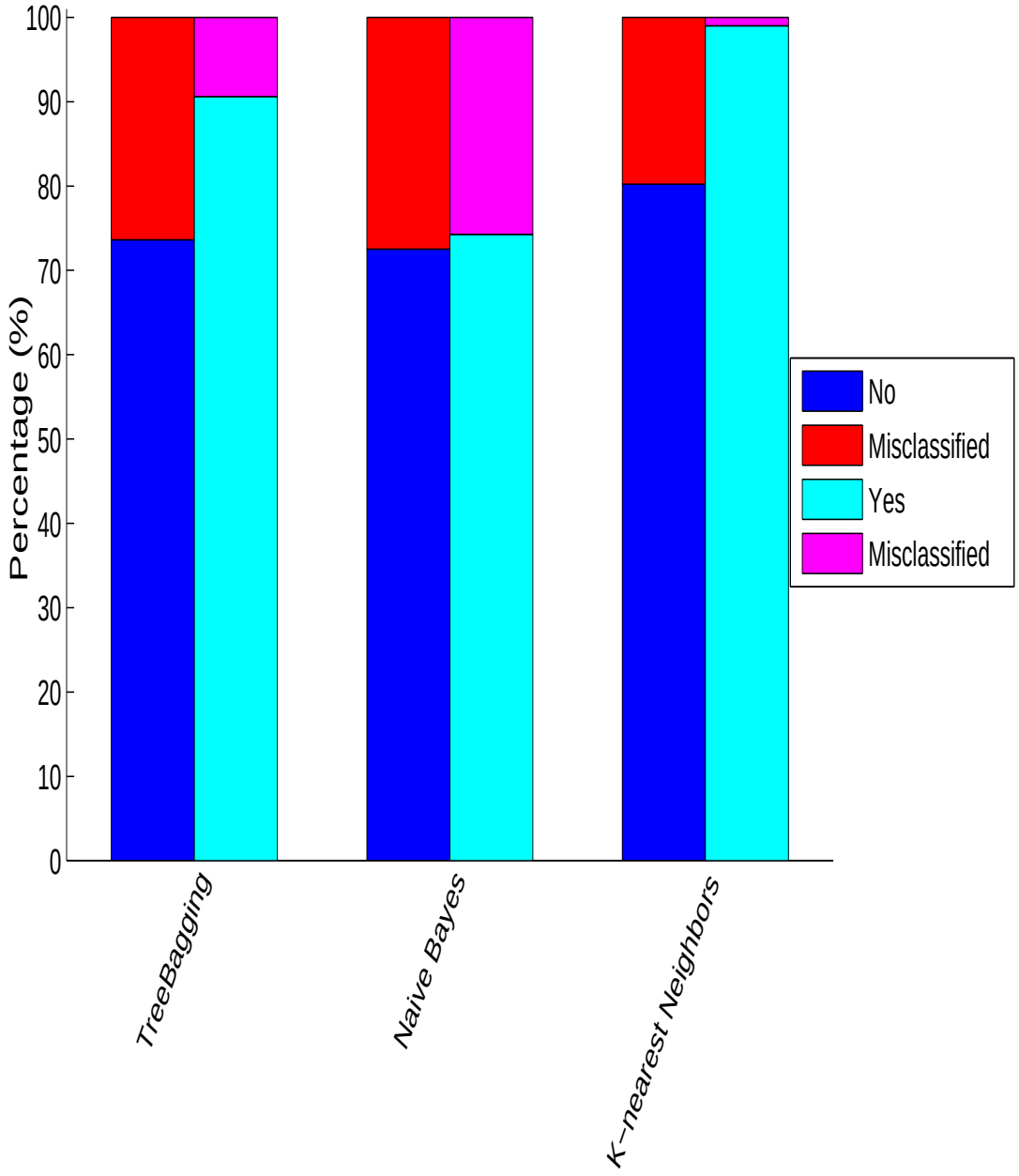


Figure 4.18: Misclassification performance bar chat for spatial strip-splitting validation using KNN,TB and NB classifiers.

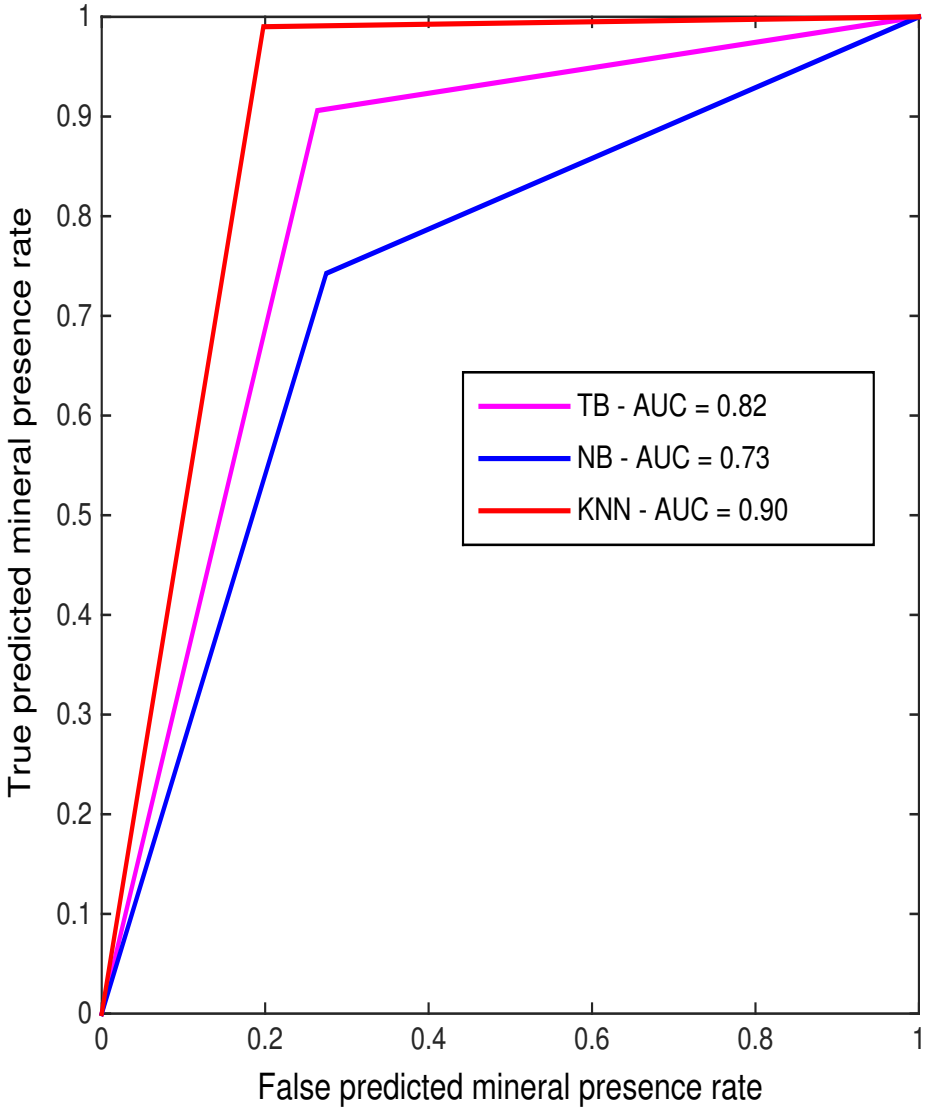


Figure 4.19: ROC performance curve plot for KNN, TB and NB algorithms using strips split.

4.6.1 Result Discussion of Novel Technique of PSM-MPM Performance Evaluation

The evaluation results of PSM-MPM predictive performance appraisal techniques through model validation indicated changes in predictive accuracy scores through the spatial strip split (SSS) method. These performances results illustrate how important the selection of training and testing data is when judging the predictive performance of a distribution model. The re-substitution technique was used for measuring the goodness of fit of a model but shows a high level of overfitting or underfitting and should be avoided. A similar impressive measure of predictive accuracy performance presented using KNN and TB, but the performance of NB was weak as judged by the random hold out the testing scheme, which still indicated a measure of PSM-MPM overfitting and underfitting.

