Applied Clay Science 46 (2009) 1–6

Contents lists available at ScienceDirect

Applied Clay Science

journal homepage: www.elsevier.com/locate/clay

Multiple data clustering algorithms applied in search of patterns of clay minerals in soils close to an abandoned manganese oxide mine

G-I.E. Ekosse [a,*], K.S. Mwitondi [b]

[a] Directorate of Research Development, Walter Sisulu University, Private Bag XI Mthatha, Eastern Cape 5117, South Africa
[b] Computing and Communication Research Group, Faculty of Arts, Computing, Engineering and Sciences, Sheffield Hallam University, Sheffield S1 1WB, UK

ARTICLE INFO

Article history:
Received 19 December 2008
Received in revised form 7 June 2009
Accepted 9 June 2009
Available online 26 June 2009

Keywords:
Data mining
Kaolinite
Model reliability
Multiple clustering algorithms
Muscovite
Over-fitting

ABSTRACT

This paper proposes a multi-level approach to data clustering and provides a novel approach to characterisation of clay soils by, effectively, looking at the same clay sample from different angles. It is shown that using this approach can help avoid detection of spurious clusters or skipping vital natural grouping in data. Muscovite, illite and kaolinite were identified by X-ray diffraction (XRD) in <4 μm fraction of soil samples obtained from the periphery of an abandoned manganese oxide mine and semi quantified as major, minor and trace. Based on information inherent in the data attributes, useful rules for grouping the samples were generated and with the aid of multiple data clustering, applied to characterize the clay minerals occurrences in the soils. The paper found that the presence of large quantities of illite and kaolinite heavily influence the formation of clusters. When the most influential variables—LJ and KJ were taken out, the resulting model showed that muscovite traces play a vital role in initial cluster building and the importance matrix of inputs suggested inter-dependence between muscovite, kaolinite and illite traces as well as between them and minor quantities of illite. Dwelling on aspects of clay mineralogy and modelling sciences, the paper marks a significant departure from the conventional approaches to clay characterisation by showing how effectively data mining methods can be adopted in the area. For a successful approach to characterisation of clay minerals in African soils, the paper recommends to set-up data repositories that will provide scientific data sources and forums in a multi-disciplinary environment. This is particularly important as capturing interesting patterns requires expert knowledge describing the emerging natural groupings.

1. Introduction

Soil mineral components and how they impact on agriculture, environment, health and other aspects of life have been widely studied using a variety of statistical methods. Bianchini et al. (2002) investigated distributions of two groups of soil samples in order to establish whether or not high heavy-metal concentrations were related to urban–industrial–agricultural activities or original lithologies. Statistical studies of over 2000 samples of soil clay fraction from Ivory Coast, Nigeria, Tanzania, and Uganda (Asadu et al., 1997) revealed significant correlations between all the data attributes and the effective cation exchange capacity (ECEC). Using conventional statistical methods, Tombale (1986) interpreted sediments of the Jwaneng area, and Ekosse and Fouche (2006a) studied abundance and distribution patterns of haematite and goethite in soils close to an abandoned manganese oxide (MnOx) mine in Kgwakgwe, both in Botswana. Other conventional statistical approaches include applications of principal components analysis (Jongman et al., 1995), and impact of noisy environmental data on canonical correspondence (McCune, 1997).

One of the main problems tackled in data mining is data clustering—a technique used in detecting naturally arising groups in data when data class labels are not known. Data clustering, which describes a set of mainly distribution-free algorithms, belongs to the data mining family of algorithms for extracting knowledge from data (Kogan, 2007; Valiant, 1984). Kogan (2007) provides mathematically rigorous approaches to data clustering with main clustering issues such as function smoothing fully addressed. A statistical version of data mining is provided by Hastie et al. (2001) and Taylor and Mwitondi (2001). In the last few years there has been growing interest in the applications of data mining algorithms in clay and soil analyses. Bianchini et al. (2002) investigated the mineralogical composition of the fine fraction of soil using X-ray powder diffraction (XRPD); and identified groups in data on the basis of their heterogeneity—which is the main idea of data clustering. Brown et al. (2001) applied neural networks in gold prospecting. Eriksson et al. (2001) used data mining procedures in monitoring early fault detection and classification as well as in identifying relationships between chemical compositions and biological properties. Taylor and Mwitondi (2001) adapted the expectation–maximisation (EM) algorithm in detecting structures in ecological data. Henderson et al. (2005) applied decision trees in

* Corresponding author.
E-mail address: gekosse@wsu.ac.za (G.-I.E. Ekosse).

