Geoderma 337 (2019) 11–21

Contents lists available at ScienceDirect

Geoderma

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

Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils

Jannis Heil[a,*], Volker Häring[b], Bernd Marschner[b], Britta Stumpe[a]

[a] *Department of General Geography/Human-Environment-Research, Institute of Geography, University of Wuppertal, 42119 Wuppertal, Germany*
[b] *Department of Soil Science/Soil Ecology, Institute of Geography, Ruhr-University Bochum, 44780 Bochum, Germany*

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Soil classification
Diffuse reflectance spectroscopy
Mid-infrared spectroscopy
Clustering
Fuzzy k-means
K-means

ABSTRACT

The amount of data in soil science increased at exponential rates over the last decades, promoted by rapid technological innovation. This development led to a better understanding of processes but also required the introduction of data mining into soil science. With diffuse reflectance Fourier transform (DRIFT) spectroscopy, one of those new methods, soil scientist could build up large spectral libraries. These libraries can expand over large, heterogeneous areas requiring classification algorithms to find subsets or patterns in the data prior to further analysis. The k-means algorithm has become one of the most frequently used algorithms for this task. However, fuzzy k-means (FKM) clustering, a fuzzy variation of k-means, is potentially better suited for spectral data. Fuzzy logic allows for class overlaps and is supposed to reflect the complex nature of soil spectra and continuous environmental variables. In this study, we collected over 1000 mid-infrared DRIFT spectra of agricultural soils from the West African savannah zone and clustered the data using k-means and FKM. Our aim was to explore the feasibility of centroid-based cluster algorithms in finding substructures in spectral data and to discuss the benefits of fuzzy clustering. We found a two-group pattern separating the data set in a northern and southern part. The clustering could primarily be explained by geology and climatic gradients. While both algorithms performed similarly well in picking up the structure, FKM could reveal a transition zone between the two clusters that was not detectable with k-means. This transition zone was explained by a gradual change in aeolian dust deposition, topography, and a change in geology. With this study, we showed the benefits of fuzzy clustering over traditional hard clustering for finding substructure in unexplored spectral data. We recommend the use of continuous classes, as they incorporate more information that could potentially improve subsequent analysis.

1. Introduction

In recent years, the sheer amount of soil, agricultural, and environmental data has increased at exponential rates through the introduction of new measurement equipment for proximal and remote sensing, and the prevalent use of automated data collection systems (Cebeci and Yildiz, 2015; Raj et al., 2018). These new techniques provide data for a better understanding of environmental processes by allowing on the one hand a high temporal resolution and/or a long term monitoring and on the other hand a high spatial resolution over large areas. These progresses in soil science over the last 20–30 years have jointly evolved into a new discipline labelled digital soil mapping (DSM) (McBratney et al., 2003).

One of these new methods facilitating DSM efforts is diffuse reflectance Fourier transform (DRIFT) spectroscopy both in the visible-near infrared (vis-NIR) and the mid-infrared (MIR) range (McDowell et al., 2012). DRIFT spectroscopy shows a high potential for identifying differences and changes in soil properties over time and space, as these methods are more rapid, cost-effective, and require minimal sample preparation compared to traditional laboratory methods (e.g., Bellon-Maurel and McBratney, 2011; McCarty et al., 2002; Reeves, 2010; Soriano-Disla et al., 2014). Spectroscopy allows for a high sample throughput and thus a high spatial resolution of samples as used in DSM (Viscarra Rossel et al., 2009). Soil spectra contain principally qualitative information about soil properties and mineralogy but have also been used to estimate soil properties quantitatively, as reviewed by Soriano-Disla et al. (2014). The complex nature of spectra with overlapping absorption bands between spectral signatures of different soils require for multivariate statistical and chemometric methods in order to interpret the spectra (Viscarra Rossel and Behrens, 2010).

[*] Corresponding author.
E-mail address: jheil@uni-wuppertal.de (J. Heil).

https://doi.org/10.1016/j.geoderma.2018.09.004
Received 29 January 2018; Received in revised form 8 August 2018; Accepted 4 September 2018

Due to its ability to obtain large numbers of samples at low cost, spectral libraries can expand over large areas incorporating soils with highly varying characteristics making an evaluation even more difficult. In these cases, classification algorithms are useful to find subsets or patterns of similar samples based on their spectral information prior to the main analysis, e.g., to improve the predictions of soil parameters on subset models (Araújo et al., 2014; McDowell et al., 2012). These different largely unsupervised classification techniques are commonly combined under the umbrella term cluster analysis (Cebeci and Yildiz, 2015). The main concept behind all cluster algorithms is to assign similar objects into a cluster based on given features that are more similar to each other than to objects in different clusters (Bezdek et al., 1984). While hundreds of algorithms based on different concepts exist and the choice of an appropriate algorithm depends on criteria such as data size, structure, and aim of the study, many studies report a good overall performance for partitioning algorithms belonging to the k-means family (Bora and Gupta, 2014; Cebeci and Yildiz, 2015; Steffens et al., 2014). Since its introduction by MacQueen (1967), k-means has become one of the most popular algorithms in exploratory data analysis. However, as a hard clustering algorithm, k-means is not suited to find overlapping classes. To take continuous data and gradual boundaries in the environment into account, fuzzy logic had been introduced into DSM in the 1990s (Burrough et al., 1997; De Gruijter and McBratney, 1988; McBratney and Odeh, 1997; Odeh et al., 1990, 1992). In clustering, this problem was handled by the introduction of fuzzy k-means (FKM, also known as fuzzy c-means) clustering by Bezdek (1981). FKM is a direct generalization of k-means hard clustering (McBratney and de Gruijter, 1992). In fuzzy, soft clustering, every object belongs to every cluster to a certain degree of membership. This makes FKM potentially interesting for the clustering of spectral data, due to the continuous and complex nature of the spectra (Viscarra Rossel et al., 2016).