The random selection involves splitting from clusters that select at a certain range based on attributes that are similar due to their closeness, but could not replicate on the test set, thereby causing overfitting or underfitting. Several data points may have, for example, same elevation values but different coordinate locations. Equally, the nearest distances to various geological attributes such as rocks types may have similar distance values even when the points are from different directions or locations. The validation of PSM through the RHO of attributes into training and testing sets failed to mitigate the effect of SAC and still return a very high predictive accuracy indicating score overfitting and underfitting, depending on the type of classifier or the learning algorithm used (i.e., either a complex or a simple algorithm). The results are evident by the predictive performance of KNN, TB and NB. In the case of KNN and TB, which are very efficient but complex, it easily learn from a mostly homogeneous values in clustered dataset through random split using the majority voting system of predictive modelling and validating with similar test sets, but this was unable to tackle overfitting when using RHO selection. Whereas in the case of NB, the range of attributes data was more independent, by spatially separating the datasets and allow the selection of training and test sets separately to improve the predictive ability of the model and address underfitting in the datasets.

The overly optimistic validation performance test results of PSM-MPM using RHO split, tested on a not entirely independent dataset, was illustrated in the diagram Figure 3.11 and in the summary results in Figure 4.17. The predictive accuracy score was measuring close to perfection (about between $97 - 98\%$) for both KNN and TB but around 67% for NB. When the PSM-MPM was tested on truly independent and spatially segregated data, however, it fared better in the intermingled strip-split approach than in the halves approach. The result of the predictive performance of spatially separated validation data was an indication of the remaining effect of SAC along the segregation lines. There was only one separation line in the half-splitting method and that account for the low accuracy score in the half split, but three in the strips approach, and this reduces the problem when the models are used to predict into a new area, as shown in Tables 4.10, 4.11 and 4.12. One explanation for the change in predictive power of the models when training and test datasets were split geographically is the absence of a genuinely independent testing data. The introduction of a rather dramatic geographical segregation effectively broke the dependence among the predictive attributes caused by SAC, thereby resulting in very poor predictive accuracy performance scores of less than 50% for all the classifiers, as shown in Tables 4.10, 4.11 and 4.12.

The results for the performance of each classifier represented by the AU-ROC value in Figure 4.19, and the predictive misclassification presented in Figure 4.18, however, both indicated a better and a more realistic predictive performance compared to the random split method and the re-substitution which overfits. The NB model also recorded a slight improvement in the predictive accuracy, as well as in the size of the area under the ROC plot, when the datasets are divided quarterly along longitude called spatial strip split. The small increase in the predictive accuracy score by the NB classifier was attributed to the most independent values among the predictive characteristics through spatial separation and reducing the effect of SAC in the dataset. The attribute independence is a fundamental assumption of the NB classification model.

The basic idea behind the SSS technique is reducing the effect of the homogeneity (dependence) that exist in a clustering arrangement when learning on the predictive data sets, which will allow the splitting of training and test datasets

to be truly two independent datasets. Since SAC is principally a concept of distance related similarity in datasets, and secondary mineral datasets are spatially distributed in clusters, the method of dealing with the problem of SAC in spatial datasets must be spatially distinct. The SSS sampling of training and test data in space allowed for a more independent dataset that results in high covariance among attributes values, rather than selecting through a RHO validation technique that results in a PSM-MPM that indicates overfitting or underfitting the datasets. It is hard to eliminate SAC in spatial data modelling while retaining high predictive accuracy, but it is possible to evaluate the performance of PSM-MPM using a carefully pre-processed method of splitting data sets into training and testing datasets to reduce the detrimental effect of SAC in PSM-MPM.

4.7 The Comparative Analysis of RHO, PCA-RHO and SSS Validation Performance Technique Results

The comparative analysis of PSM-MPM performance based on the three approaches used in this research was aimed at determining the ideal approach to adopt when developing a PSM-MPM that generalises and possess high predictive accuracy with consideration for overfitting and underfitting.

The comparative analysis discusses the interpretation of the various experimental results which include the three techniques deployed to model and evaluate the performance of PSM-MPM and attempts to optimises the predictive accuracy that considers the effect of SAC leading to overfitting and underfitting. Table 4.13 represent the PSM-MPM comparative performance scores, which include: accuracy and error rate of KNN, TB and NB algorithms. The techniques of RHO using the original dataset without preprocessing and the attribute subset selection have an over-exaggerated predictive accuracy score between $96 - -97\%$ and $98 - -99\%$ for TB and KNN classifiers respectively. The NB classifier, however, shows low predictive scores between $59 - -67\%$. These scores clearly indicate overfitting and underfitting of the data by the classifiers since the techniques still allow random selection of spatial attributes leading to poor separation of training

and test datasets; by allowing the datasets to be either too independent or entirely dependent depending on the classifier's method of learning and sampling selection between training and test set.

Only the SSS technique that split attribute data spatially in a quarterly or longitudinal dimension shows a more positive predictive accuracy scores of 85% and 93% for TB and KNN. But gained a significant improvement in a the predictive accuracy of NB from 59% and 67% to 74% which is more optimistic (i.e., accuracy scores that is not outrageously high). The predictive score surpasses the performance achieved by the RHO validation technique used on the original dataset without preprocessing and the method of preprocessing to select attribute best subset. The performance of PSM-MPM produced using the SSS sampling method in this work, does not only determine an optimistic predictive score concerning overfitting problem as agreed by Bahn & McGill (2013) but also improved predictive performance regarding the *underfitting*.

The results clearly show that the spatial separation of training and test dataset is the ideal approach to sampling and validating model performance in spatially distributed dataset, such as secondary mineral deposits. Because it allows for a truly spatial independent as well as the intermingling of attribute datasets with real correlation to be split in both the training and testing set that helps in generalisation and not to be randomly split, since that could lead to bias selection that is overly influenced by high SAC alone.

Table 4.13: Comparative predictive performance analysis table of RHO, SSS and PCA-RHO techniques for PSM-MPM using TB, KNN and NB in percentage (%).

Technique	Algorithm	Accuracy	Error	Remarks
RHO	TB	97	3	Pessimistic High/Overfitting
	KNN	98	2	Pessimistic High/Overfitting
	NB	67	33	Low/Underfitting
PCA-RHO	TB	96	4	Pessimistic High/Overfitting
	KNN	99	1	Pessimistic High/Overfitting
	NB	59	41	Pessimistic Poor/Underfitting
Quarter-SSS	TB	85	15	Optimistic high
	KNN	93	7	Optimistic High
	NB	74	26	Optimistic & Improved

4.8 The Contributions of the Thesis

The contributions of the thesis as implemented in this chapter are as follows:

- The thesis established a unique and systematic technique for acquiring secondary mineral occurrence attribute datasets. These characteristics included the identifying and collection of secondary mineral presence deposit location of presence and absence as coordinate points; lithological positions and dimension; spatial components such as relative distances between mineral deposit points and geological features and determined the relationship among attributes presented in a form applicable to ML classification algorithm.

  The work designed and implemented a systematic approach to a geodata collection, by first conducting a geological survey of all the mining points in an attempt to tackle data paucity in the PYGR. The desired data collected included a geological map of the area and the coordinate location of points for all the existing mining areas (i.e., latitude, longitude and elevation). Other data collected included the historical mining information of the area, particularly in respect to the presence or absence of minerals. The analogue data sets collected are rarely useful for scientific research such as this unless converted to digital format. Using some specialist equipment like Global Positioning System (GPS) tools, the exact position of the mining pits were obtained and plotted on the geological map of the area before converting the map data to digital. By using GIS tools, the predictive spatial attribute data extracted from the established mining locations depicted as points alongside other geological or geographic features such as rock type, size and spatial distances between the mining points and the geological features were all extracted. The data collection procedure is unique as it uses a systematic method applicable to areas with geodata paucity. It is also the first time an attempt has been made properly to document the secondary mineral occurrence data of the PYGR area in a digital format. The systematically collected and recorded digital geodata of the PYGR was used to build a PSM-MPM. The data collection process was possible because of the ability to translate digital map data into binary or numeric weighted values in GIS

and export the predictive data in an ML support format into an algorithm using data mining or ML classification software such as MATLAB, R or WEKA, used to build PSM-MPM. The use of GIS for data processing was vital because it can handle (i.e., manipulate and store) geospatial data effectively in a spatial frame of reference (i.e., space dimension) and specially presented as labelled point dataset for spatial modelling using supervised ML techniques (Ibrahim & Bennett, 2014a,b).

