Combining wavelets with statistical inference to map the mineralogical composition of pedological features from synchrotron X-ray diffraction data

Claude Manté, Daniel Borschneck, Cristian Mocuta, Romain van den Bogaert, David Montagne, Cécilia Cammas, Sophie Cornu

Combining wavelets with statistical inference to map the mineralogical composition of pedological features from synchrotron X-ray diffraction data.

Claude Manté · Daniel Borschneck · Cristian Mocuta · Romain van den Bogaert · David Montagne · Cecilia Cammas · Sophie Cornu

the date of receipt and acceptance should be inserted later

Abstract Clay translocation is among the most widespread processes in soils. It is generally identified by the presence of clay coatings at the macroscopic or microscopic scales. Nevertheless, several authors demonstrated that clay coatings have different origins, which renders the attribution of individual coatings to a particular process difficult. Therefore, their characterization at the microscopic scale is necessary. Modern synchrotron technics allow mapping of the mineral composition of soil thin sections by X-ray diffraction with a lateral resolution of a few micrometres that is compatible with the size of the clay coatings. However, the use of this technic raises a certain number of technical difficulties when clay minerals and small pixel size are considered. Therefore a preliminary feasibility analysis was performed on soil coating obtained experimentally in a soil column

Claude Manté
Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU,IRD, MIO,UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France
Tel.: +330486090631
E-mail: claude.mante@gmail.com

Daniel Borschneck
Aix Marseille Université, CNRS, IRD, INRAE, Coll France, CEREGE, Aix-en-Provence, France

Cristian Mocuta
Synchrotron SOLEIL, L'Orme des Merisiers, Saint-Aubin, BP 48, Gif-sur-Yvette 91192, France

Romain van den Bogaert
Aix Marseille Université, CNRS, IRD, INRAE, Coll France, CEREGE, Aix-en-Provence, France

David Montagne
AgroParisTech, INRAE, Université Paris-Saclay, UMR ECOSYS, 78850 Thiverval-Grignon, France

Cecilia Cammas
AgroParisTech, INRAP, UMR Archéologie des sociétés méditerranéennes, 78850 Thiverval-Grignon, France

Sophie Cornu
Aix Marseille Université, CNRS, IRD, INRAE, Coll France, CEREGE, Aix-en-Provence, France

experiment. We developped a mathematical/statistical method to automatically identify and map the minerals present in synchrotron X-ray diffraction maps. This method combines a subtraction of the background signal, an identification of the diffraction peaks and the attribution of the obtained peaks to the different minerals phases. The robustness of the method was tested for the two first steps that were critical. We then showed that most of the minerals present in the considered samples could be identified, including clay minerals and Fe oxides; however, their relative proportions were difficult to estimate within the experimental conditions. As a conclusion, we proposed here a new mathematical method for the data treatment of synchrotron X-ray diffraction. Our study shows that this method could be applied on natural heterogeneous samples with data of poor quality to perform robust qualitative analysis. Moreover, with the recently available fast data acquisition schemes, large amounts of datasets (diffractograms) can be rapidly acquired, and are expected to benefit of the automatic data treatment approach proposed here.

Keywords Signal processing · Filtering · Mineralogy · Clay translocation · Pedogenesis

Mathematics Subject Classification (2010) 92F99 · 62-07 · 60G35 · 65D10

1 Introduction

Clay translocation is among the most widespread processes in soils. This process is defined as a substantial vertical transfer of particles from a horizon, called eluviated horizon, to another horizon referred to illuviated horizon. Lessivage is generally identified in soils by the presence of clay coatings and fillings at the macroscopic or microscopic scales (Jamagne, 1973; McKeague et al., 1980; Thompson et al., 1990; Gunal and Ransom, 2006; Gutiérrez-Castorena et al., 2007). Clay translocation has been considered as occurring mostly under boreal conditions and consequently as a fossil process in temperate conditions (Van Vliet-Lanoë, 1990). Recent research however clearly demonstrated that this process is still active nowadays in temperate climate (Montagne et al., 2013; Sauzet et al., 2016) and that its intensity is particularly sensitive to anthropic or climatic global changes (Montagne and Cornu, 2009; Montagne et al., 2016). As increasing intensity of clay translocation may correspond to increasing risks of erosion and compaction in surface soil horizons, to increasing transfer of soil contaminants associated with eluviated soil particles but also to increasing long term C sequestration in subsoil horizons (Torres-Sallan et al., 2017), it seems thus necessary to better understand its mechanisms notably at the microscopic scale.

Indeed while few studies deal with the characterization and quantification of clay coatings by image analysis to our knowledge (Terribile and FitzPatrick, 1992; Protz and VandenBygaart, 1998; Sauzet et al., 2016, 2017), none of them proposes a mineralogical characterization of illuvial clay features. Clay coatings being generally 60 to 200 μm thick (Dalrymple and Theocharopoulos, 1987; Thompson et al., 1990; Sauzet et al., 2017), the selective sampling of clay coatings to characterize their mineralogical composition with laboratory X-ray diffractometers is generally not feasible. The mineralogical composition of clay coatings and fillings therefore

remains largely unknown, making it impossible to specify the mineralogical selectivity of the translocation process or its dependence on the conditions of formation of clay coatings. Micromorphological studies have however demonstrated that the composition of clay coatings and infillings depends on their deposition conditions. Pure micro-laminated and limpid clay coatings are associated with a homogeneous and low energy flow of water under permanent plant cover while "dusty" clay coatings containing organic and silt-sized particles are associated with a flow of high energy water under cultivation (Fedoroff and Courty, 1994; Jongmans et al., 2001; Kuhn et al., 2010)... Modern synchrotron technics allow mapping mineral composition on soil thin sections by (monochromatic) X-ray diffraction (XRD) with a lateral resolution of a few micrometres that is compatible with the size of the clay coatings. Classically in the literature, synchrotron XRD is either used on powder to detect mineral phases that are hardly detectable by laboratory XRD apparatus, as iron oxides for example (Schulze and Bertsch, 1995; Sumoondur et al., 2008; Lopano et al., 2011; Yusiharni and Gilkes, 2012; Tsao et al., 2013), or in association to micro X-ray fluorescence (XRF) or micro-EXAFS to identify phases bearing trace elements on thin sections (Dillmann et al., 2002; Manceau et al., 2005; De Nolf et al., 2009; Fan and Gerson, 2011; Stroh et al., 2014; Sutton et al., 2017; Siebecker et al., 2018), among others. While dealing with small investigated volumes (small X-ray beams and thin section samples), the probed sample cannot be anymore assimilated to a random orientated powder. As a consequence of the few number of illuminated crystallites (Gueriau et al. (2020), in press), the Bragg diffraction condition is scarcely satisfied and not all the peaks are recorded for a given mineral phase (Manceau et al., 2005; Stroh et al., 2014; Siebecker et al., 2018). To solve this problem, most authors sum up individual XRD-pattern on a chemically homogeneous zone of interest (Sutton et al., 2017, among others) or use larger X-ray beams, e.g. of size of several tenth of μm (Dillmann et al., 2002; De Nolf et al., 2009; Sutton et al., 2017) in order that the probed volume resembles more to a random orientated crystalline powder. Also, XRD-maps are rarely presented with the notable exception of De Nolf et al. (2009) who used a 60 μm pixel size. In addition, while clay minerals as kaolinite and illite were detected by synchrotron XRD, mainly on powder samples (Manceau et al., 2005; Tsao et al., 2013; Ndzana et al., 2019), in the case of orientated clay coating, depending on the orientation axis, detection of these mineral might be more challenging. At last under these conditions (few number of crystallites in the X-ray probed volume), classical data treatment approaches as Rietveld refinement and other full pattern fitting methods that were proved very efficient to quantify the mineral composition of a sample cannot be used as mentioned by Siebecker et al. (2018) "often, peak-matching is challenging because minerals can have overlapping peaks, and the number and intensity of the peaks can be disproportionate to the reference database". In addition, due to the large amount of XRD-patterns acquired automatic processing procedure is required if mapping mineral is one of the pursued objective. The aim of this paper was thus to study the feasibility of mapping the mineralogy of clay coatings by synchrotron X-ray diffraction. To this aim, we analyzed coatings from thin sections obtained in an experiment of in vitro pedogenesis on soil columns, performed by Cornu et al. (2014). Due to the high number of diffractograms acquired, we developed an automatic processing procedure that combines a subtraction of the background signal (both from thin section mount and from mathematical modeling), an identification of the diffraction peaks and

an attribution of the obtained peaks to different minerals phases in order to draw various contrasts maps.

2 Material and methods

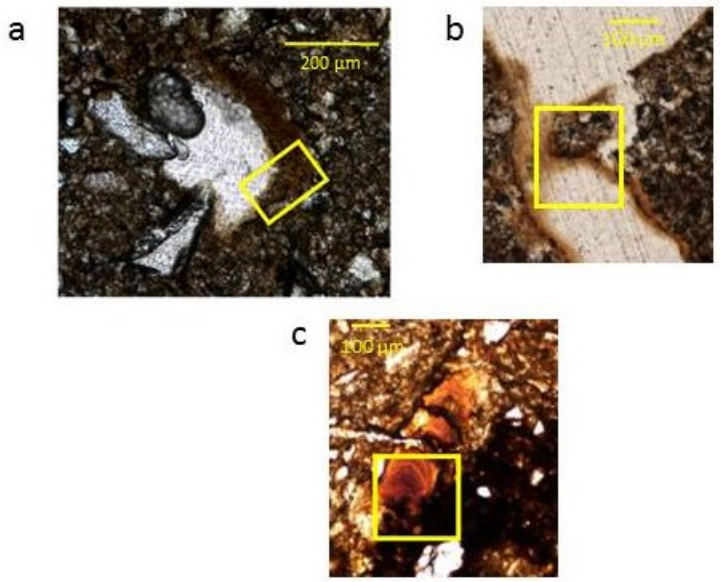
2.1 Experimental design

Cornu et al. (2014) performed clay translocation laboratory experiments on two E-horizons of Luvisols developed in loess with contrasting pH and clay mineralogy. For eluviation, an E-horizon, called L1, containing smectites and with a pH of 6.2 was chosen, while for illuviation, a E-horizon, called L2, containing no smectite and with a pH of 7 was chosen. Undisturbed cylinders (with height and diameter of 15 cm) were sampled at a depth between 35 and 50 cm. To monitor illuviation (particle deposition), complex columns were built consisting of an undisturbed L1-cylinder overlaying an undisturbed L2-cylinder. Thirty rain events were applied to the soil columns using a rain simulator. The considered L2 cores contained quartz, feldspars, kaolinite, illite, chlorite, iron oxides (goethite, ferrihydrite notably) and accessory minerals, and eventually smectite coming from the upper L1 cylinders. For more details see Cornu et al. (2014).