While cluster algorithms are widely applied to spectral data in soil studies, whether to improve prediction models (e.g., Araújo et al., 2014) or to determine soil classes for DSM (e.g., Žížala et al., 2017), we do not know of any study that compares hard and fuzzy clustering approaches in a case study. Therefore, we attempted to find substructures in a large data set of West African agricultural soils based on over 1000 collected MIR spectra. For classification, we used two different centroid-based cluster algorithms out of the broad spectrum of methods: the very common and more conventional hard clustering algorithm k-means and FKM as a clustering approach based on fuzzy logic. As this is an exploratory data analysis, we tried to find substructure in the data first and afterwards tried to explain underlying structures based on local climate, geography, and available soil data. We compared the results of both algorithms and discussed the advantages of fuzzy classification compared to conventional hard classes. The aims of this study were: (i) to explore the feasibility of centroid-based cluster algorithms in finding substructures in an unexplored spectral data set, (ii) to explain the clustering internally by spectral interpretations and with external data using environmental and soil variables, and (iii) to discuss the benefit of fuzzy compared to hard classes.

2. Material and methods

2.1. Soil sampling and analysis

The study area expands over the savannah zone of western Africa incorporating parts of northern Ghana and Burkina Faso. Soil samples were taken during four different sampling campaigns between 2013 and 2015. The sampling sites encompass urban agricultural sites in the cities of Tamale, Ghana and Ouagadougou, Burkina Faso and the surrounding peri-urban and rural areas, as well as covering wide parts of rural areas in the Northern, Upper West, and Upper East Regions of Ghana (Fig. 1). Altogether, there were 1084 soil samples summarizing the four sampling campaigns. Further details about the different campaigns can be found in Table 1.

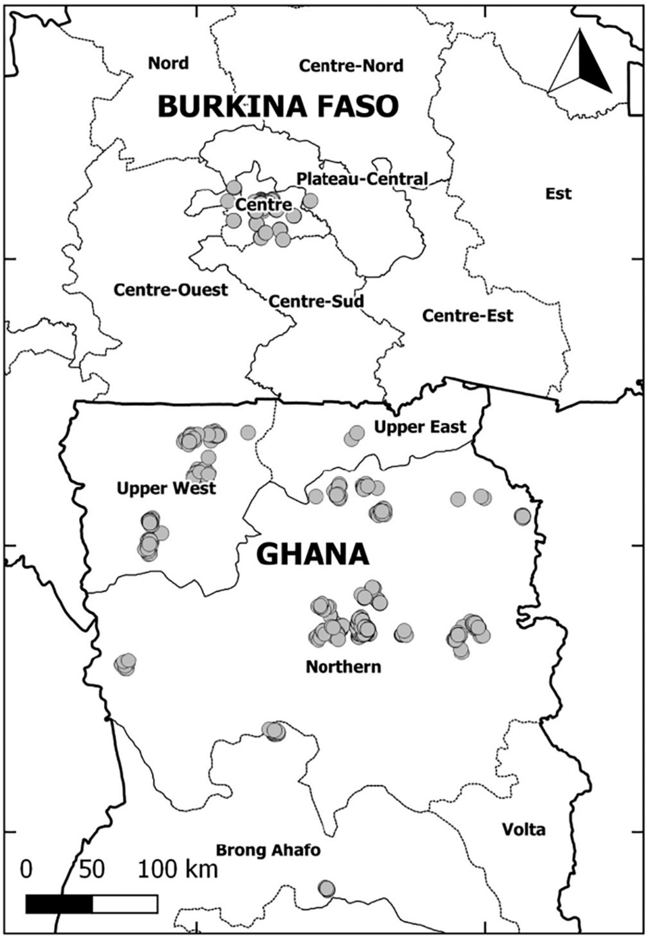

Fig. 1. Overview over the study area in Burkina Faso and Ghana with the location of the sampling locations.

Table 1
Overview over the different sampling campaigns.

Study	Year	Location	n	Characteristic
A	2013	Tamale and Ouagadougou	419	Urban, peri-urban, and rural sites
B	2014	Tamale and Ouagadougou	294	Peri-urban sites only
C	2014	Northern Ghana	294	Rural maize fields only
D	2015	Northern Ghana	77	Rural maize fields only

Soil samples were taken at 0–20 cm depth in three replicates per plot. All replicates were air-dried after sampling, pooled, gently disaggregated, and sieved to < 2 mm for further analysis (Bellwood-Howard et al., 2015). Afterwards, samples were finely ground (< 200 μm) for 2 min in an agate mill (Fritsch GmbH, Idar-Oberstein, Germany) for spectral measurements.

For selected reference samples, soil properties were determined in the laboratory. A total of 163 samples (40 from Burkina Faso and 123 from Ghana) were analysed by wet chemistry according to methods described in Häring et al. (2017) and Bellwood-Howard et al. (2015). Soil properties determined were soil texture, soil pH, soil organic carbon, Fe oxides, and cation exchange capacity (CEC).

All sample sites were georeferenced and loaded into a geographic information system (GIS) for spatial analysis of the obtained data. For further interpretation, additional geographic and geologic information was loaded into the GIS. We obtained climatic information for the study