- The thesis developed point-based predictive spatial models for secondary mineral potential mapping (PSM-MPM) using standard ML classification algorithms and evaluate their predictive performance through RHO data sampling validation technique (Ibrahim & Bennett, 2014b).

  The work designed and developed a point-based PSM-MPM using standard ML classification algorithms capable of capturing the spatial relationships among mineralisation attributes by learning the distribution pattern of the secondary mineral deposits of the PYGR, represented as points, with other mineralisation features to make the prediction of potential mineral occurrence points. Part of the uniqueness of this contribution is in the design and implementation of the PSM-MPM as a point-based data approach mapping from GIS, in contrast to the traditional area-based or polygonal data approach. The method combines both theoretical relationships derived from the various literature on mineral distribution components and points statistical analysis using techniques of distance point distribution, to determine the distribution pattern of the secondary cassiterite mineral deposit distribution. This analysis is part of the exploratory data analysis to determine data applicability to modelling before deploying ML classification algorithms to train from point observations, and validate with the test datasets. Individually, a survey of seven supervised ML algorithms was used to build the PSM-MPM because of their popularity (i.e., well-documented) and their ability to model binary dataset. The selected algorithms include; Naive Bayesian (NB), Bagged Decision Tree– Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) and Logistic Regression (LGR). However, only three classifiers were

selected out of the seven, for further evaluation due to apparent an exaggerated high predictive accuracy scores (overfitting) and a poor predictive accuracy scores (underfitting) associated to the geospatial secondary mineral data distribution. The three classifiers selected for evaluation implemented are KNN, TB and NB using RHO sampling or splitting validation technique.

- The thesis determined the detrimental effect of spatial attribute and SAC to predictive performance of PSM-MPM through firstly, the deliberate isolation of the spatial component of the datasets and secondly, the simulation of real mineral distribution datasets without SAC.

  The work evaluated the importance of predictive spatial attributes to the performance of PSM-MPM, by comparing the predictive accuracy scores of each classifier produced using the following: all the attributes datasets, no spatial attributes and attribute datasets only, then compare their performances to show the importance of spatial characteristics in PSM-MPM. The evaluation technique used a style that first eliminated all the space attribute in the datasets and validated the predictive accuracy of the model produce, with the results of the model developed containing spatial attributes in the datasets. This predictive performance evaluation investigated the effect of the spatial components of the dataset to model performance. The distribution datasets are mostly dependent on each other due to proximity in space, leading to high SAC among the attribute data values, causing the classifiers to show overfitting or underfitting. The result indicates that the spatial attribute datasets are very relevant to the predictive accuracy of the PSM-MPM.

  The work equally developed a method that investigated the relevance of SAC and spatial distribution as a predictive component of the PSM-MPM, using simulated mineral distribution datasets obtained from the PYGR. The secondary cassiterite mineral dataset was simulated to generate a new set of data that eliminated the spatial distribution components inherent in the data sets and used the new dataset as test sets for validation of the predictive performance result of the PSM-MPM using ML classification algorithms. The result of the evaluation identified the importance of spatial

correlation among the predictive attribute data, for making a better judgement about model performance and generalisation due to the effect of SAC in the dataset.

- The thesis developed and implemented a PCA-RHO, a method of data preprocessing that selects the best attribute subset of the original data structure using principal component analysis (PCA), before applying the RHO. The method attempts to assess the performance accuracy of the ML classifiers used to develop a PSM-MPM that can eliminate the adverse effect of SAC leading to overfitting and underfitting.

  The PCA technique deployed an attributes data preprocessing to select only the best predictive attributes subsets. The process involved standardisation of the mineralisation variable values obtained at similar relative scale and prevented some variables from becoming prevailing due to overriding large measurement units. A combination of correlation coefficient using heat map plot, eigenvalues and the contribution of PCs variable factors was used as a yardstick to select attributes most important subsets using PCA. Since the problem of overfitting and underfitting in spatial characteristics dataset was attributed to high correlation among attributes data in space (SAC) causes greater similarities and violates the data attribute independence. The PCA preprocessing was employed first as against the application of using real raw dataset, to reduce the size dimension of the natural datasets before applying the RHO sampling technique, in an attempt to correct the violation of attribute independence caused by SAC.

  By selecting only the attributes that are most important and removing highly dependence attributes that may influence the performance of the ML classifiers, the technique of data preprocessing using PCA was found to show high predictive accuracies for TB and KNN (exaggerated precision) and a poor predictive accuracy for the NB classifier. The results of the predictive performance accuracies using RHO splitting with and without data preprocessing are similar and failed to address the problem of overfitting and underfitting. The predictive performance shows that the technique of PCA has failed to tackle the issue of SAC, unable to mitigate the adverse

effect of both overfitting and underfitting in spatially distributed dataset and specifically in PSM-MPM.

- The thesis proposed a four-way assessment performance technique for PSM-MPM performance evaluation that evaluated the conventional methods of data splitting for evaluation with a new method of spatial splitting of attribute datasets into training and test set for model performance assessments. The technique demonstrated through an expanded pseudo-codes (algorithm 1 - 4) designed in this work, that evaluated the PSM-MPM predictive performance. The method compares the predictive accuracy scores produced from each of the three selected classifiers (KNN, TB and NB) using four different data sampling method in the four different PSM-MPM validation methods. The four-way sampling approach included: re-substitution, random holdout, half longitudinal spatial split and quartered longitudinal spatial strip split (SSS) techniques. The method of spatial splitting to asses PSM-MPM performance, offers a better approach to detecting the detrimental effect of overfitting and underfitting, associated to spatial data or affected by SAC in ML classification, and presented the ideal predictive accuracy scores for PSM-MPM (Ibrahim & Bennett, 2014a).

  The method of attribute data sampling along space that involves the stripping of attribute data into training and testing for validation, is an extension of the work of Bahn & McGill (2013), the work was used to test the predictive performance of distribution models by evaluating the performance of animal species distribution in a geographical locations. The method used in the previous work considers species distribution data as a continues spatially distributed dataset (Bahn & McGill, 2013); the work was found to account for SAC, which allows predictive characteristics data to be more independent during model validation, to test for the presence of overfitting only. However, the technique adopted in this work uses point based ML approach to modelling and prediction of secondary mineral distributed dataset, which is a discrete spatially distributed dataset, the validation sampling technique of splitting the training and test data spatially ensures the

mineralisation attributes in training, and test dataset are relatively independent and spatially separated. Since the primary cause of SAC in secondary mineral deposit dataset is the proximity or nearness of data attributes values in space, as proposed by the Tobler's law of geography. SAC is considered, therefore, a primary challenge in PSM-MPM causing attribute values to be closely similar, leading to *overfitting and underfitting*, which the traditional random holdout or cross-validation has not been able to detect or tackle adequately in ML classification.

The Spatial Strip Splitting (SSS) technique was considered to be the most appropriate technique in PSM-MPM validation approach among the listed techniques. The method validates the work of Bahn & McGill (2013) in a different domain of datasets (i.e., mineral occurrence distribution data) by evaluating PSM-MPM performance due to overfitting in KNN and TB performance, the technique also addresses the effect of underfitting by limiting the the accuracy of PSM-MPM to a more optimistic and a more reliable performance and also optimising the predictive accuracy of poorly performing NB classifiers, as shown in the implementation of SSS technique results. The primary effect of SAC that leads to poor model performance or overfitting and underfitting is in the distribution arrangement of the dataset. An SSS sampling technique of validation is considered the most ideal and novel approach to PSM-MPM predictive performance evaluation for ML classification at the expense of the RHO technique of validation sampling often used for assessing the performance of mineral potential modelling (Porwal, 2006). The novel SSS method of sampling training and test set for model performance validation was found to determine that SAC mitigates the performances of PSM-MPM and causes not only overfitting but also *underfitting* respectively. The technique of SSS equally showed promising performance better than the method PCA-RHO, that supports data pre-processing. The results of the comparative analysis of PSM-PMP performance evaluation techniques that involves RHO, PCA-RHO and the quarter longitudinal SSS to determine the ideal method that best determines and detects the detrimental effect overfitting and underfitting of the PSM-MPM due to SAC in the dataset.