2.2 Thin sections fabrication, and observation by optic microscope

On four L2 cores (named R, S, AR and P), undisturbed blocks (80 mm x 50 mm) were sampled both at the top and at the bottom of the core (named /T and /B respectively), for thin sections. The blocs were air-dried for 2 months, then oven-dried at 40°C for 2 weeks, and impregnated under a vacuum of -5 kPa with a polyester resin diluted to 30% by volume with a styrene monomer (Bruand et al., 1996). Eight uncovered thin sections, 50 x 50 mm in size and 25 μm thick, were prepared on a glass slide and observed by using a petrographic microscope Nikon Eclipse 50ipol optical microscope in both plane polarized light (PPL) and crossed polarized light (XPL) at the magnification of 20x, 40x, 100x and 200x. Three different microscopic units (MU) are observed: the soil matrix (skeleton and plasma), voids and voids coating made of fine material, and of varying thickness. The skeleton is composed of sandy and silty particles embedded in a finer brownish silty-clay plasma containing some fine brown organic particles common in E horizons. The dark hue of the matrix is typical of Fe and/or Mn oxides. The skeleton includes mainly quartz grains. Pores are mainly channels, which are the traces of burrowing by soil macrofauna and roots, and vughs.
Three types of coatings may be distinguished: a dark brown, dusty and slightly birefringent one (DSB) (Fig 1a), a limpid and birefringent one (LB) (Fig 1b), and a highly birefringent laminated one (HBL) (Fig. 1c). Increasing birefringence is interpreted as increasing orientation of the clay particles inside the clay coatings or infillings, and is linked to decreasing energy of the water flows. Limpid and dusty coatings are respectively associated with pure and impure clay deposits. Impure clay coatings contain mineral or organic particles mobilised from the soil surface. The DSB and LB coatings are generally thin and may be observed from the top to the base of the soil column (Table 1) suggesting that they may have been

Fig. 1 Micromophology of (a) a dusty and (b) a clear clay coating surrounding a pore, and (c) a clay cutane with internal lamination. The yellow squares are the regions-of-interest that have been scanned with μXRD.



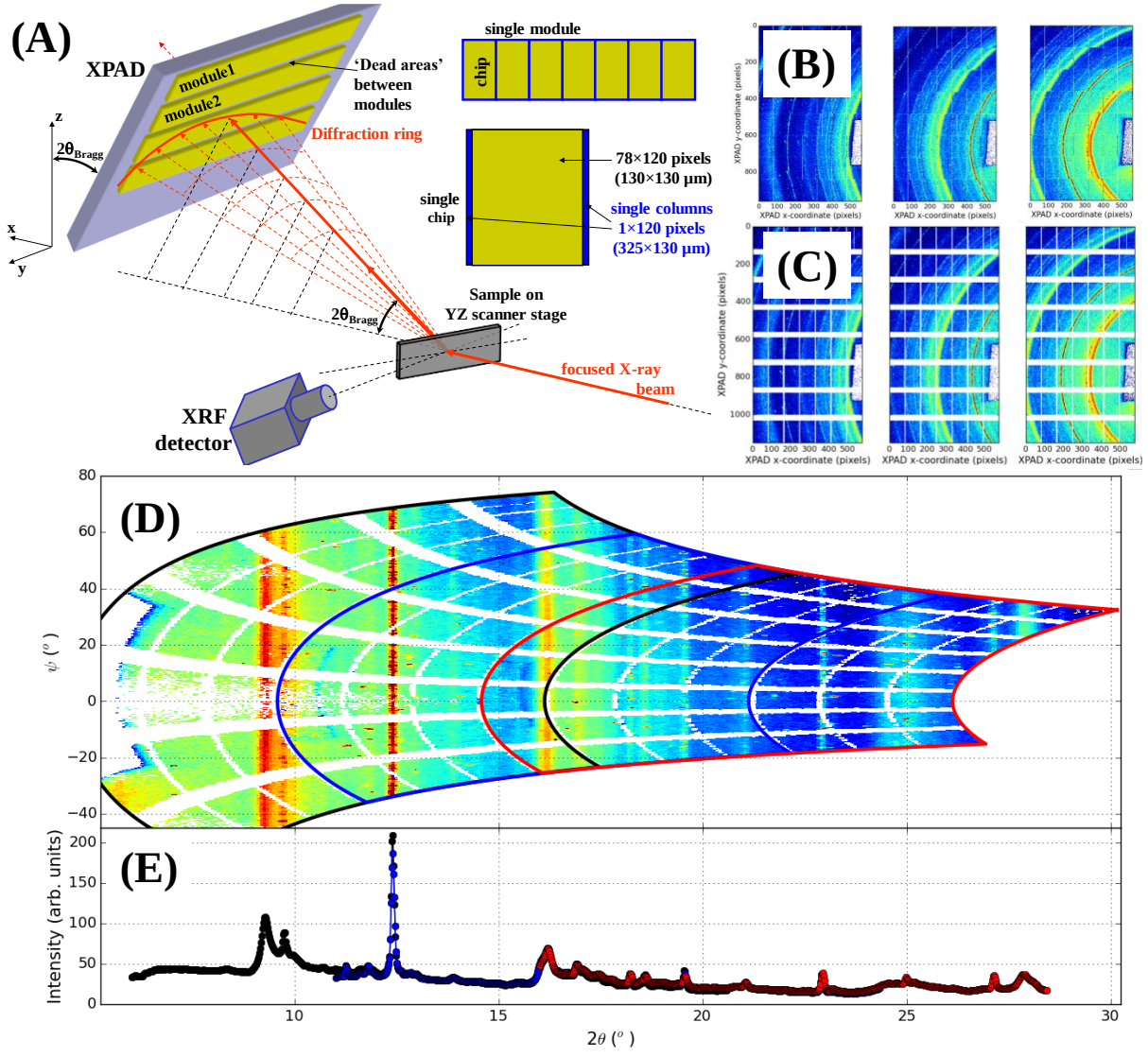
formed during rain simulations and thus may contain smectites. Their boundaries are diffuse as the clay fill the irregularities of the edge of the pores. Highly birefringent, thick and microlaminated coatings and infilling with iron staining (HBL) are mostly observed at the base of the soil column and are typical of illuviated horizons suggesting that they may have been formed under field conditions.
On each thin section, zones containing the portions of the different MU: matrix, pore and one particular kind of clay coating were selected for X-ray diffraction characterization. Three thin section zones are more specifically studied here in order to target the different kinds of clay coatings and finally identify potential relationships between clay coatings morphology, conditions of deposition and mineralogy. Image analysis (supervised classification) allowed the segmentation of the zones in the three MU as already performed by Montagne et al. (2007).

2.3 Synchrotron-$\mu\,XRD$ and $\mu\,XRF$ analysis

The three selected zones of coatings were mapped by coupling $\mu\,XRD$ and $\mu\,XRF$ on the DiffAbs beamline of the SOLEIL Synchrotron (France) in transmission mode at 17.2 keV X-ray photons energy (i.e. $\lambda = 0.72\,\mathring{A}$), and a beam of 10$\mu$m per 6$\mu$m (FWHM, horizontal x vertical direction) with a photon flux of several

1010 ph./s in the focused beam. The experimental setup is sketched in figure 2-A: the sample (see details above) was placed vertically on a YZ scanner stage (μm accuracy), with its surface normal rotated by about 20° with respect to the incident X-ray beam, to avoid shadowing effects for measuring the XRF signal. This small angle has as consequence a slight increase of the beam footprint on the sample, in the horizontal direction, by about 6.5 percents. The XRF silicon drift detector (SDD, Vortex®-EM X-ray Detector from Hitachi, with 50 mm^2 active area) is placed in the horizontal plane, at 90° from the incident X-ray beam. Since the X-ray beam is polarized in the horizontal plane, this particular geometry insures that the elastic scattering is minimized (zero in theory). The area detector used for μXRD analysis is a hybrid pixel area detector, XPAD 3.2 that works in a single-photon-counting pixel mode. It consists of 8 modules assembled like shown in the figure 2. Each module consists of 7 hybrid integrated circuits (chips) on a single 500 μm thick silicon sensor (Medjoubi et al. 2012) which rear face pixelized and each pixel is coupled via 'bump-bonding' to an electronic counting device in a dedicated circuit. Each chip consists of 80 x 120 pixels of 130 μm size, except the first and last columns if situated at the junction with the next chip. They have a size precisely 2.5 times larger along the neighbor chip direction. Due to the module assembling on a flat plane, an inter-module gap of about 27 pixels (about 3.5 mm of dead area) is observed. Thus, the total area covered by the detector is 75.1 x 149.3 mm^2. Knowing this detector geometry in detail, and combining it with reference X-ray diffraction images (ex. XRD taken on calibration powders), reliable image and geometry corrections can be carried out to extract the position in space, and then in the angular space, for each pixel (figure 2). The detail of these corrections is described in several references (He, 2009; Le Bourlot et al. 2012; Mocuta et al., 2013). For this experiment, the sample-detector distance used was of 369.9 mm, yielding to an angular opening of about 0.02° per pixel. For all the reported results on thin sections, a counting time of 4 s was used; XRD and XRF data are acquired simultaneously, thus on precisely same region of the sample. The choice of the acquisition time is a compromise between signal statistics (for detecting XRD, mostly) and total duration to measure one sample, in order to remain compatible with the possibility of measuring several samples during the allocated beamtime at synchrotron. Since the XPAD detector analysed angular domain is of about 10 - 15°, three different acquisitions were performed for 3.5 -17; 9.3-24.7; 19.3-33.4 degrees ranges in 2θ. Glass plus resin areas were also analysed to have a background signal as well as $< 2\mu m$ fraction samples in powder, in order to verify that the peak of interest were detected in the analytical configuration used (see Appendix A). All the XPAD data were pre-processed for image correction as mentioned above (Le Bourlot et al. 2012; Mocuta et al., 2013). The 2D diffraction patterns (XPAD images)s were transformed into diffractograms (intensity vs. $2\,\theta$ curves) using Python codes (Fig. 2). 'Background' like signal (obtained on glass and resin alone) was systematically subtracted. One can note that both the XRD signal and the 'background' have rather noisy baselines, so this subtraction is generally insufficient and further processing of the data was necessary (see below).

