Geoderma 337 (2019) 413–424

Contents lists available at ScienceDirect

Geoderma

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

Pre-treatment of soil X-ray powder diffraction data for cluster analysis

Benjamin M. Butler[a,*], Andrew M. Sila[b], Keith D. Shepherd[b], Mercy Nyambura[b], Chris J. Gilmore[c], Nikolaos Kourkoumelis[d], Stephen Hillier[a,e]

[a] The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, UK
[b] World Agroforestry Centre (ICRAF), P.O. Box 30677-00100 GPO, Nairobi, Kenya
[c] School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK
[d] Department of Medical Physics, School of Health Sciences, University of Ioannina, 45110 Ioannina, Greece
[e] Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), SE-75007 Uppsala, Sweden

ARTICLE INFO

Handling Editor: David Laird

Keywords:
Soil mineralogy
X-ray powder diffraction
Cluster analysis

ABSTRACT

X-ray powder diffraction (XRPD) is widely applied for the qualitative and quantitative analysis of soil mineralogy. In recent years, high-throughput XRPD has resulted in soil XRPD datasets containing thousands of samples. The efforts required for conventional approaches of soil XRPD data analysis are currently restrictive for such large data sets, resulting in a need for computational methods that can aid in defining soil property – soil mineralogy relationships. Cluster analysis of soil XRPD data represents a rapid method for grouping data into discrete classes based on mineralogical similarities, and thus allows for sets of mineralogically distinct soils to be defined and investigated in greater detail. Effective cluster analysis requires minimisation of sample-independent variation and maximisation of sample-dependent variation, which entails pre-treatment of XRPD data in order to correct for common aberrations associated with data collection.

A 2^4 factorial design was used to investigate the most effective data pre-treatment protocol for the cluster analysis of XRPD data from 12 African soils, each analysed once by five different personnel. Sample-independent effects of displacement error, noise and signal intensity variation were pre-treated using peak alignment, binning and scaling, respectively. The sample-dependent effect of strongly diffracting minerals overwhelming the signal of weakly diffracting minerals was pre-treated using a square-root transformation. Without pre-treatment, the 60 XRPD measurements failed to provide informative clusters. Pre-treatment via peak alignment, square-root transformation, and scaling each resulted in significantly improved partitioning of the groups ($p < 0.05$). Data pre-treatment via binning reduced the computational demands of cluster analysis, but did not significantly affect the partitioning ($p > 0.1$). Applying all four pre-treatments proved to be the most suitable protocol for both non-hierarchical and hierarchical cluster analysis. Deducing such a protocol is considered a prerequisite to the wider application of cluster analysis in exploring soil property – soil mineralogy relationships in larger datasets.

1. Introduction

X-ray powder diffraction (XRPD) is a widely employed method in the study of complex mineral mixtures, with the varied mineral assemblages of soils providing particularly apposite examples (Bish, 1994; Dixon and Schulze, 2002). Conventional approaches to the assessment of soil mineralogy by XRPD typically involve a first stage of identification of the minerals present and a subsequent stage that seeks to quantify the relative abundance of the different minerals identified in the soil. In such a conventional approach the first stage of identification is typically made by an iterative process using reference databases and monographs that tabulate data for the diffraction patterns of different minerals that may be encountered (ICDD, 2016; Brindley and Brown, 1980; Dixon and Schulze, 2002; Harris and White, 2008). This tabulated reference data usually records peak positions in terms of their 'd-spacings' in Ångstrom (or nanometers) together with their relative intensities (0–100%). Despite the availability of automated search match procedures and other software tools to aid the identification stage, round robin evidence suggests that the process still relies heavily on the experience of the analyst to correctly identify (Raven and Self, 2017) the minerals in samples like soils.

Compared to mineral identification, the next conventional step of mineral quantification is widely acknowledged as a much more complex task (Omotoso et al., 2006). Quantification seeks to relate variation

* Corresponding author.
E-mail address: benjamin.butler@hutton.ac.uk (B.M. Butler).

https://doi.org/10.1016/j.geoderma.2018.09.044
Received 15 May 2018; Received in revised form 19 September 2018; Accepted 22 September 2018