A comprehensive four-way assessment of PSM-MPM performance which also include the SSS technique of attribute data splitting for model performance evaluation is a significant contribution to this work because it helps to detect the presence and limit the detrimental effect of both overfitting and underfitting in PSM-MPM due to SAC through the predictive performance of the classifiers. The process leads to the adjustment of over-exaggerated predictive model scores to a more optimistic predictive accuracy scores, as well as optimised the predictive performance of some poorly performing ML classifiers better than other validation splitting techniques methods such as the RHO. The work showed the effectiveness of the spatial splitting of datasets approach in detecting the adverse effect of overfitting as well as underfitting in ML classification.

4.9 Summary

The spatial and non-spatial exploratory data analysis conducted here, tested for complete spatial randomness in the occurrence of mineral dataset and indicated that non-random mineral distribution patterns are capable of deploying learning techniques such as ML to develop a predictive model. The non-randomness of mineral occurrence distribution patterns confirms that the distributed point data is truly representative of the mineral formations in the PYGR area and that the patterns can be learnt. It is also an indication that a systematic process is the result of the points and mineral occurrence represented by points does not exist by chance (i.e., non-random occurrence). There is also a strong spatial correlation between geological features of granite rocks and mineral deposit locations, represented as points in the PYGR area. This correlation implies the likelihood that granite rocks are the primary source of cassiterite deposits in the area, in agreement with the literature describing the main source of the mineral type (i.e., secondary cassiterite). Therefore, the lithology (granite rock) was considered to be an important predictive attribute of the geological component of mineralisation in the PYGR and forms the basis for the spatial distance analysis between data points and the geological features to extract the potential attributes data for predictive mineral deposits potential mapping.

Statistical tests of hypotheses for the presence of SAC in spatially distributed datasets such as the secondary mineral occurrence point data is imperative to evaluate its effect on the spatial modelling of a given data location using Complete Spatial Randomness (CSR) tests in Point Pattern Analysis (PPA). The measurement of spatial proximity between point locations or data arrangements and the mean cross product of terms are crucial in determining the high and low areas. The high and low values of the datasets show the extent of SAC in the data set when closely located. The presence of SAC in distributive data is determined by the test of hypothesis for CSR, which rejected the null hypothesis but showed that the data distribution is in clusters, given the value of the variance of mineral points greater than the mean.

The research results emphasise the importance of spatial correlation and spatial characteristics as predictive attributes in modelling. The findings in this chapter also showed that the absence of spatial correlation or interaction among predictive attributes reduces the predictive performance and the generalisation of the PSM-MPM to new and unseen datasets significantly. Although most ML problems tend to deal with natural and unknown data knowledge, the simulation of such natural datasets to reflect the desired composition can be used to validate the result of an expected output to make stronger inferences. Identifying the essential attribute in the PSM-MPM building helps to build a simple, better and more robust model that fits the purpose to which it is intended. The work also clearly shows the importance of SAC and spatial components (attributes) in modelling spatial distribution data. The validation for the importance of spatial interaction or spatial correlation on PSM-MPM performance was assessed using simulated data to show the importance of inclusion of spatial distribution attributes and association (i.e., SAC) in the generalisation and predictive performance of PSM-MPM.

The absence of a more rigorous approach in current or popularly used model evaluation practices has far-reaching consequences. The assessment was achieved through model selection and judging the confidence in the model performance results, i.e., predictive accuracy change. By applying a method of testing that offers a more optimistic evaluation, the case of any test on a partially independent test data will lead to the selection of overfitting. Such models give a false insight as

to which factors are relevant to the distribution model, which can, in turn, lead to model predictions of which conditions are suitable for the PSM-MPM. The results of this experiment suggested that secondary mineral distribution modellers must always determine the extent to which SAC is presented in a distributive spatial dataset and exercise caution when using such data for modelling or predictions, especially under radically changed conditions such as exploring the mineral potential of a given area (i.e., geospatial data).

The approach to model testing and evaluation used for assessing the effectiveness of a model should equal the intended purpose. For a PSM-MPM to be robust and predict into new areas well, it will need to be thoroughly tested based on independent attributes devoid of noise, and the results of the performance evaluated. Considering the extent to which spatial data are associated (distance-wise), allows for modelling using truly independent and spatially segregated data.

The work also shows that it is imperative that the current and most widely used methods of testing or validating a target entity in distribution models namely; re-substitution, RHO tests and cross-validation, lead to estimates of predictive performance that are affected by the presence of SAC. A comparison of the conventional methods of the testing models listed above to a more rigorous test that accounted for the presence of SAC used a novel spatial splitting techniques known as quarterly longitudinal spatial strip split (SSS) was employed. The presence of SAC prevents RHO data from being truly independent (regardless of the method employed in data collection) and creates a false sense of predictive ability in models.

A novel *spatial strip splitting* or SSS sampling validation technique was used to sample training and test data that better handle overfitting of PSM-MPM for the secondary mineral occurrence of the PYGR area. Notwithstanding the importance of SAC in PSM-MPM, it is imperative to evaluate the predictive performance results or accuracy of the PSM-MPM to avoid accepting over-exaggerated classification performance scores or, on the other hand, a poor PSM-MPM performance. Attribute data sub-selection based on data dimension reduction using PCA has equally failed to tackle the effect of SAC and overfitting but further increases the predictive performance score of the classifiers or ML algorithms (i.e., TB and KNN), but worsened the performance of the poor classifier as expressed

by the NB performance. The SSS resulted in a more optimistic performances score for TB and KNN while improving the scores of the NB algorithm. The performances indicated the ability to tackle the detrimental effect of SAC leading to overfitting in both TB and KNN, as well as underfitting for NB classifier.

The direct application of learning unrestricted mineral attribute datasets by selected classifiers (i.e., KNN, TB and NB), suspected of overfitting and underfitting using RHO selection of training and test set was analysed, against the application of restricted attributes that select best attributes subset using PCA. The PCA selection saw an increase in the predictive accuracy of KNN and TB –i.e., from 98% to 99% and 96% to 98% respectively. There was a further decline, however, in the NB from 67% to 59%. It shows that, although there was an increase in the predictive accuracy of the KNN and TB, there was also a decline in the predictive accuracy score of NB algorithm.

Finally, a comparative analysis of predictive performances for the three techniques used for building PSM-MPM, which included: RHO, novel SSS and PCA for attribute best subset selection were conducted before implementing RHO. The results obtained showed that the novel SSS technique is the best approach that mitigates the effect of overfitting and underfitting, as well as maintained high predictive performance among best-performing algorithms (i.e., TB and KNN) and improve the less performing ones too (i.e., NB).

Chapter 5

Conclusions, Summary and Future Work

This chapter summarises the entire findings of this thesis, highlighting the major achievements, limitations, conclusion of the thesis, and suggests some possible new directions for future work in this research area.

5.1 Conclusions

The empirical findings and contributions of this work were detailed in chapter 4. To achieve the set objective, a point based PSM-MPM designed and developed using standard supervised ML classification algorithms was validated, using different proposed model validation techniques that involve the re-substitution, standard RHO, half longitudinal split quarter longitudinal spatial strip splitting (SSS) method and RHO with pre-processing of attribute data using PCA. These model validation techniques deployed that divide the predictive attributes datasets into training and test set for ML classification was considered as part of the primary contribution of this work. From the results obtained in this work, the techniques of the spatial splitting of the predictive attributes into training and test set detects the presence of overfitting better in the distribution datasets.