Fig. 2 (A) Illustration of the synchrotron set-up for XRD and XRF mapping with an X-ray microbeam: the sample is mounted on a micron-accuracy scanner stage, and surrounded, as shown, by the two detectors. The schematics of the XPAD detector (modules and chips assembly) is also shown, after Mocuta et al. (2013); Powder diffraction data acquired on a powder sample ($< 2\,\mu m$ fraction of the L2 soil) using the XPAD detector, for three 2θ positions (of about 3.5, 13.5 and 23.5 degrees respectively) and the data conversion procedure, to obtain XRD Intensity vs. 2θ datasets : (B) raw XPAD images (Log10 color scale, from blue to red); (C) the previous images are first transformed to mask the inter-module gaps and broader inter-chip pixels; (D) then the angular coordinates of each pixel are calculated from the known particular geometry and reported in maps with the horizontal axis being the scattering angle (2θ) and the vertical one being ψ, the azimuth measured on the XRD ring, zero being its apex point, situated in the xz plane (cf. panel (A)). In this representation, the XRD rings appear as vertical lines (2θ = constant). The borders of the 3 XPAD images / datasets patched to generate this figure are shown by the colored contours (black, blue and red respectively); (E) The XRD (I vs. 2θ) curve is obtained from each dataset by performing a ψ azimuthal regrouping (histogram like), and shown using the previously mentioned colors (black, blue, red).



2.4 Automatic identification of XRD peaks (pixel scale)

As mentioned in the introduction, at the pixel scale, only few, if any, XRD peak may be encountered, which may be very small in intensity. Their automatic identification thus requires the identification of relevant peaks, after removing both the noise and the baseline.

2.4.1 Removal of the remaining baseline

To detect the peaks, the remaining part of the background signal either linked to the glass slide mount or to the detection apparatus, had to be removed. At first sight, this is similar to what was done by Antoniadis et al. (2007, 2010) for mass spectrometry data. These authors modeled each spectrum Y by the equation:

$$Y(x) = L(x) + S(x) + \varepsilon \tag{1}$$

where x is the molecular mass, L is the baseline (some smooth function), S is the genuine signal (peaks associated with proteins) and ε is a white noise. Technically, Antoniadis et al. (2007, 2010) used first the stationary wavelet transform algorithm (Mallat, 1999, Section 10.2.3) to pick out the noise, and then the R package of penalized quantile smoothing spline COBS (Ng and Maechler, 2007) to estimate the baseline. We adapted their method to the specific nature of our data. Unlike Antoniadis et al. (2007, 2010), who considered mass spectrometry peaks having a quantitative meaning, the peaks in XRD-patterns have a semi-quantitative meaning. In addition, the signal over noise ratio of our data is much lower than in Antoniadis et al. (2007, 2010).
To detect significant peaks, we propose a two-steps strategy:

1. baseline correction, thanks to COBS: $Y \mapsto \left(Y - \widehat{L}\right)$, where $\widehat{L}$ denotes the baseline estimate
2. stochastic denoising, thanks to an original method (see Section 2.4.2):

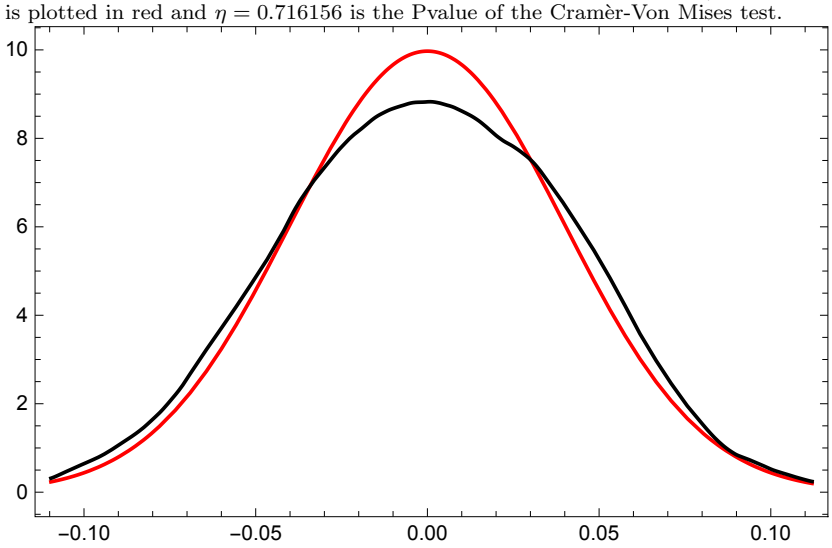
$$\left(Y - \widehat{L}\right) \mapsto \left(Y - \widehat{L}\right) - \hat{\varepsilon} \approx S. \tag{2}$$

Like in Antoniadis et al. (2007, 2010), the smoothing parameter λ of COBS (Ng and Maechler, 2007) was automatically chosen according to the Schwartz Information Criterion, but we chose as a quantile level $\tau = 0.5$ (median regression).

Remark 1 *For the first angular domains, ranging from 3.5 to 17 degrees of angles, because of the steepness of all the associated patterns near the origin (see figure 5), it was impossible to obtain a satisfying fit with median regression in a standard way. This case has been processed in a special way detailed in Appendix B.*

While Antoniadis et al. (2007, 2010) chose a rather large number of knots ($K = 60$) for fitting L (adapted to the length of their series: about 18000 points (Coombes et al., 2003)), we used smaller values: $K = 10$ for the angular domains ranging from 9.3 to 24.7 and 19.3 to 33.4 degrees of angle respectively, and $K = 20$ for the angular domain ranging from 3.5 to 17 degrees of angles (remember that K controls the smoothness of L). We could this way determine a sufficiently smooth estimate $\widehat{L}$ of the baseline, and separate it from $S + \varepsilon$ (see Equation 1). We finally obtained for each angular domain and pixel i the **detrended XRD signal** $\mathcal{D}_i := Y_i - \widehat{L_i}$ from the original diffractogram Y_i.

Fig. 3 Kernel density estimate of the noise distribution (in black), for a fixed pixel from a zone, and a fixed angular domain; the associated Gaussian distribution ($\mu = 0$, $\sigma = 0.0400082$) is plotted in red and $\eta = 0.716156$ is the Pvalue of the Cramèr-Von Mises test.



2.4.2 An inference method to identify significant peaks in $\mathcal{D}_i$

Due to the size of the analyzed pixel, each diffraction signal Y_i as well as the detrended one $\mathcal{D}_i$ were noisy and of low intensity. We thus considered as a diffraction peak only those higher than a given number of times the standard deviation of the noise. Consequently, we had to estimate the standard deviation (SD) σ of the noise ε before eliminating it thanks to the three-σ rule evoked by Pukelsheim (1994).

Stage 1: estimation of σ

We drew from each zone a random sample of N pixel positions and obtained the required estimate of σ, **for each angular domain**, according to the following algorithm (detailed in Appendix C):

1. remove the noise from each detrended signal, thanks to stationary wavelets denoising: $\mathcal{D}_i \mapsto \mathcal{D}_i - \widehat{\varepsilon_i}$
2. check the normality of the noise
3. robustly estimate its standard deviation σ.

A typical distribution of noise is displayed on figure 3. The normality of the noise was practically never rejected (Cramèr Von Mises test of level 0.05), in accordance with model (1); consequently the general hypothesis

$$(H): \ \varepsilon^m = \mathcal{N}(0, \sigma^m)$$

of a Gaussian noise affecting each angular domain m, for each pixel, was accepted.

Stage 2: binarization

Impulse noise can be observed on the XRD-pattern and must thus be discarded. Thus, once the SD of the noise affecting the m^{th} angular domain of the studied zone, σ^m, had been estimated, we binarized the detrended XRD-patterns by using the following two-steps algorithm.

1. Detect **stochastic peaks under (H)** by binarizing each $\mathcal{D}_i$ thanks to formula (6) of Appendix D, according to some fixed threshold η_0
2. keep only those stochastic peaks which can be **simultaneously** considered as local maxima of the smoothing cubic spline $\mathscr{S}\left(\mathcal{D}_{\mathrm{i}}, \widehat{\sigma^{\mathrm{m}}}\right)$ whose smoothing parameter λ is determined from $\widehat{\sigma^m}$ by minimizing the Unbiased Risk Estimate of Wahba (1990), implemented in R by Wang and Ke (2004).

Remark 2 *In the first step above, we fixed* $\eta_0 := P\left(|Z| > 3\ \widehat{\sigma^m}\right)$ *as a threshold (three-*σ *rule of Pukelsheim (1994)) , where* $Z = \mathcal{N}\left(0, \widehat{\sigma^m}\right)$ *and* $\widehat{\sigma^m}$ *is the robust estimation provided by Stage 1; this may seem severe, but notice that high-energy physicists, for instance Collective (2007), currently use* 5 σ *as a threshold.*

Finally, each $\mathcal{D}_i = \left\{\left(a_1^i,\, d_1^i\right), \cdots, \left(a_J^i,\, d_J^i\right)\right\}$ was binarized according to the rule:

$$\mathcal{D}_i \mapsto \mathcal{B}_i := \left\{\left(a_1^i,\, b_1^i\right), \cdots, \left(a_J^i,\, b_J^i\right)\right\}$$

where $b_k^i = 1$ if a_k^i is the position of some retained peak and otherwise $b_k^i = 0$.

2.5 From peaks to images of minerals: presence indicators

The method proposed above allowed to identify the XRD-peaks for each pixel. These peaks must then be associated with minerals (through "theoretical" binary reference spectra (see Appendices E.1 and E.2) associated with minerals), and mapped.