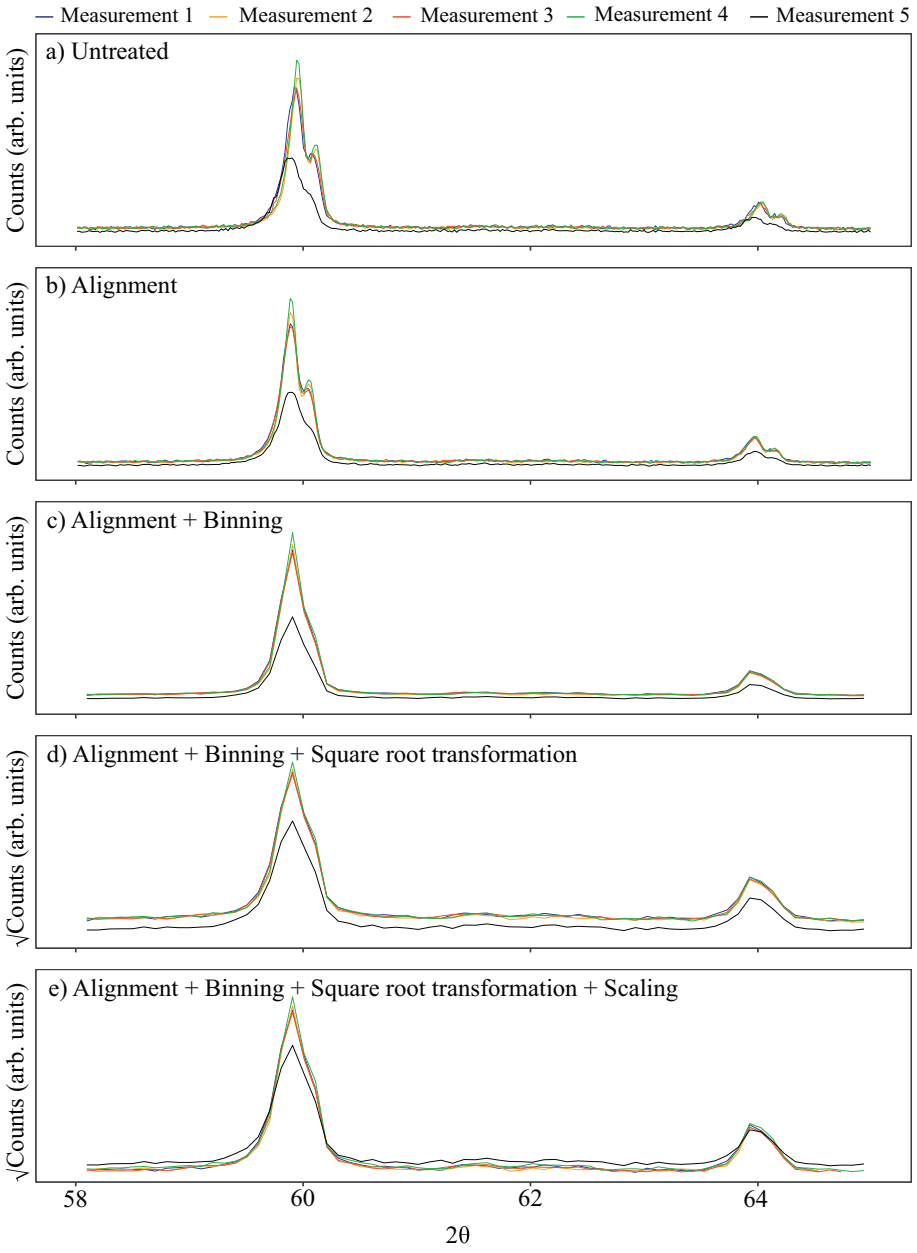

Fig. 1. Effects of sequential pre-treatment steps on XRPD data illustrated using data from a single soil sample measured once by five different personnel. The 2θ range has been reduced to 58–65° to aid comparison between pre-treatments. Untreated data (a) display variations in peak alignment, noise, and signal intensity. Pre-treatment via alignment (b) results in suitably aligned peaks, and causes a slight smoothing of the data due to a linear spline interpolation used to harmonise the aligned data to the same 2θ scale (Section 2.3). Subsequent binning of the data (c; bin width = 5) acts to further reduce the noise whilst retaining sufficient mineralogical information, and simultaneously acts as a form of data reduction. Subsequent pre-treatment by square root transformation (d) re-scales the data so that minor peaks become emphasised relative to major peaks. Lastly, subsequent pre-treatment by scaling (e; mean centering) corrects for most of the variation in signal intensity between samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in the intensity of either individual mineral peaks or of the full patterns (i.e. all peaks) of each mineral to its concentration in the sample, usually expressed in weight %. Such quantitative analyses are most often made on multiple samples, so that variation in mineral abundance can be compared from one soil sample to another and ultimately interpreted in relation to soil properties and functions that are dependent on soil mineralogy. Again the procedures that may be used to perform quantitative mineralogical analyses of soils rely heavily on the experience of the analyst to ensure that the results obtained are reliable and fit for purpose (Raven and Self, 2017).

The data of modern XRPD methods which all these procedures require is a precisely measured digital diffraction pattern typically comprised of discrete 'Bragg' diffraction peaks varying in intensity (y), expressed for example in counts, along an experimental axis (x) usually expressed in degrees 2θ. The 'Bragg' peaks from crystalline mineral phases rise above a background, which may also include diffuse scattering from X-ray amorphous phases. For example organic matter and volcanic glass can be common amorphous phases in many soils (Dixon and Schulze, 2002).

In recent years the availability of digital XRPD soil data has increased, and attempts are now being made to generate datasets containing thousands of spatially referenced XRPD measurements [e.g. those collected for the National Soil Inventory of Scotland (NSIS) and the Africa Soil Information Service (AfSIS)]. Since many soil properties are closely related to soil mineralogy (Butler et al., 2018; Newman, 1984), such datasets in combination with computational data analysis represent unique opportunities to advance the understanding of the role of soil minerals in governing or influencing many soil properties, processes and functions.

In data-rich cases like NSIS (Butler et al., 2018) and AfSIS (Towett et al., 2015), computational methods of XRPD data analysis become particularly attractive because they are time-efficient and do not necessarily require any classical expert interpretation of the XRPD patterns, at least not until the final stages of such an analysis. This may seem like the analysis is initially disconnected from the crystallographic origins of the data, but this is the foundation of data-driven, or digital, approaches to soil mineralogy (Butler et al., 2018; Hillier and Butler, 2018). Cluster analysis is one such 'digital' approach that can aid the