The presence of SAC caused high similarity of attribute values in space such as the values of elevation or nearest closest distance of several mineral data points to a certain geological factor like rock type in the dataset. The subsequent level

of overfitting and underfitting were determined by the various degree of predictive accuracies presented by the classifiers used; which is KNN, TB and NB. The technique of spatial splitting offers a unique and better procedure that ascertain the severity of overfitting and underfitting in the performance of various ML classifiers, over the traditional RHO validation technique alone. The method was used to extract novel features that deal with SAC in spatially distribution datasets that cause overfitting and underfitting, by enhancing the attribute spatial data independence, for a proper model performance validation. The lack of real spatial independence among the secondary mineral predictive attributes values has been found to be the primary cause of spatial autocorrelation (SAC); a phenomenon established in the secondary mineral distribution data obtained from PYGR. An extension work to the space splitting method of model validation in this work is the determination of the effect of both overfitting and *underfitting* in spatially distributed discrete data set. This effect is evident in the moderation and optimisation of predictive performance for the KNN and TB while the NB experience predictive performance improvement used for building PSM-MPM.

A statistical and spatial data analysis of the attribute data procedure was conducted first, to determine the spatial relationship between mineral occurrence and another geological attribute. The Point Pattern Analysis (PPA) was used to measure the distribution pattern of the mineral occurrence represented as point's distribution, to measure the spatial relationship between the mineral occurrence points and the nearest distance to rocks in the study area. The results of a complete spatial random (CSR) test concluded a non-random distribution of the mineral occurrence point data with a further analysis using quadrat that indicated a more likely clustered points distribution on the map scale than a linear distribution pattern.

The spatial analysis test using Kolmogorov-Smirnov test confirmed the presence of spatial correlation among the predictive attributes extracted from the datasets, such as the relative nearest distances between the point of mineral occurrence and the geological rocks type. The correlation test was done to justify the selection or inclusion of spatial attributes of nearest distances between geological attributes (source) and occurrence points (deposit location), as part of the overall predictive attributes dataset selected to build the PSM-MPM. Although the of

effect SAC inherent in the mineral distribution dataset constitute the challenges faced by the ML classification when modelling spatially distributed data such as the secondary mineral deposits. The adverse effect of SAC was identified by the exaggerated high and sometimes very poor predictive accuracy scores, indicating a possibility of overfitting and underfitting performance respectively, by the used ML classifiers.

The result of overfitting, as well as underfitting, are often determined by the poor predictive performance of the PSM-MPM when validated. The predictive performance assessment was determined by the predictive accuracy scores of the classifiers through performance validation of the training data on the test set by the supervised ML classification. The predictive performance shows a case of overfitting in both KNN and TB algorithms, but underfitting by the NB algorithm or classifier. The approach was conducted using firstly, the use of attribute data for modelling without any preprocessing, while the second method uses the attribute best subset selection that reduces the number of predictive attribute to best fewer sets. The two approaches were referred to as the technique of RHO with the original dataset, in its natural form and the PCA-RHO methods for the purpose of this research. The performance of PSM-MPM using these methods (i.e., RHO and PCA-RHO), were analysed and compared to the spatial split techniques result. The spatial splitting and SSS sampling technique, presents the best plan that handles SAC and limit the detrimental effect overfitting and underfitting in the PSM-MPM. The method attempts to improve heterogeneity of attribute datasets values in the entire data set, by reducing the over dependence of attribute data spatially, while retaining some relative amount of SAC in the dataset such that both the training set contains similar but not exactly same values of attributes, that can be transferable. The results of the experiment show the uniqueness of the procedure of splitting training and validating sets that best determine the presence of overfitting and underfitting in spatially distributed datasets such as the PSM-MPM developed using point based approach to ML classification.

The comprehensive achievements recorded in this work are summarised based on the set objectives as follows:

- A systematic approach of combining GIS, statistics and spatial analysis of

secondary mineral data, that leads to acquisition and mineralisation attribute data extraction, such as the geophysical, geological and geospatial data that involves obtaining the analogue sets of data and therefore requires a specialise skills in GIS, Excel and statistics to acquire and convert the spatial datasets containing the occurrence of captured geological survey due to acute paucity of datasets and converting into a supervised ML classification acceptable format. Other mineralisation attributes extracted systematically and used for modelling include lithology type; spatial components of relative distances between mineral deposit points, other geological features, statistics and GIS to pre-process and analyse the collected data points value in the form suitable for applying ML algorithms. Note that, it is not the data collection that was considered an achievement but the unique technique adopted to acquire it and presents it in a supervised ML classification format.

- Determined the geological features of the PYGR, which include lithological components of fifteen different types of granite rocks, which are tested to be spatially associated with or are indicative of mineralisation. A spatial analysis of mineral occurrence points (i.e., both of mineral data presence and absence) with the associated nearest distances of granite rock features was conducted to determine the distribution pattern of the cassiterite mineral distribution points. Other geological features associated to the mineralisation identified in this work that included some associated attributes of the mineral occurrence points, such as elevation and slope of the coordinate points (latitude and longitude) were also determined. The sizes and perimeters of all the rocks in the PYGR area were also established as part of the mineralisation indicators or as predictive attributes of the PSM-MPM.

- Designed and developed a point data predictive spatial model for mineral potential mapping (PSM-MPM) using seven standard ML classification algorithms. Specifically, the seven supervised ML algorithms used are Naive Bayes (NB), Bagged Decision Tree or Tree-Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbour (KNN) and Logistic Regression (LGR). The performance

of all the seven classifiers was validated using standard random holdout (RHO) sampling technique. A performance evaluation of the classifiers conducted through the comparison of the results of model validation was used. Two best-performing algorithms (i.e., TB and KNN) and one worst performing algorithm (i.e., NB) measured by their predictive accuracy scores and the value of the area under the receiver operating characteristics (AU-ROC), were selected for further evaluations.

- Developed a point data-driven approach to assessing the effect of either the presence or absence of spatial attributes and spatial distribution components (SAC) in mineral distribution datasets. By first simulating the original mineral dataset that excluded the SAC and then, the deliberate isolation of the spatial attributes in the datasets. The results of the performance assessment test show that both spatial attributes and its components are critical predictors required for developing PSM-MPM. The predictive performance results indicated that the absence of the spatial attribute sets in PSM-MPM presented poor predictive performance score, while the complete elimination of SAC using the synthetic data generated through simulation, lead to a far worst predictive performance equal only to predicting from a random guess (i.e., failed to predict).

- Proposed and implemented a new method of PSM-MPM performance evaluation, that involves model performance validation evaluation technique through data sampling of training and test set. The method was used for PSM-MPM performance evaluation due to both *overfitting and underfitting*; a phenomenon often associated with spatial distribution datasets, as presented in the secondary mineral distribution datasets used in this work. The spatial splitting technique used for training and validation in an ML classification is considered to be the ideal approach to PSM-MPM performance evaluation at the expense of the traditional RHO sampling method. The RHO is often used for splitting training and test set randomly, to test the performance of model when predicting mineral potential mapping (Porwal, 2006). The technique of SSS sampling of training and test set as shown

in the developed algorithm 4, detect and limit the effect of both the overfitting and *underfitting* caused by SAC, by ensuring less homogeneity between the attributes training and test sets through the attributes spatial splitting in a spatially distributed point data, such as the secondary mineral distribution datasets.

- Conducted a comparative performance analysis between the novel SSS sampling technique, designed in this thesis with the standard RHO technique and PCA method of attribute subset selection. The results indicated that the SSS method has the better ability than the RHO validation, and PCA attribute data preprocessing that selects the best attribute subset (i.e., PCA-RHO). The ability of the SSS split training and test set in an ML classification to best determine overfitting and underfitting was achieved by the spatial separation of attributes values that enhances true independent but similar attributes that reduces the over dependent of similar attributes values due to SAC in both the TB and KNN classifiers and improves the predictive accuracy of the NB classifier. The method of SSS was considered to be the ideal approach to sampling or splitting training and test data set using ML classification, for PSM-MPM performance based on the comparative analysis of the three techniques mentioned. The PCA-RHO technique still incorporate the SAC components completely that hardly represent true independent training and test sets that are transferable.

5.2 Limitations of the Thesis

Although the work achieved some certain level of success through some of the achievements already highlighted, it was not without some limitation among which include:

- The lack of digital map data of the PYGR area: The manual method of analogue data collection through the physical, geological survey used for obtaining and recording the location of mineral deposits, introduces human and equipment errors into the datasets. The process of reading values from equipment such as the coordinate location of mineral data set from the GPS

and recording them manually before transferring onto the digitised geological map in GIS may introduce errors into the dataset. Whereas, with a fully digitised geological map data values are collected and process with greater precision or minor error recorded.