2.5.1 Mineral phases: identification and image

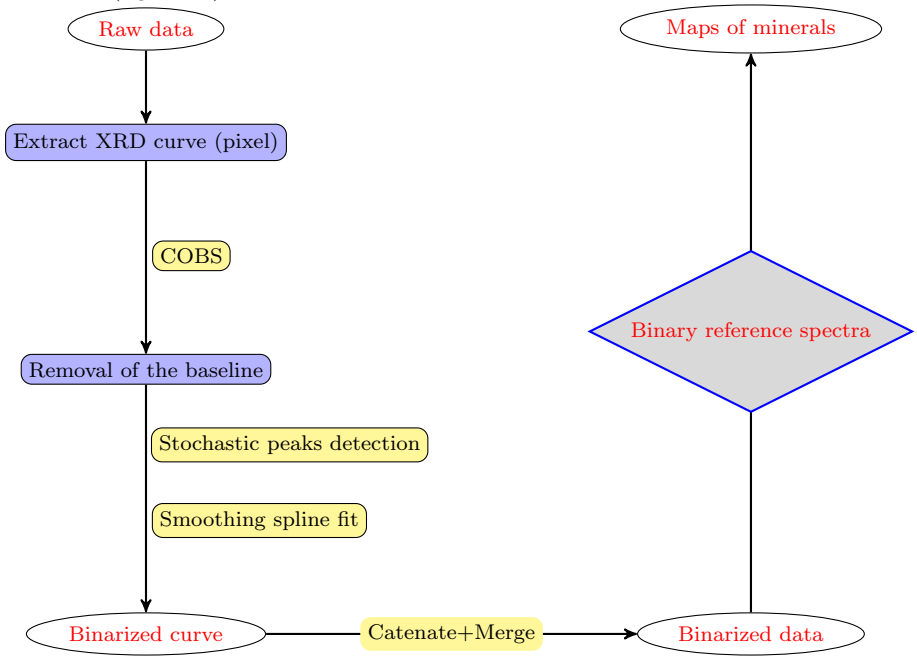
For each analyzed thin sections, XRD-patterns were averaged by (MU pore, matrix, clay coating or cutane), based on the attribution of each pixel to one of these features by image analysis, as described above. The obtained XRD-patterns were then interpreted classically and a list of XRD peaks encountered was established and reported in Table 1. Peaks that were common to several minerals were differentiated from peaks belonging to a single mineral, also called exclusive peaks (Table 1). Each mineral is characterized by a family of diffraction peaks, which were defined as **characteristic intervals** (because of possible experimental shift). Since these intervals are different from one mineral to another, it was necessary to construct a common system $\mathcal{T}$ of **disjoint** intervals such that any original spectrum could be reconstituted in $\mathcal{T}$ (the construction of $\mathcal{T}$ is detailed in Appendix E.1). In concrete terms, $\mathcal{T}$ consists in this study of 442 intervals. Since some XRD-peaks may be associated to several minerals while other may be exclusive of a single mineral, a ponderation was applied to give more weight to exclusive peaks than to peaks shared by several minerals (see Appendix E.1). We

then defined indicators assigning to each peak a weight depending on the number of minerals that shares it. This point is detailed in Appendix E.3, where are defined the **Presence Index** of some mineral s at some position (x_i, y_i), $PI(s; x_i, y_i)$, and its **Weighted Presence Index**, $WPI(s; x_i, y_i)$, which takes into account the **exclusiveness** of each peak.

2.6 Computational aspects: a digest

On figure 4, we synthesize by a flowchart the sequence of processing proposed. The left side of this figure corresponds to a *Mathematica* package, calling two R packages: COBS (Ng and Maechler, 2007) for baseline removing, and assist (Wang and Ke, 2004) for fitting the residual of COBS by some smoothing spline. A second *Mathematica* package has been dedicated to the right part of figure 4: conversion of binarized DRX data into maps of presence indicators of minerals of interest, thanks to reference spectra provided by experts. Another preliminary *Mathematica* package has been written, to build the family $\mathcal{T}$ of intervals (see Appendix E) associating 2 θ coordinates with characteristic intervals. The whole processing lasts around an hour per image (depending on its size, its signal to noise ratio, *etc*).

Fig. 4 First part (left side): sketch of processing at the pixel level, for each angular domain; the resulting binary curves are then merged and compared with reference spectra of minerals of interest (right side).



3 Results and discussion

3.1 Stability of the peak detection and identification process

In the above described approach two key steps were identified: the background removal method, and the choice of the number of σ as a threshold for peak detection.

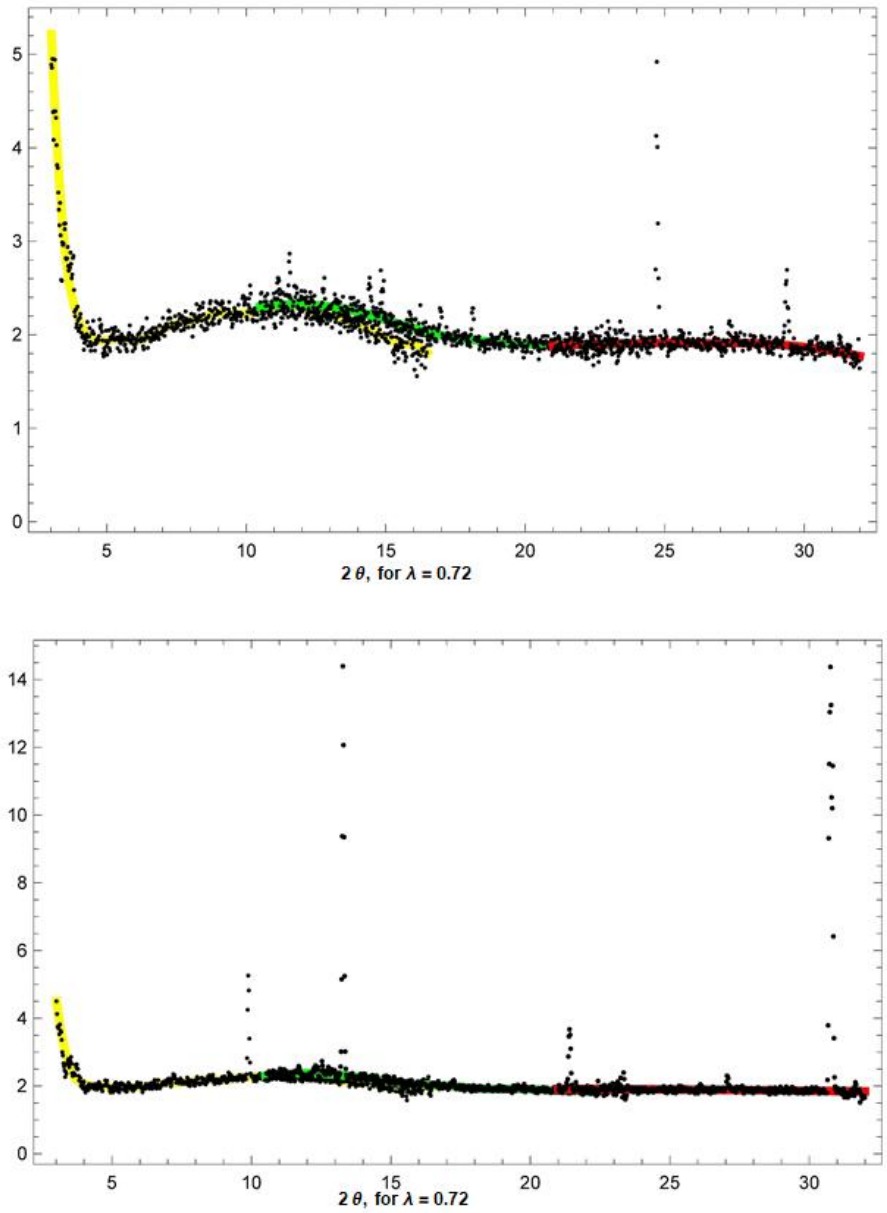
3.1.1 Removal of the background

As seen on figure 5, the background of the obtained XRD-pattern may be quite complex and varies according to the angular domain considered. It comprises a decrease of exponential type of the signal from 3 to 5° in 2θ and a large bulge ranging from 7 to 15° in 2θ (figure 5). The exponential decrease in small angles is classical (principally due to air and small particles scattering and other phenomenons like Compton scattering, fluorescent radiation... (Gaast and Vaars, 1981). The bulge corresponds to the signature of the glass slide (amorphous material) on which the soil thin section was mounted. The subtraction of an experimental XRD-pattern obtained for a zone of the thin section containing only the resin from the raw XRD-patterns before pre-treatment was used to remove this bulge. It however results in an increasing noise. That is why a supplementary mathematical removal of the background was performed. Both the exponential decrease and the residual bulge were thus removed mathematically from the XRD-signal for the automatic analysis of the XRD-patterns of the individual pixels as shown on figure 5 in red, green and yellow. As seen on this figure, the chosen algorithm was robust enough to remove the background whatever its shape for the different angular domains considered.

3.1.2 Binarization strategy: choice of the threshold k σ

In order to differentiate diffraction peaks from noise, the results of Pukelsheim (1994) was used for determining stochastic peaks. This type of threshold is classical in signal analysis but can range from as high as five SD, as in high-energy physics Collective (2007), or as low as two sigmas in geochemistry. On figure 6, we reported the peaks identified for two XRD-patterns using 2, 3 or 3.5σ as a threshold. Some of the peaks identified with 2σ do not really come out of the background level and would not have been interpreted as peaks by an expert (see for examples peaks at 21.2, 28.2 and 31.5 degrees in 2θ in figure 6a and at 10.8, 11, 18.2, 21.8, 22.8 and 23.2 degrees in two-theta in figure 6b). Two sigmas seemed thus an insufficient threshold for diffraction peak detection. Using 3σ, for the two considered pixels, only significant diffraction peaks were selected (see figure 6c and d). Moving to a 3.5σ threshold, small peaks considered as significant by an expert, as the peak at 11.7 degrees in 2θ on figure 6f, are not identified anymore by our procedure. At last, while considering 4σ, no peak is detected, in both cases (results not shown). We thus conclude that 3σ was the optimal threshold to be used for the peak detection.