0169-1317/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.clay.2009.06.018

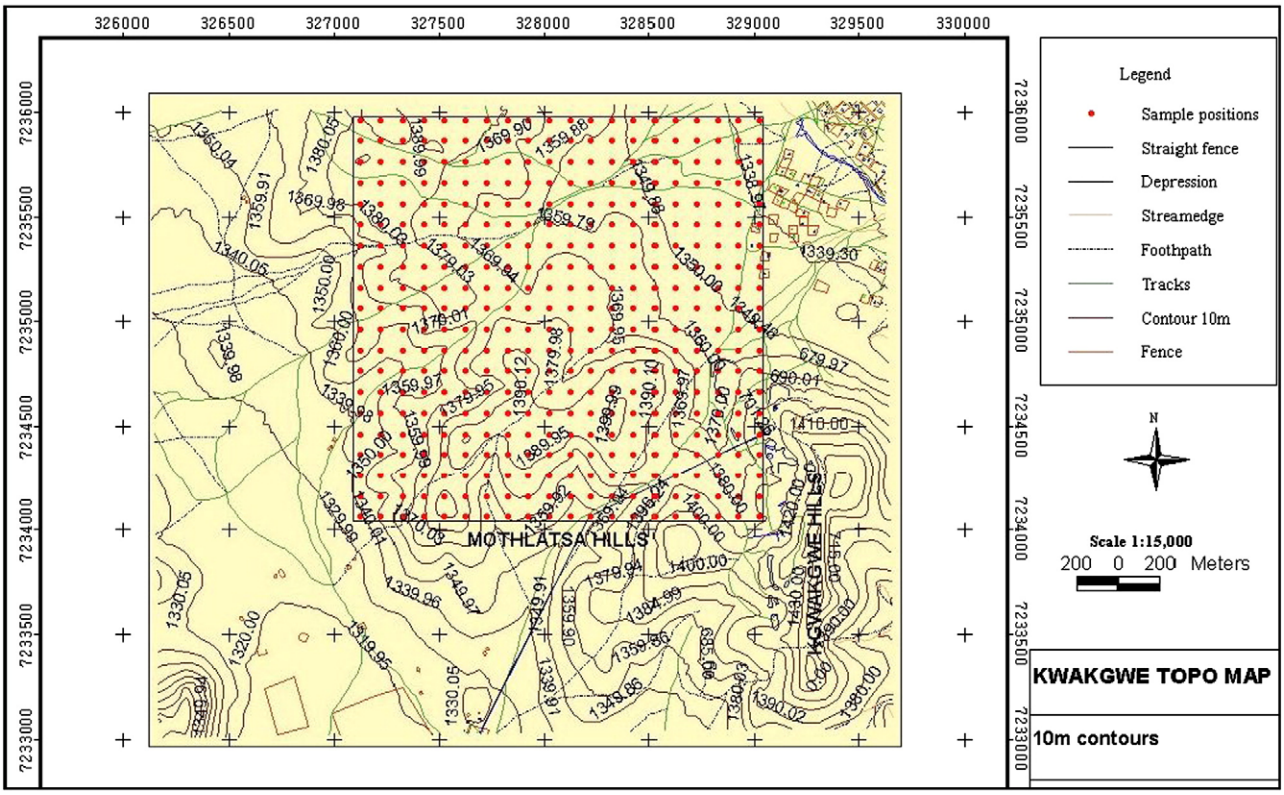

Fig. 1. Map of the study area showing sampling positions.

constructing soil property predictions by accessing the Australian national soils point database. Ekosse and Fouche (2007a,b) applied association rules to elicit inherent relationships among Mn minerals in soils.

Previous mineralogical study of 400 soil samples from the Kgwakgwe area revealed weak correlation coefficients except between kaolinite and illite which was moderate (Ekosse and Fouche, 2006b). A single clustering model was fitted and generated four clusters but the study fell short of addressing core issues such as the way the cluster formation was influenced by the distributions of the data and consequently how the presence/absence of any one of the clay minerals influenced the presence/absence of the other. Apparently, applying a single clustering model to 400 samples from a confined geographical area impaired the validity of its conclusions. The validity of results from data clustering models exclusively depend on their ability to capture interesting groupings in data and on the corresponding expert knowledge in providing useful descriptions of those groupings. The mechanics of typical data clustering algorithm stipulate that the initial number of clusters be known prior to its initialisation which is paradoxical because it is supposed to determine the number of distinguishable clusters. Hence, one of the most frequently asked questions relates to how many data clusters to start with for which, unfortunately, there are no theoretical rules. Typically, cluster searching is guided by trial and error, expert knowledge and exploratory data analysis (EDA) results and quite often data clustering proceeds by combining the three. This paper seeks to address some of these issues.

The paper takes a novel approach to soil clay minerals characterization using multiple data clustering algorithms. It proposes an optimal approach to combining data resources, tools/techniques and technical skills for the purposes of capturing and describing natural clusters in clay minerals input data. The algorithms, based on the philosophy of K-Means algorithm (McQueen, 1967), are used to search for naturally arising structures in clay minerals data obtained from soil samples within the periphery of an abandoned manganese oxide mine. In particular, it advances a combination of data clustering techniques as suitable interpretive methods in elucidating on the mutual influence of the clay minerals on each other—which is central towards the characterization of Mn contaminated soils at Kgwakgwe.

2. Materials and methods

2.1. Background

This paper forms part of a wide study in understanding environmental mining impact on soils close to an abandoned MnOx mine which has been extensively studied by Ekosse (2008) and Ekosse and Fouche (2006a,b), Ekosse and Fouche, 2007a, 2007b) and kaolin deposit (Ekosse, 2001). Both mine and deposit are located in the Kgwakgwe area, 4 km south of the Kanye township in southeastern Botswana, between latitudes 24°59' and 25°02', and longitudes 25°17' and 25°20' covering approximately 4 km^2 in a 2 km by 2 km space as shown in Fig. 1. Four hundred soil samples were collected from the area in Fig. 1 using both random techniques highlighted in Jewell et al. (1993) and judgmental techniques described in Crépin and Johnson

Table 1
Ten out of the 400 samples in a coded tabular illustration.

Sample no.	Muscovite				Illite				Kaolinite			
	None	+	++	+++	None	+	++	+++	None	+	++	+++
1	0					4				7		
2	0				0				0			
3		1				4				7		
4	0					4			0			
5		1				4			0			
6	0							6	0			
7		1			0							9
8	0				0						8	
9		1			0							9
10	0				0					7		

(Note: +, ++ and +++ correspond to trace, minor and major presence respectively of muscovite, illite and kaolinite).