- Paucity of mineralisation attribute: A limitation due to the absence of some known predictive attribute dataset that is often indicative of mineral occurrence. Specifically, the lack of attributes such as the geochemical components of cassiterite deposits and spatio-temporal components attributes were missing among the predictive attributes data used. Although, the basic elements of cassiterite mineralisation are the mostly the spatial aspects of the process of formation that excludes chemical transformation, the inclusion of such attributes which are common in most mineral composition or occurrences would have aided the robustness of our model to apply to other mineral deposits that require such attribute. Since the more mineralisation attributes used, the more robust and transferability the model.

- Class assumptions: Due to the absence of enough available negative data to represent the non-mineralised location, a general assumption was made to include all other mineralise location apart from the target mineral deposits found in the study area were assumed to be the non-mineralised location. This general assumption could be avoided if there were enough negative instances needed to conduct a supervised ML classification. Only two classes or labels used in PSM-MPM are considered mineralised and non-mineralised. The PSM-MPM is only deemed to be applicable in predicting the strict presence of cassiterite or other minerals and not presence. Therefore, this is a limitation to the PSM-MPM, and a more robust and precise PSM-MPM that will predict the occurrence of mineral deposits will require an exclusive area where cassiterite occurred or not.

5.3 Summary

This work clearly highlights the effectiveness of an ML classification method to model and predict spatial distribution phenomena such as the secondary cassiterite

mineral deposit distribution, presented as points in space (map). The PSM-MPM was developed using for the first time, the secondary mineral cassiterite data obtained from the PYGR can predict areas where mineral deposits are found based on the current mining points data. Some standard supervised ML classification algorithms were deployed to test their performances; they include Naive Bayesian (NB), Tree-Bagging (TB), Decision Tree (DT), Support Vector Machine (SVM), Discriminant Analysis (DA), K-Nearest Neighbours (KNN) and Logistic Regression (LGR). Consideration was given, however, to only three algorithms that include TB, NB and KNN for performance evaluation of the PSM-MPM produced, to determine the effect of SAC to predictive accuracies of the classifiers. Since secondary mineral distribution data represents a spatial distribution datasets, that are often affected by SAC due to the spatial arrangement of the mineral occurrence point's data together (clustered).

While the work of this thesis explicitly shows the importance of SAC and spatial predictive attributes dataset through simulation, the detrimental effect of SAC among predictive attribute dataset leads to overfitting and underfitting performance by the three supervised ML classifiers. The ML predictive model performance evaluations are often determined by the comparison of the performance of all the classifiers, based on the test set also referred to as the validation set. The traditional method of splitting training and test mineral occurrence dataset for model performance evaluation is the random holdout (RHO) or cross-validation technique. The distributive arrangement of the secondary cassiterite mineral occurrence datasets was determined using statistics and GIS. The data arrangement of the spatial dataset is very vital to establish the applicability of the dataset to modelling and prediction. The work examines the various method of predictive performance assessments that determines the detrimental effect of SAC and seek to mitigate such adverse effect leading to overfitting and underfitting.

Predicting spatial distribution must take into consideration the concept of SAC associated with spatial attributes to avoid both exaggerated high accuracy score and poor predictive accuracy by the ML classifier. The results of various ML predictive modelling and performance evaluations carried out in this work clearly indicated that spatial attributes and their autocorrelation (SAC) are major factors that determine the performance of an ML classifier used to build PSM-

MPM. The concept of Tobler's first law is the theory behind the spatial association that defines the predictive modelling of mineral point distribution data, such as the secondary mineral deposits of the PYGR in Nigeria. The work highlights a method of PSM-MPM performance validation evaluation for spatially distributed datasets. The performance assessment procedure considers the spatial separation when splitting training and test datasets, to reduce spatial dependency among the predictive attributes. The concept of space splitting increases the covariation among predictive attribute dataset, which gradually decreases similarity of values among predictive attributes to a more decent predictive performance accuracy (an optimistic precision). Recalled that the presence of SAC in point distribution data is a concept of attribute similarities at certain distance intervals that decay with an increase in distance separation. It is, therefore, logical that the splitting of training and test dataset for model validation be conducted using distance intervals separation such as the new SSS approach to conducting model performance assessment. The SSS has been established as a novel technique for PSM-MPM validation on secondary mineral of the PYGR. The conventional methods of assessing the ML classification performance using supervised ML classifiers that include the KNN, TB and NB algorithms in this work, are the random holdout (RHO) and re-substitution. The work also applied a data preprocessing approach using PCA that selects the best attribute subsets and uses RHO to split the training and test set randomly. All the methods deployed except the SSS technique presented an unreliable predictive performance accuracy scores, and failed to limit the effect of overfitting for KNN and TB as well as underfitting for the NB classifier.

Finally, from the results achieved in this work, it was very clear that despite the high sensitivity of PSM-MPM to SAC, prediction of spatial distribution dataset such as the secondary mineral occurrence distribution datasets is made with standard ML classification using a point data approach of labelling, consisting of both spatial and non-spatial predictive attributes. Care must always be considered, when developing and evaluating PSM-MPM performance, by taking account of the presence or otherwise of SAC. In other words, exploratory data analysis that checks for the presence of SAC in the datasets through distribution pattern test should be conducted to test for overfitting and underfitting before implementing

the PSM-MPM. The evaluation of PSM-MPM must always be subject to performance assessment and appraisal of the spatial distribution component that deals with SAC, a consequence of overfitting and underfitting of ML classifiers must be determined before implementing PSM-MPM.

5.4 Future Work

The approach to modelling mineral deposits conducted in this work considers existing mineral occurrences as points. Future work could compare point-based with area-based (polygonal) or grid approaches to predicting mineral deposits potential and their performances, to determine the best approach to modelling and predicting secondary mineral deposit occurrence. Since part of the limitation of PSM-MPM is the lack of adequate mineralisation predictive datasets that includes historical and geo-mineral deposits data, it is necessary to consider more historical geo-mining data of secondary mineral occurrence data for implementing PSM-MPM. The inclusion of historical mineral occurrence attribute data such as years of formation, type of mining conducted, as well as the quality or quantity of minerals discovered, could add to the robustness of the PSM-MPM and lead to performance assessment due to other forms of autocorrelation such as the spatio-temporal autocorrelation attributes of the mineralisation.

An extension of the SSS sampling validation technique of PSM-MPM implemented in this thesis can be used to determine the optimal or best extent of strips or separation of attributes data (i.e., number of stripping) or alternative directional spatial split, such as latitudinal strip split that may reduce or eliminate the adverse effect of SAC in the spatially distributed datasets, to avoid overfitting or underfitting by supervised ML classifiers. An accurate measure of spatial split (distance wise) that leads to spatial independence of attributes data values in space may enhance true data independence and allow real correlation to be present in both training and test sets sampling for validation. Such an experiment when conducted successfully, can results in establishing a better method of validating spatial predictive model performance for model generalisation in supervised ML classification and help to reduce if not eliminate the over-dependence of attributes

data values in space. The aim of this work did not include eliminating overfitting and underfitting but limiting their effect by addressing SAC.

Finally, regarding applications, many possible directions can be explored from various spatially distributed datasets, for modelling and predictions within the context of ML classification, using point distribution datasets. The implementation of spatial strip split (SSS) method of model validation deployed in this thesis, would help in the area of geo-mining and mineral prospecting, an implementation PSM-MPM that optimises performance as well as mitigates the effect of SAC developed in this thesis can be considered by ML software engineers, in an embedded or working system for the mining industries and other local miners. A working mineral potential predictive system when developed, can be used for the discovery of potential mineral deposit locations, especially in the PYGR area of Nigeria and other places, which is the primary focus of this research. The model would also help the Nigerian authority with a better land policy in the PYGR area.