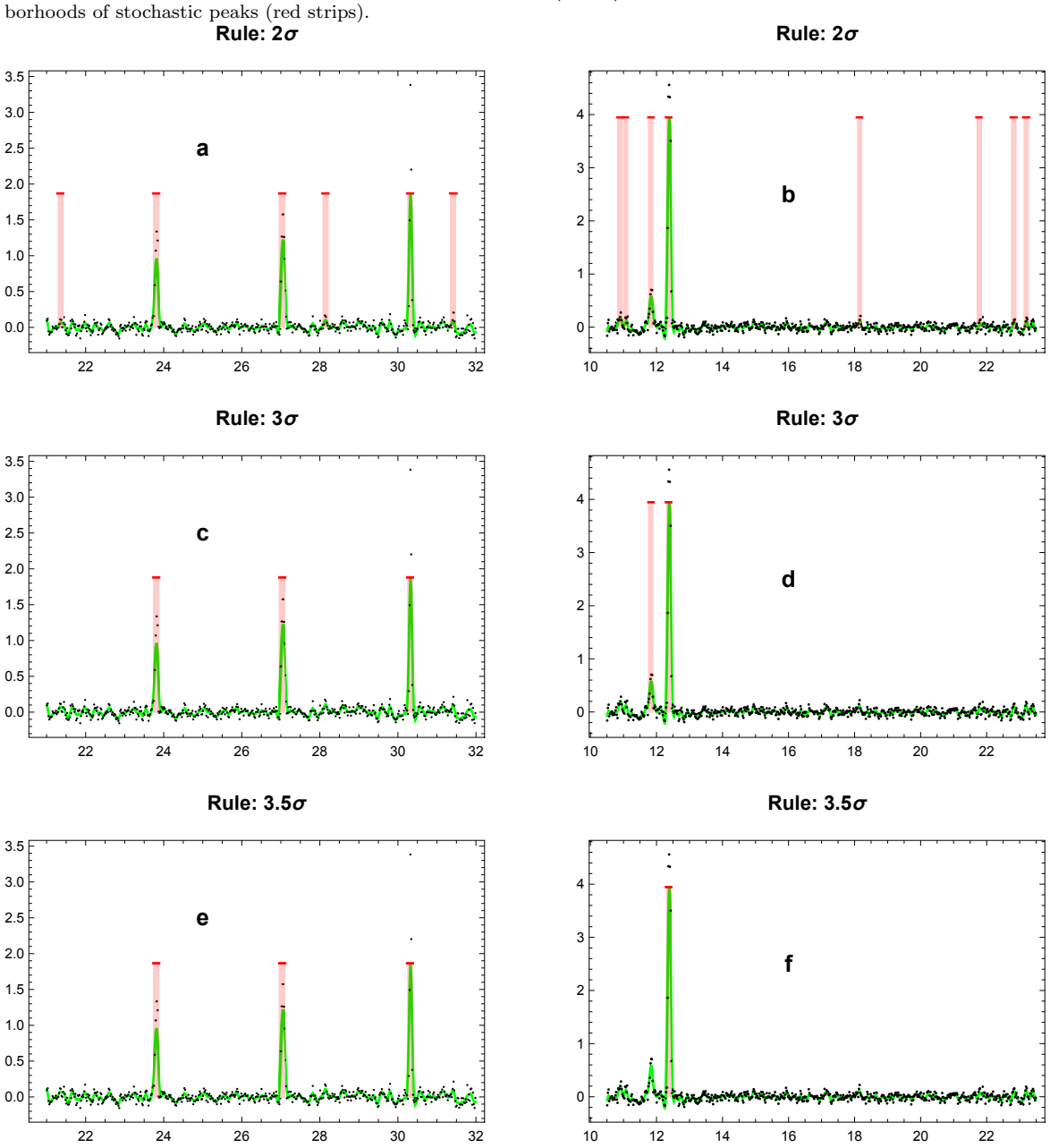
Fig. 5 Two examples of XRD-patterns. Raw data (dots) and fitting of the background by COBS of angular domains 1 (yellow), 2 (green) and 3 (red).



3.2 Identification of the mineral on the mean XRD-patterns of the different soil microscopic units for the three studied zones

Average XRD-patterns for the analyzed zones shows that the main notable fea-

Fig. 6 Peaks detection for two diffractograms (columns). On each panel are superimposed: the detrended data (black points) $\mathcal{D}$, the smoothing spline $\mathscr{S}\left(\mathcal{D}, \widehat{\sigma^{\mathrm{m}}}\right)$ (green) and the neighborhoods of stochastic peaks (red strips).



ture of these XRD-patterns is a bulge for 2θ values ranging from 7 to 17° that is characteristic of the glass slide on which the thin sections are mounted. Few small peaks are also observed. They may be considered either as artefacts due notably to

Fig. 7 Analysis of a dusty clay coating surrounding a pore from S/T: a- micromorphology picture; b- classification of the different considered zones (pore, coating, matrix) by image analysis; c to i: maps of the probability of presence the different minerals WPI of different minerals identified on l, the average XRD-patterns for the different considered zones, with c- feldspars; d- quartz, e- goethite; f- maghemite; g- kaolinite; h- illite; i- smectite. The green lines represent the borders between the different pedological features (pore, coating and matrix). Blue stands for the absence of the mineral when orange is its maximal probability of presence. Maps j and k are μXRF maps for Fe and Zn respectively.

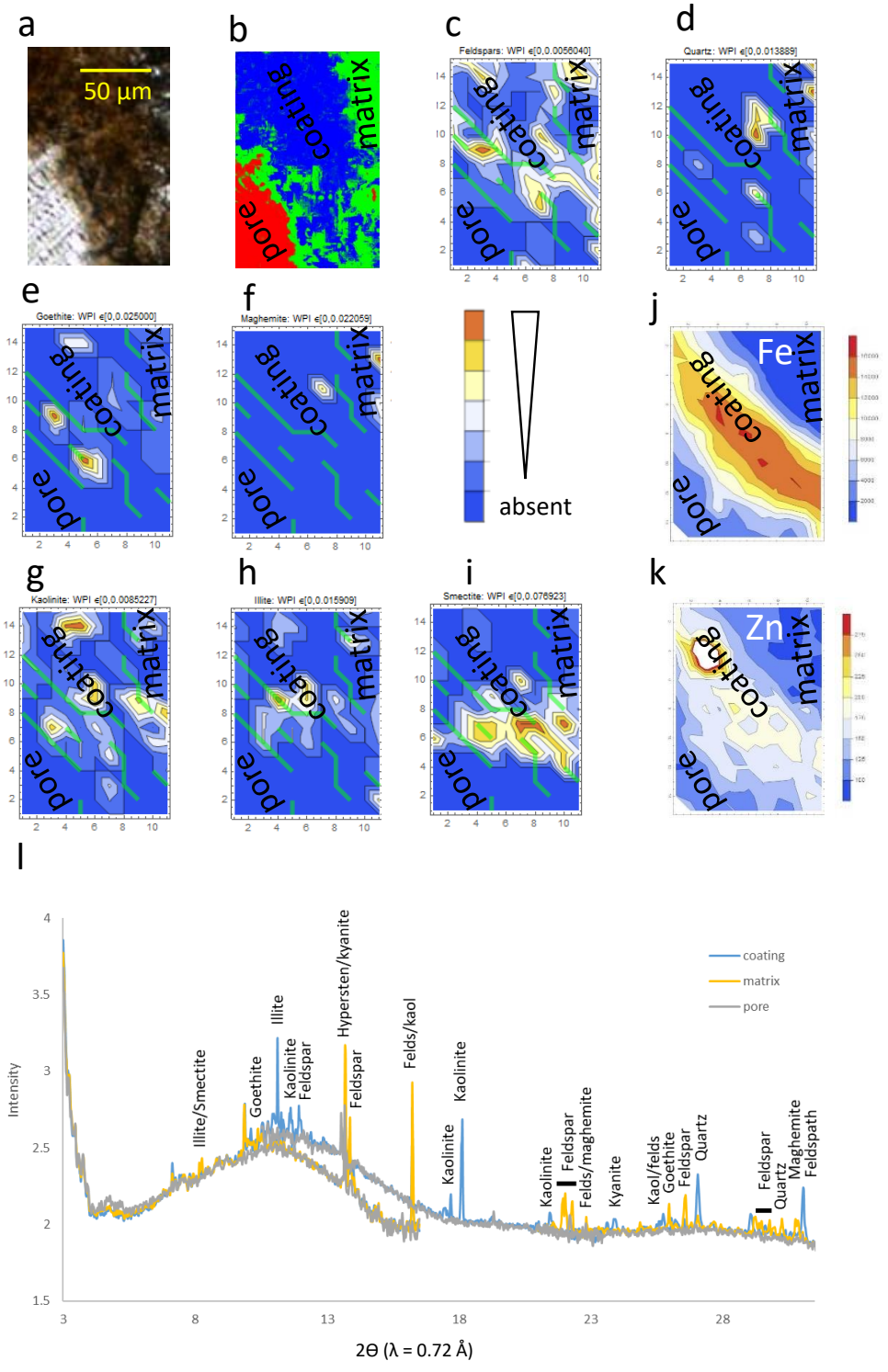


Fig. 8 Analysis of a limpid clay coating surrounding a pore from R/T-3: a- micromorphology picture; b- classification of the different considered zones (pore, coating, matrix) by image analysis; c to i: maps of the probality of presence of the different minerals identified on l, the average XRD-patterns for the different considered zones, with c- feldspars; d- quartz; e- goethite; f- maghemite; g- kaolinite; h- illite; i- smectite. The green lines represent the borders between the different pedological features (pores, coating and matrix). Blue stands for the absence of the mineral when orange is its maximal probability of presence. Maps j and k are μXRF maps for Fe and Zn respectively.