References

Agarwal, P. (2007). Walter Christaller: Hierarchical Patterns of Urbanization. *Centre of Spatially Integrated Social Science*. 36

Agterberg, F., Bonham-Carter, G.F., Wright, D. *et al.* (1990). Statistical pattern integration for mineral exploration. *Computer applications in resource estimation prediction and assessment for metals and petroleum*, 1–21. 17, 20, 26

Araujo, M.B. & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688. 30

Bahn, V. & McGill, B.J. (2013). Testing the predictive performance of distribution models. *Oikos*, **122**, 321–331. xvi, 9, 24, 26, 28, 30, 80, 86, 87, 88, 92, 131, 138, 143, 144

Bateman, M.A. (1951). The formation of mineral deposits. *Wiley*. 2

Bennett, B. (1996). The application of qualitative spatial reasoning to GIS. In R. Abrahart, ed., *Proc First Int. Conf. on GeoComputation*, vol. I, 44–47, Leeds. 17

Berman, M. (1977). Distance distributions associated with poisson processes of geometric figures. *Journal of Applied Probability*, 195–199. 72

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. *Oxford: Clarendon Press*. 2, 20

REFERENCES

Bishop, C.M. (2006). *Pattern recognition and machine learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 38

Bonham-Carter, G. (1985). Statistical association of gold occurrences with LANDSAT-derived lineaments, Timmins-Kirkland Lake Area, Ontario. *Canadian Journal of Remote Sensing*, **11**, 195. 72

Bonham-Carter, G. (1994). *Geographic Information Systems for geoscientists: modelling with GIS*, vol. 13. Pergamon press. 2, 4, 16, 17, 20, 22, 24, 26, 45, 48

Bonham-Carter, G. & Agterberg, F. (1990). Application of a microcomputer-based Geographic Information System to mineral potential mapping. *Microcomputers in Geology*, **2**, 49–74. 20, 21

Boots, B.N. & Getis, A. (1988). *Point pattern analysis*, vol. 10. SAGE publications Newbury Park, CA. 1, 16, 19, 22, 34, 35, 36, 69, 70

Bowden, P. & Jones, J. (1978). Mineralization in the younger granite province of Northern Nigeria. *Metallization Associated with Acid Magmatism*, **3**, 179–190. 18

Bowden, P. & Kinnaird, J.A. (1978). *Younger Granites of Nigeria: A Zinc-rich Tin Province*. 5

Bowden, P., Bennett, J., Kinnaird, J.A., Whitley, J., Abaa, S. & Hadzigeorgiou-Stavrakis, P.K. (1981). Uranium in the Niger-Nigeria younger granite province. *Mineralogical Magazine*, **44**, 379–389. 16

Bradley, A.P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, **30**, 1145–1159. 27, 41, 78

Bratko, I. (2001). *Prolog (3rd Ed.): Programming for Artificial Intelligence*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 38

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32. 37, 40

Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A. (1984). *Classification and Regression trees*. CRC press. 27, 40

Campbell, A., Hollister, V., Duda, R. & Hart, P. (1982). Recognition of a hidden mineral deposit by an artificial intelligence program. *Science*, **217**, 927–929. 20

Carranza, E., Van Ruitenbeek, F., Hecker, C., Van der Meijde, M. & Van der Meer, F. (2008). Knowledge-guided data-driven evidential belief modeling of mineral prospectivity in Cabo de Gata, SE Spain. *International Journal of Applied Earth Observation and Geoinformation*, **10**, 374–387. 21

Carranza, E.J.M. & Hale, M. (2000). Geologically constrained probabilistic mapping of gold potential, Baguio district, Philippines. *Natural Resources Research*, **9**, 237–253. 21

Carranza, E.J.M. & Hale, M. (2001). Geologically constrained fuzzy mapping of gold mineralization potential, Baguio district, Philippines. *Natural Resources Research*, **10**, 125–136. 21

Carranza, E.J.M. & Hale, M. (2002). Spatial association of mineral occurrences and curvilinear geological features. *Mathematical Geology*, **34**, 203–221. 72, 107

Carranza, E.J.M. & Hale, M. (2003). Evidential belief functions for data-driven geologically constrained mapping of gold potential, Baguio district, Philippines. *Ore Geology Reviews*, **22**, 117–132. 16, 21

Carranza, E.J.M., Mangaoang, J.C. & Hale, M. (1999). Application of mineral exploration models and GIS to generate mineral potential maps as input for optimum land-use planning in the Philippines. *Natural Resources Research*, **8**, 165–173. 21, 44

Carranza, E.J.M., Owusu, E.A. & Hale, M. (2009). Mapping of prospectivity and estimation of number of undiscovered prospects for lode gold, southwestern Ashanti Belt, Ghana. *Mineralium Deposita*, **44**, 915–938. 77

Cheng, J. & Greiner, R. (1999). Comparing Bayesian Network Classifiers. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 101–108, Morgan Kaufmann Publishers Inc. 39

Connor, E.F. & Simberloff, D. (1979). The assembly of species communities: chance or competition? *Ecology*, 1132–1140. 36

Cooper, G.F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, **9**, 309–347. 39

Dacey, M.F. (1964). Modified poisson probability law for point pattern more regular than random 1. *Annals of the Association of American Geographers*, **54**, 559–565. 36

Dacey, M.F. & Tung, T.h. (1962). The identification of randomness in point patterns. *Journal of Regional Science*, **4**, 83–96. 36

Danso, S.O. (2006). An exploration of classification prediction techniques in data mining: The insurance domain. *Master Degree Thesis, Bournemouth University*. 41

Dark, S.J. (2004). The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. *Diversity and Distributions*, **10**, 1–9. 4

Diebold, F.X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, **33**, 1–1. 28

Diggle, P.J. (1983). Statistical analysis of spatial point patterns. 1, 19, 35, 36, 70

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, **12**, 53–64. 32

Dormann, C.F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138. 33

Duda, R.O., Hart, P.E. & Stork, D.G. (1995). Pattern classification and scene analysis 2nd ed. 19, 20, 39

Duda, R.O., Hart, P.E. & Stork, D.G. (2012). *Pattern classification*. John Wiley & Sons. 41

Ekosse, G.I. & Mwitondi, K.S. (2015). Principal component analysis to evaluate the spatial variation of major elements in kaolin deposit. *Bulletin of the Chemical Society of Ethiopia*, **29**, 41–51. 21, 22

Falconer, J.D. (1912). Nigerian tin; its occurrence and origin. *Economic Geology*, **7**, 542–546. 17, 18

Fielding, A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, **24**, 38–49. 87

Fortin, M.J. & Dale, M.R.T. (2005). *Spatial analysis: a guide for ecologists*. Cambridge University Press. 24, 32, 33, 34

Fowler, J., Cohen, L. & Jarvis, P. (1990). Practical Statistics for Field Biology John Wiley and Sons. 70

Geoscience Australia, C. (2007). Australia's identified mineral resources. Tech. rep. 18

Geoscience Australia, C. (2013). Australia's identified mineral resources. Tech. rep. 18

Goodchild, M.F. (1987). A spatial analytical perspective on geographical information systems. *International journal of geographical information system*, **1**, 327–334. 4

Greig-Smith, P. (1979). Pattern in vegetation. *The Journal of Ecology*, 755–779. 36

Greig-Smith, P. (1983). *Quantitative plant ecology*, vol. 9. Univ of California Press. 70

Griffith, D.A. (2013). *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer Science & Business Media. 4, 31

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., OVERTON, J., Aspinall, R., Hastie, T. *et al.* (2006). Making better biogeographical predictions of species distributions. *Journal of Applied Ecology*, **43**, 386–392. 4, 24

Haldar, S. (2013). Chapter 1 - mineral exploration. In S. Haldar, ed., *Mineral Exploration*, 1 – 21, Elsevier, Boston. xv, 19

Hampe, A. (2004). Bioclimate envelope models: what they detect and what they hide. *Global Ecology and Biogeography*, **13**, 469–471. 30

Han, J., Kamber, M. & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier. 27, 41, 78

Hepinstall, J.A. & Sader, S.A. (1997). Using Bayesian statistics, Thematic Mapper satellite imagery, and breeding bird survey data to model bird species probability of occurrence in Maine. *Photogrammetric Engineering and Remote Sensing*, **63**, 1231–1236. 39