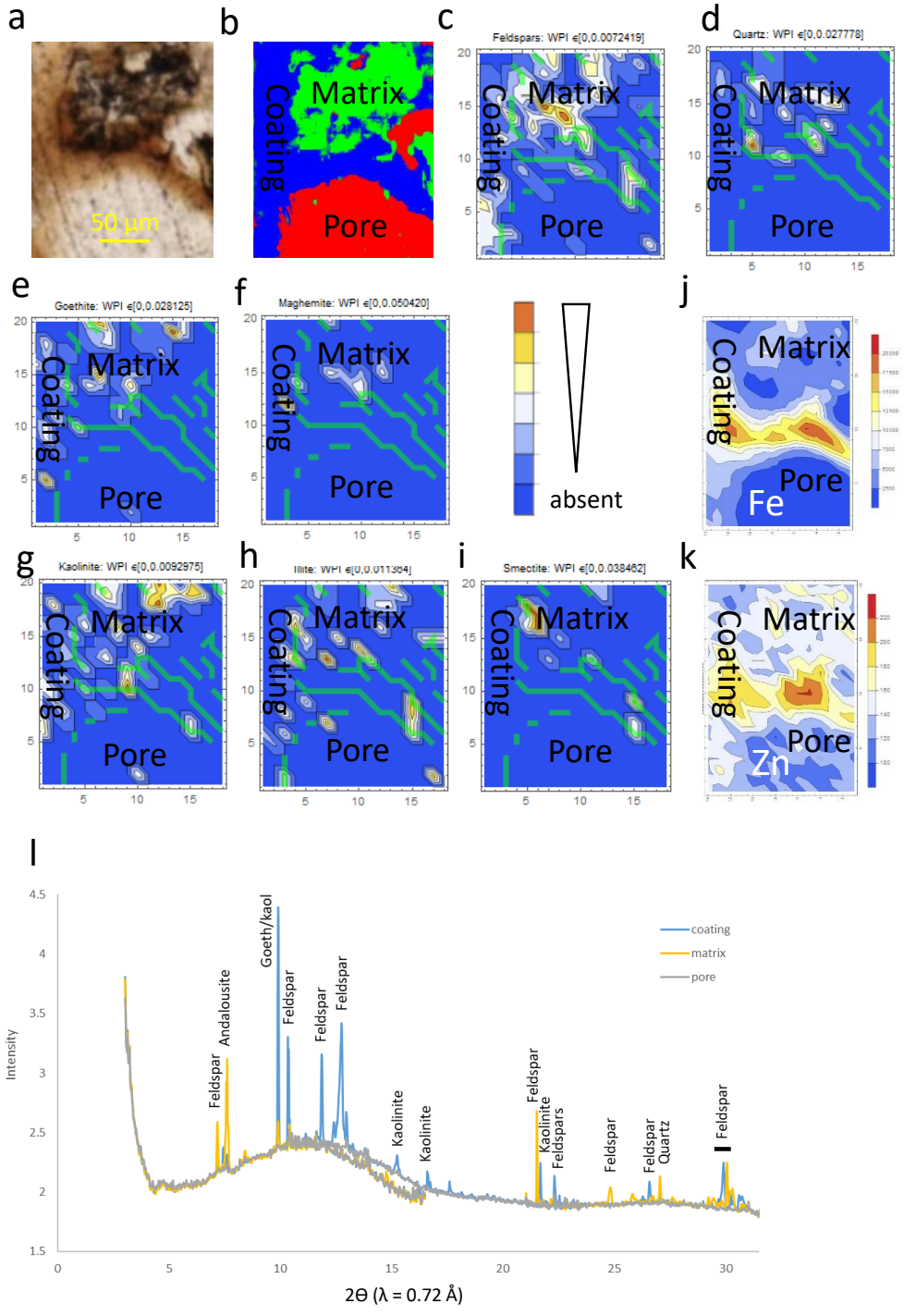
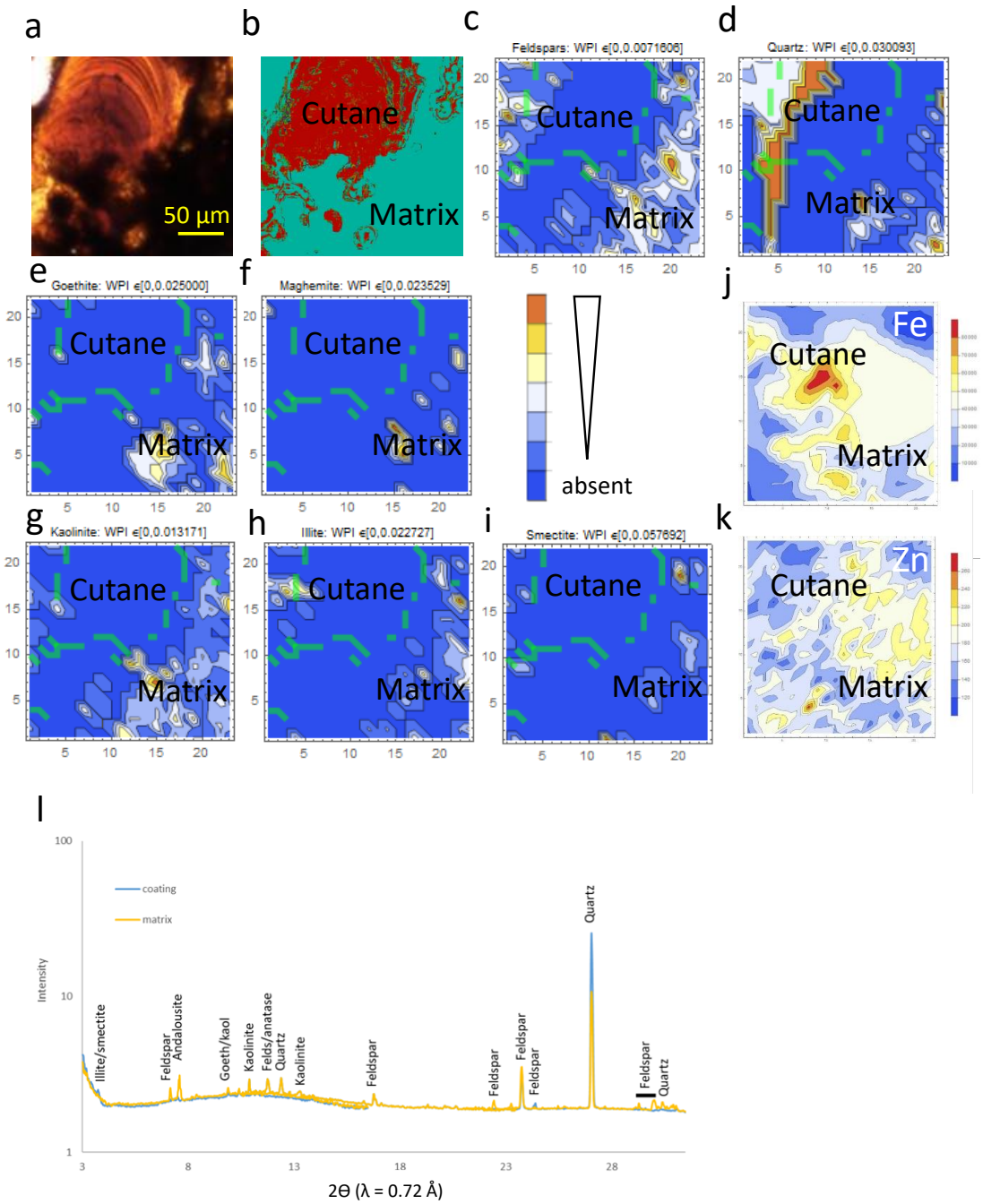


Fig. 9 Analysis of a laminated coating from AR/B: a- micromorphology picture; b- classification of the different considered zones (pore, coating, matrix) by image analysis; c to i: maps of the probability of presence of the different minerals identified on l, the average XRD-patterns for the different considered zones, with c- feldspars; d- quartz, e- goethite; f- maghemite; g- kaolinite; h- illite; i- smectite.The green lines represent the borders between the different pedological features (pores, coating and matrix). Blue stands for the absence of the mineral when orange is its maximal probability of presence. Maps j and k are μXRF maps for Fe and Zn respectively.



Feature type	Studied zone	Kaolinite	Goethite	Illite	Feldspars	Quartz	Smectite	Tot. nb. of obs. peaks
Nb. of peaks observed exp.	All	52(30)	14(7)	17(11)	150(119)	12(7)	6(1)	
Limpid coating	R/T-1	1 (0)	0	0	7 (0)	0	0	7
	R/T-2	16(9)	2(0)	4(2)	22(15)	7(5)	0	42
	R/T-3	3(3)	1(1)	1(0)	15(12)	3(3)	0	22
	S/B	7(4)	3(1)	9 (7)	30(23)	3 (3)	3(1)	30
Laminated coating	AR/B	3(2)	1(0)	2 (1)	8(8)	1(1)	1 (0)	14
	P/B	5(2)	0	0	7(4)	0	0	9
Dusty coating	AR/T	6 (5)	3 (2)	0	11(11)	1 (1)	1 (0)	22
	P/T	10(4)	2(1)	3(2)	23(18)	2(2)	1(0)	31
	S/T	11(7)	2(1)	2(1)	18(13)	1(1)	2(0)	31
Matrix	AR/B	8 (3)	3 (1)	0	15 (13)	3 (3)	0	30
	AR/T	10 (8)	3 (2)	0	14 (12)	4 (4)	1 (0)	30
	P/B	9(5)	1(1)	1(1)	12(8)	1(1)	0	20
	P/T	5 (2)	3 (2)	1 (1)	30 (28)	3(3)	0	41
	R/T-1	8(3)	2(0)	2(0)	29(21)	4(3)	0	36
	R/T-3	6(3)	2(0)	3(3)	18(17)	2(2)	0	31
	S/B	10(7)	4(1)	1(1)	35(28)	4 (3)	1 (0)	53
	S/T	6(2)	4(3)	3(1)	20(13)	2(2)	3(1)	33
Pore	P/T	0	0	1(1)	8 (8)	2 (2)	0	11
	R/T-1	0	0	0	4(2)	0	0	4
	R/T-2	2(2)	2(1)	3(2)	3(2)	1(0)	0	9
	R/T-3	0	0	0	3(3)	0	0	3
	S/B	3(2)	1(0)	1(1)	5(4)	2(1)	0	11
	S/T	1(1)	0	0	1(1)	0	0	4

Table 1 Number of peaks observed experimentally for each mineral on all the studied thin sections in the angular domain 3.5 - 33.4 degrees in 2θ ($\lambda = 0.72\mathring{A}$) and number of peaks observed experimentally for each mineral for the different zones. Under bracket, the number peaks experimentally detected and not shared among several minerals (called thereafter exclusive peaks).

an underestimation of the noise in certain zones, or resulting from an overestimation of the pore surface by the image analysis including grains located at the edges of these areas. The mineral maps of the two zones including pores (Figures 7 and 8 c to e) also show that almost no minerals are detected in the zones considered as pores with some exceptions of very small intensity. For other zones however, more peaks were detected due to an overestimation by image analysis of the pore zone (see pore zones of P/T, S/B and R/T-2 in Table 1). The total number of peaks detected remains however largely lower than in the matrix or in the clay coatings (Table 1).

The average XRD-patterns of the matrix always exhibit a large number of peaks, many of them being associated to feldspars and quartz (Figures 7j to 9j and Table 1). Kaolinite, goethite are also systematically observed on the average XRD-patterns of the matrix while illite is only encountered in the matrix of some of the analyzed zones. Peaks of feldspar are the most numerous with the highest intensities; this might be due to the fact that feldspar is a mineral family with a high amount of possible diffraction peaks (150, see Table 1), it is also a mineral common in the studied soil. Surprisingly, quartz that is known to represent about 80 percents in mass of these soils was not dominating the XRD-signal, probably due to the much lower number of possible diffraction peaks in the angular domain of 3.5 to 33.4 in 2θ (Table 1) and the difficulty in detecting the crystallite size of the quartz with the small size beam used for mapping with X-ray synchrotron

as already observed by others (Stroh et al. (2014); Siebecker et al. (2018), Supplementary materials 2 and 3). Smectite is rarely detected without ambiguity, as expected for the matrix of this soil chosen to be depleted in smectite.
In the coating zones, the average XRD-patterns show a comparable number of diffraction peaks to that observe for the matrix with the notable exception of R/T-1 (Table 1) that is very thin with probably a small amount of matter on the section according to the optic microscopy observations (data not shown). As for the matrix, feldspars and kaolinite were systematically detected in the coatings whereas goethite, illite and quartz are observed in most of them. Finally, smectite is rarely unambiguously detected (Table 1). No clear distinction is observed between the mineral composition of the two types of coatings while dusty coatings (DSB) were thought to be richer in grains (feldspars and quartz, Fedoroff and Courty (1994); Jongmans et al. (2001); Kuhn et al. (2010)). In addition, while it was expected to identify more clay minerals in coatings than in the matrix, this was not clearly observed (Table 1). These two surprising results may be due to changes in the relative abundance of the mineral species but not in the presence/absence of specific minerals and possibly to imprecisions in the segmentation of images MUs. In the case of the laminated coating (HLB), only a few peaks were identified (Table 1 and Figure 9j) despite the clear evidence of a large amount of matter on the considered zones of the thin sections (see Figure 8a). The number of peaks associated to feldspars is for example closed to those observed in pores and considerably lower than those of the soil matrix (Table 1) suggesting that the diversity and very likely the abundance of feldspar is limited in the HLB. Quartz shows a similar behaviour (Table 1). Contrastingly, peaks associated with clay minerals and more particularly with kaolinite, are more systematically detected than in pores (Table 1). Smectite, that has been only very rarely detected, seems entering in the mineralogical composition of the AR/B HLB. These combined results suggest that HLB are very likely enriched in various clay minerals and impoverished in quartz and feldspar grains as expected. It is however clear that the total number of peaks detected in the HLB remains surprisingly low. This is interpreted as resulting from the marked orientation of the clay minerals in HLBs as deduced from its high level of birefringence (Figure 9a). Therefore, due to their orientated nature, it seems very difficult to characterize with certainty the nature of the clay minerals contained in HLBs, while the same minerals being randomly orientated when present in the matrix may be identified with the same technic (Schulze and Bertsch, 1995; Lopano et al., 2011; Yusiharni and Gilkes, 2012; Tsao et al., 2013; Sumoondur et al., 2008).

3.3 Spatial distribution of the XRD-pattern in the different soil microscopic unit (MU) of the three studied zones

The μXRD-maps confirm reasonably well the analysis obtained from the different MU average XRD-patterns: the pore, the matrix and the coatings (Figures 7 to 9 c to i). Pores are almost free of mineral (Figures 7 and 8 c to i) as already observed from average XRD-patterns and in good coherence with the very low Fe and Zn concentrations (Figures 7 and 8 j and k) used here as geochemical tracers for irons oxides and for the clay minerals respectively as already demonstrated in comparable studies (Cornu et al., 2007). The matrix is characterized by the

highest abundances in feldspar and quartz (Figures 7 to 9 c and d) and by a patchy distribution of clay minerals like kaolinite and illite and of iron oxides, especially maghemite (Figure 7 to 9 e to i). The coatings, unsurprisingly rich in clay minerals as suggested by high Zn concentrations (Figures 7 to 8k), are characterized by i) the presence of kaolinite, illite and of most of the detected smectite in the HLB (Figures 7 to 9 g, h and i); ii) almost no feldspar nor quartz but in the dusty coating (DSB) or at the borders of the limpid (LB) and the HLB coatings (Figures 7 to 9 c and d) and iii) few goethite and almost no maghemite (Figures 7 to 9 e, f) despite very high Fe contents (Figure 7 to 9j). The DSB is characterized by the identification of quartz and feldspars grains, which is in good agreement with microscopic observations made for this type of coatings (see section 2.2.) and with their conditions of formation (Fedoroff and Courty, 1994; Jongmans et al., 2001; Kuhn et al., 2010). Smectite is in addition more frequently detected in this coating than in LBs (Figures 7 and 8i). Since DSBs are supposed to developed after rains on bare soils (Fedoroff and Courty, 1994; Jongmans et al., 2001; Kuhn et al., 2010), which correspond to the conditions of the rain experiment performed, and since this coating contains smectite that are absent from the surrounding matrix, it could have been formed by eluviation of material coming from the upper core that contain smectite during the rain experiment. More systematic analysis of this type of coating should be performed in order to definitely conclude on that point. Finally, the ability of μXRD to detect clay minerals in clay coatings decreases from DSBs to LBs and finally to HLBs despite similarly high Zn contents (Figure 7 to 9k). It is very likely related to an increasing orientation of clay particles as shown by increasing birefringence and resulting from conditions of formation characterized by slower water flows. the μXRD maps were relatively successful in identifying changes in the mineralogical composition not only between the soil matrix and soil features like but also between DSB and LB coatings. On the contrary, only few Fe-bearing minerals was identified in the different clay coatings despite homogenously high Fe concentrations in clay coatings and despite the identification of patches of goethite or maghemite in the soil matrix for lower Fe concentrations (Figures 7 to 9j). This apparent contradiction results from that fact that in this type of soil, a large part of Fe is mainly present as poorly crystalline oxides as shown by sequential extractions by Montagne et al. (2008) and that maghemite is more of an inherited mineral present in parent loess material and thus in the matrix.

4 Conclusion

Synchrotron μXRD on thin sections of soils resulted in mineralogical maps of high complexity due to the small size of the analysed pixel compared to the size of the mineral grains and the orientated nature of the clay coatings that partially extinguish the diffraction peaks of the clay minerals. It leads to the detection of a few diffraction peaks in the most favourable case. Despite this complexity, our approach, combining various signal processing methods (wavelets, quantile regression, smoothing splines, statistical inference), was shown to be efficient and robust enough to automatically identify and map minerals from synchrotron XRD large datasets obtained on naturally highly heterogeneous soil materials. Most of the minerals present in the considered samples could be identified, including clay minerals with the notable exception of smectite that was rarely identified with cer-

tainty. μXRD mapping was found to be complimentary to average XRD-patterns and particularly helpful to characterize changes in mineralogical assemblages. Although mostly qualitative, the approach proposed here successfully differentiated the mineralogical composition of the soil matrix from that of soil features like clay coatings. It results in one of the first characterization of the spatial variability of clay mineral assemblages in clay coatings. In the studied case, the clay coatings were found to contain not only smectite as generally hypothesized but also kaolinite and illite. It suggests that the clay translocation process, if selective of clay minerals species, is not selective enough to produce monospecific clay coatings. The development of quantitative approach thus seems necessary to definitively state on the potential clay mineral selectivity of the translocation process. The mineralogical composition of clay coatings was found to be sensitive to their conditions of formation as shown by dusty coatings richer in feldspar, in quartz or in smectite than the limpid ones.
Moreover, the recently available fast data acquisition schemes, yielding now large amounts of datasets (e.g. several $10^4 - 10^6$ diffractograms), requires automatic methods dedicated to the automatic identification and mapping of minerals as the one developed in this study.

5 Acknowledgments

This research was conducted in the framework of the Agriped project (ANR-10-BLANC-605) supported by the French National Research Agency (ANR). The authors are grateful to the SOLEIL synchrotron for experimental time allocation and to C. Le Lay for thin section fabrication.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Antoniadis A, Lambert-Lacroix S, Letué F, Bigot J (2007) Nonparametric preprocessing methods and inference tools for analyzing time-of-flight mass spectrometry data. Current Analytical Chemistry 3, 2:127–147

Antoniadis A, Bigot J, Lambert-Lacroix S (2010) Peaks detection and alignment for mass spectrometry data. Journal de la Société Francaise de Statistique 151, 1:17–37

Bruand A, Cousin I, Nicoullaud B, Duval O, Bégon J (1996) Backscattered electron scanning images of soil porosity for analysing soil compaction around roots. Soil Sci Soc Am J 60:895–901

Collective (2007) Appendix a. 95% cl and 5σ discoveries. J Phys G: Nucl Part Phys 2, Physics Performance:1485–1490

Coombes K, Fritsche H, Clarke C, Chen Jn, Baggerly K, Morris J, Xiao Lc, Hung MC, Kuerer H (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clinical Chemistry 49, 10:1615–1623

Cornu S, Montagne D, Maguin F, Lay C, Chevallier P, Cousin I (2007) Influence of human impacts on Albeluvisol analysed by X-ray microfluorescence: Relative evolution of the transforming front at the tongue scale. The Science of the total environment 377:244–54, DOI 10.1016/j.scitotenv.2007.01.083

Cornu S, Quénard L, Cousin I, Samouëlian A (2014) Experimental approach of lessivage: quantification and mechanisms. Geoderma 213:357–370

Dalrymple J, Theocharopoulos S (1987) Intrapedal cutans — lateral differences in their properties and their spatial clustering. Geoderma 41(1):149 – 180, DOI https://doi.org/10.1016/0016-7061(87)90033-4

De Nolf W, Jaroszewicz J, Terzano R, Lind OC, Salbu B, Vekemans B, Janssens K, Falkenberg G (2009) Possibilities and limitations of synchrotron X-ray powder diffraction with double crystal and double multilayer monochromators for microscopic speciation studies. Spectrochimica Acta Part B: Atomic Spectroscopy 64(8):775–781, DOI 10.1016/j.sab.2009.06.003, URL http://www.sciencedirect.com/science/article/pii/S0584854709001426

Dillmann P, Neff D, Mazaudier F, Hoerle S, Chevallier P, Beranger G (2002) Characterisation of iron archaeological analogues using micro diffraction under synchrotron radiation. Application to the study of long term corrosion behaviour of low alloy steels. J Phys IV France 12(6):393–408, DOI 10.1051/jp4:20020248, URL http://dx.doi.org/10.1051/jp4:20020248, publisher: EDP Sciences

Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90:1200–1224

Fan R, Gerson AR (2011) Nickel geochemistry of a Philippine laterite examined by bulk and microprobe synchrotron analyses. Geochimica et Cosmochimica Acta 75(21):6400–6415, DOI 10.1016/j.gca.2011.08.003, URL http://www.sciencedirect.com/science/article/pii/S0016703711004595

Fedoroff N, Courty M (1994) Organisation du sol aux échelles microcopiques. In: Bonneau M, Souchier B (eds) Constituants et propriétés du sol., Masson, Paris, pp 349–375

Gaast SJVD, Vaars AJ (1981) A method to eliminate the background in X-ray diffraction patterns of oriented clay mineral samples. Clay Minerals 16(4):383–393, publisher: Cambridge University Press

Gueriau P, Réguer S, Leclercq N, Cupello C, Brito PM, Jauvion C, Morel S, Charbonnier S, Thiaudière D, Mocuta C (2020) Visualizing mineralization processes and fossil anatomy using synchronous synchrotron X-ray fluorescence and X-ray diffraction mapping. preprint, PaleorXiv, DOI 10.31233/osf.io/vw68e, URL https://osf.io/vw68e

Gunal H, Ransom M (2006) Clay illuviation and calcium carbonate accumulation along a precipitation gradient in kansas. Catena 68:59–69, DOI 10.1016/j.catena.2006.04.027

Gutiérrez-Castorena M, Ortiz-Solorio C, Sánchez-Guzmán P (2007) Clay coatings formation in tepetates from texcoco, mexico. Catena 71:411–424, DOI 10.1016/j.catena.2007.03.017

Jamagne M (1973) Contribution à l'étude pédologique des formations loessiques du Nord de la France. PhD thesis, Faculté des sciences agronomiques de l'état, Gembloux, Belgique

Jongmans A, Pulleman M, Marinissen J (2001) Soil structure and earthworm activity in a marine silt loam under pasture versus arable land. Biology and Fertility of Soils 33(4):279–285, DOI 10.1007/s003740000318