Hines, C., Adams, G., Brosnahan, J., Djuth, F., Sulzer, M., Tepley, C. & Van Baelen, J. (1993). Multi-instrument observations of mesospheric motions over arecibo: comparisons and interpretations. *Journal of atmospheric and terrestrial physics*, **55**, 241–287. 36

Hunt, E.B., Marin, J. & Stone, P.J. (1966). Experiments in induction. 40

Ibrahim, A.M. & Bennett, B. (2014a). The assessment of machine learning model performance for predicting alluvial deposits distribution. vol. 36, 637 –

642, complex Adaptive Systems Philadelphia, {PA} November 3-5, 2014. 9, 15, 26, 28, 37, 40, 70, 77, 85, 86, 87, 88, 97, 131, 140, 143

Ibrahim, A.M. & Bennett, B. (2014b). Point-based model for predicting mineral deposit using gis and machine learning. In *Proceedings of the 2014 First International Conference on Systems Informatics, Modelling and Simulation*, SIMS '14, 83–88, IEEE Computer Society, Washington, DC, USA. 17, 18, 24, 27, 33, 37, 45, 55, 78, 86, 97, 140

Ibrahim, A.M., Bennett, B. & Campelo, C.E. (2015a). *Predictive expert models for mineral potential mapping*, 3161–3168. IGI Global, Hershey, PA, USA, iD: 112744. 15, 38

Ibrahim, A.M., Bennett, B. & Isiaka, F. (2015b). The optimisation of bayesian classifier in predictive spatial modelling for secondary mineral deposits. *Procedia Computer Science*, **61**, 478–485. 97

Isaaks, E.H. & Srivastava, R.M. (1989). An introduction to applied geostatistics. 34

Kemmerly, P.R. (1982). Spatial analysis of a karst depression population: clues to genesis. *Geological Society of America Bulletin*, **93**, 1078–1086. 36

Kissling, W.D. & Carl, G. (2008). Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, **17**, 59–71. 4, 24, 33

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, 1137–1143, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 42, 78

Kohavi, R. & John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273–324. 83

Langley, P. (1996). *Elements of machine learning*. Morgan Kaufmann. 38

Langley, P., Iba, W. & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the National Conference on Artificial Intelligence*, 223–223, JOHN WILEY & SONS LTD. 39, 83

Lary, D.J. (2010). *Artificial Intelligence in geoscience and remote sensing*. INTECH Open Access Publisher. 14, 80

Lary, D.J., Alavi, A.H., Gandomi, A.H. & Walker, A.L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, **7**, 3 – 10, special Issue: Progress of Machine Learning in Geosciences. 14

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673. 4, 32, 33

Legendre, P., Dale, M.R., Fortin, M.J., Gurevitch, J., Hohn, M. & Myers, D. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**, 601–615. 33

Legendre, P.L. & Legendre, L. (1998). L. 1998. numerical ecology. *Second English Edition. Amsterdam Elsevier Science*. 4, 32, 34

Lennon, J.J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, **23**, 101–113. 33

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, **72**, 445–463. 33

Liebhold, A. & Gurevitch, J. (2002). Integrating the statistical analysis of spatial data in ecology. *Ecography*, **25**, 553–557. 4, 24

MathWorks Documetation, H. (2015). `http://uk.mathworks.com/help/stats/examples/simulating-dependent-random-variables-using-copulas.html`. 81

McConnell, H. & Horn, J. (1972). Probabilities of surface karst. *Spatial analysis in geomorphology. Harper & Row, New York*, 111–133. 36

Miller, H.J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, **94**, 284–289. 33, 72

Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1st edn. 15, 38

Mwitondi, K., Moustafa, R. & Hadi, A. (2013). A data-driven method for selecting optimal models based on graphical visualisation of differences in sequentially fitted roc model parameters. *Data Science Journal*, **12**, WDS247–WDS253. 23, 43, 79

Mwitondi, K.S. & Said, R.A. (2013). A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. *Journal of Statistics Applications and Probability*. 44, 79

Neville, J., Adler, M. & Jensen, D. (2003). Clustering relational data using attribute and link information. In *Proceedings of the text mining and link analysis workshop, 18th international joint conference on Artificial Intelligence*, 9–15. 24

Pardos, Z.A. & Heffernan, N.T. (2010). Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. *Journal of Machine Learning Research W & CP*. 40

Pastor, J. & Turaki, U. (1985). Primary mineralization in nigerian ring complexes and its economic significance. *Journal of African Earth Sciences (1983)*, **3**, 223 – 227, alkaline ring complexes in Africa. 5, 55

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. 39

Perry, J., Liebhold, A., Rosenberg, M., Dungan, J., Miriti, M., Jakomulska, A. & Citron-Pousty, S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography*, **25**, 578–600. 34

Porwal, A.K. (2006). Mineral potential mapping with mathematical geological models. *ITC PhD Dissertations*, **130**. 4, 17, 21, 27, 28, 37, 73, 89, 144, 153

Quigley, J.M. (1998). Urban diversity and economic growth. *The Journal of Economic Perspectives*, 127–138. 36

Rahbek, C., Gotelli, N.J., Colwell, R.K., Entsminger, G.L., Rangel, T.F.L. & Graves, G.R. (2007). Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 165–174. 4, 24

Rigol-Sanchez, J., Chica-Olmo, M. & Abarca-Hernandez, F. (2003). Artificial Neural Networks as a tool for mineral potential mapping with GIS. *International Journal of Remote Sensing*, **24**, 1151–1156. 20, 21

Ripley, B.D. (1991). *Statistical inference for spatial processes*. Cambridge University Press. 37

Rogers, C.S. & Brown, J.J. (1974). Shopping center financing. *UMKC L. Rev.*, **43**, 1. 36

Russell, S. & Norvig, P. (2005). AI a modern approach. *Learning*, **2**, 4. 15

Sawada, M. (2001). Global Spatial Autocorrelation indices–Moran's I, Geary's C and the General Cross-Product Statistic. *Laboratory of Paleoclimatology and Climatology, Dept. Geography, University of Ottawa,(Mimeo)*. 34

Scheres, S.H. & Chen, S. (2012). Prevention of overfitting in cryo-em structure determination. *Nature methods*, **9**, 853–854. 28

Segurado, P., Araújo, M.B. & Kunin, W. (2006). Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444. 30

Shekhar, S., Lu, C., Tan, X., Chawla, S. & Vatsavai, R. (2001). A visualization tool for spatial data warehouses. *Geographic data mining and knowledge discovery*, 73. 16, 72

Simberloff, D. & Connor, E.F. (1981). Missing species combinations. *American naturalist*, 215–239. 36

Sokal, R.R. & Oden, N.L. (1978). Spatial autocorrelation in biology: 1. methodology. *Biological Journal of the Linnean Society*, **10**, 199–228. 24, 33

Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A. & Džeroski, S. (2010). Estimating vegetation height and canopy cover from remotely sensed data with machine learning. *Ecological Informatics*, **5**, 256–266. 25, 29

Stojanova, D., Ceci, M., Appice, A., Malerba, D. & Džeroski, S. (2011). Global and local spatial autocorrelation in predictive clustering trees. In *International Conference on Discovery Science*, 307–322, Springer. 24, 25

Tobler, W.R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 234–240. 27, 33

Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, **74**, 519–530. 72

Townsend Peterson, A., Papeş, M. & Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography*, **30**, 550–560. 30

Walker, A.R., Pham, B. & Moody, M. (2005). Spatial bayesian learning algorithms for geographic information retrieval. In *Proceedings of the 13th annual ACM International Workshop on GIS*, 105–114, ACM. 2

Wang, W. & Cheng, Q. (2008). Mapping mineral potential by combining multi-scale and multi-source geo-information. In *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, vol. 2, II–1321, IEEE. 39

Wilsher, W., Herbert, R., Wullschleger, N. & Naicker, I. (1993). Towards intelligent spatial computing for the earth sciences in South Africa. *South African Journal of Science*, **89**, 315–315. 36

ZHOU, J. & CIVCO, D.L. (1996). Using genetic learning neural networks for spatial decision making in GIS. *Photogrammetric Engineering and Remote Sensing*, **62**, 1287–1295. 20