Kuhn P, Aguilar J, Miedema R (2010) Textural pedofeatures and related horizons. In: Stoops G, Marcelino V, Mees F (eds) Interpretation of micromorphological features of soils and regoliths, Elsevier, Amsterdam, pp 217–250

Lopano CL, Heaney PJ, Bandstra JZ, Post JE, Brantley SL (2011) Kinetic analysis of cation exchange in birnessite using time-resolved synchrotron X-ray diffraction. Geochimica et Cosmochimica Acta 75(14):3973–3981, DOI 10.1016/j.gca.2011.04.021

Mallat S (1999) A wavelet tour of Signal Processing. Academic Press

Manceau A, Tommaseo C, Rihs S, Geoffroy N, Chateigner D, Schlegel M, Tisserand D, Marcus MA, Tamura N, Chen ZS (2005) Natural speciation of Mn, Ni, and Zn at the micrometer scale in a clayey paddy soil using X-ray fluorescence, absorption, and diffraction. Geochimica et Cosmochimica Acta 69(16):4007–4034, DOI 10.1016/j.gca.2005.03.018, URL http://www.sciencedirect.com/science/article/pii/S0016703705002528

McKeague JA, Guertin RK, Valentine KWG, Belisle J, Bourbeau GA, Howell A, Michalyna W, Hopkins L, Page F, Bresson LM (1980) Note: estimating illuvial clay in soils by micromorphology. Soil Science 129(6):386

Mocuta C, Richard MI, Fouet J, Stanescu S, Barbier A, Guichet C, Thomas O, Hustache S, Zozulya A, Thiaudière D (2013) Fast pole figure acquisition using area detectors at the DiffAbs beamline – Synchrotron SOLEIL. Erratum. Journal of Applied Crystallography 46, DOI 10.1107/S0021889813027453

Montagne D, Cornu S (2009) Do we need to include soil evolution module in models for prediction of future climate change? Climatic Change 98(1):75, DOI 10.1007/s10584-009-9666-3, URL https://doi.org/10.1007/s10584-009-9666-3

Montagne D, Yahiaoui M, Cousin I, Le Forestier L, Cornu S (2007) Quantification of soil volumes in the eg & bt-horizon of an albeluvisol using image analysis. Can J Soil Sci 87:51–59

Montagne D, Cornu S, Forestier LL, Hardy M, Josière O, Caner L, Cousin I (2008) Impact of drainage on soil-forming mechanisms in a French Albeluvisol: Input of mineralogical data in mass-balance modelling. Geoderma 145:426–438, DOI 10.1016/j.geoderma.2008.02.005, URL https://hal-insu.archives-ouvertes.fr/insu-00286941

Montagne D, Cousin I, Josière O, Cornu S (2013) Agricultural drainage-induced Albeluvisol evolution: A source of deterministic chaos. Geoderma s 193–194:109–116, DOI 10.1016/j.geoderma.2012.10.019

Montagne D, Cousin I, Cornu S (2016) Changes in the pathway and the intensity of albic material genesis: Role of agricultural practices. Geoderma 268:156–164, DOI 10.1016/j.geoderma.2016.01.019

Ndzana GM, Huang L, Zhang Z, Zhu J, Liu F, Bhattacharyya R (2019) The transformation of clay minerals in the particle size fractions of two soils from different latitude in China. CATENA 175:317–328, DOI 10.1016/j.catena.2018.12.026, URL http://www.sciencedirect.com/science/article/pii/S0341816218305629

Ng P, Maechler M (2007) A fast and efficient implementation of qualitatively constrained quantile smoothing splines. Statistical Modelling 7, 4:315–328

Protz R, VandenBygaart A (1998) Towards systematic image analysis in the study of soil micromorphology. Sciences of Soils 3(1):34–44, DOI 10.1007/s10112-998-0004-0

Pukelsheim F (1994) The three sigma rule. The American Statistician 48, 2:88–91

Renyi A (1966) Calcul des probabilités. Dunod

Sauzet O, Cammas C, Barbillon P, Étienne M, Montagne D (2016) Illuviation intensity and land use change: quantification via micromorphological analysis. Geoderma 266:46–57

Sauzet O, Cammas C, Gilliot J, Bajard M, Montagne D (2017) Development of a novel image analysis procedure to quantify biological porosity and illuvial clay in large soil thin sections. Geoderma 292:135 – 148

Schulze DG, Bertsch PM (1995) Synchrotron X-Ray Techniques in Soil, Plant, and Environmental Research. In: Sparks DL (ed) Advances in Agronomy, Academic Press, pp 1–66, DOI 10.1016/S0065-2113(08)60537-4

Siebecker M, Chaney R, Sparks D (2018) Natural speciation of nickel at the micrometer scale in serpentine (ultramafic) topsoils using microfocused X-ray fluorescence, diffraction, and absorption. Geochemical Transactions 19, DOI 10.1186/s12932-018-0059-2

Stroh J, Schlegel MC, Irassar EF, Meng B, Emmerling F (2014) Applying high resolution SyXRD analysis on sulfate attacked concrete field samples. Cement and Concrete Research 66:19 – 26, DOI https://doi.org/10.1016/j.cemconres.2014.07.015, URL http://www.sciencedirect.com/science/article/pii/S0008884614001550

Sumoondur A, Shaw S, Ahmed I, Benning LG (2008) Green rust as a precursor for magnetite: an in situ synchrotron based study. Mineralogical Magazine 72(1):201–204, DOI 10.1180/minmag.2008.072.1.201

Sutton SR, Lanzirotti A, Newville MG, Rivers ML, Eng PJ, Lefticariu L (2017) Spatially Resolved Elemental Analysis, Spectroscopy and Diffraction at the GSECARS Sector at the Advanced Photon Source. Journal of environmental quality DOI 10.2134/jeq2016.10.0401

Terribile F, FitzPatrick EA (1992) The application of multilayer digital image processing techniques to the description of soil thin sections. Geoderma 55(1):159 – 174, DOI https://doi.org/10.1016/0016-7061(92)90011-U

Thompson ML, Fedoroff N, Fournier B (1990) Morphological features related to agriculture and faunal activity in three loess-derived soils in France. Geoderma 46(4):329 – 349, DOI https://doi.org/10.1016/0016-7061(90)90023-3

Torres-Sallan G, Schulte RPO, Lanigan GJ, Byrne KA, Reidy B, Simó I, Six J, Creamer RE (2017) Clay illuviation provides a long-term sink for C sequestration in subsoils. Sci Rep 7(1):1–7, DOI 10.1038/srep45635, URL https://www.nature.com/articles/srep45635, number: 1 Publisher: Nature Publishing Group

Tsao T, Chen Y, Sheu H, Tzou Y, Chou Y, Wang M (2013) Separation and identification of soil nanoparticles by conventional and synchrotron X-ray diffraction. Applied Clay Science 85:1–7, DOI 10.1016/j.clay.2013.09.005

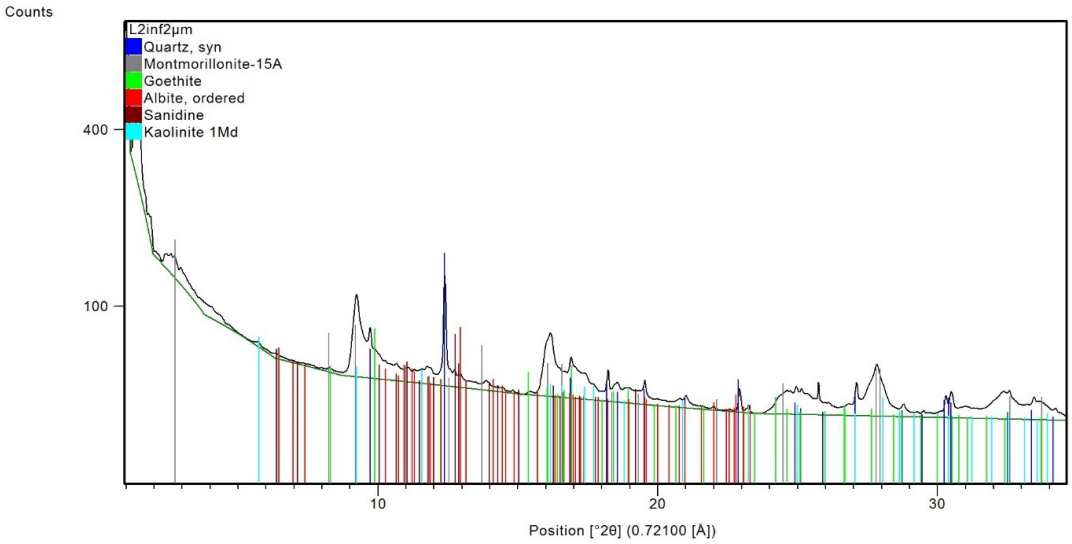
Van Vliet-Lanoë B (1990) The genesis and age of the argillic horizon in Weichselian loess of northwestern Europe. Quaternary International 5:49–56, DOI 10.1016/1040-6182(90)90024-X, URL http://www.sciencedirect.com/science/article/pii/104061829090024X

Wahba G (1990) Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, DOI 10.1137/1.9781611970128

Wang Y, Ke C (2004) Assist: a suite of S functions implementing spline smoothing techniques. University of California, Santa Barbara pp 1–127

Yusiharni E, Gilkes RJ (2012) Changes in the mineralogy and chemistry of a lateritic soil due to a bushfire at Wundowie, Darling Range, Western Australia. Geoderma 191:140–150, DOI 10.1016/j.geoderma.2012.01.030

Fig. 10 Synchrotron XRD-pattern of the $< 2\,\mu m$ fraction (in powder) of the L2 core (acquisition time: 100s)



Appendices

Notations and special terms

In this part, $\#E$ will denote the number of elements of some finite set E. We will say that a matrix is of size $r \times c$ if it has r rows and c columns, and the Hadamard (i.e. elementwise) product of two vectors of dimension p will be denoted: $u \circ v := (u_1 v_1, \cdots, u_p v_p)$. As usual, $\mathbb{N}$ will denote the set of positive integer numbers, $\mathbb{Q}$ the set of rational numbers and $\mathbb{R}$ the set of real numbers, and the logical connectors ***and*** (resp. ***or***) will be denoted $\wedge$ (resp. $\vee$).

A Methodological test on the $< 2\,\mu m$ fraction of the considered soil.

The fraction $< 2\,\mu m$ of the soil was analysed in powder mode by synchrotron XRD in the same configuration than the studied thin section in order to test whether the mineral of interest, notably the clay minerals could be detected in the configuration used. Two main differences with respect to the thin section analysis are to be noted: a) longer XRD acquisition time (100 s) and b) during the acquisition, the sample is continuously moved laterally in Y and Z direction over a $\pm 500\,\mu m$ range. As a consequence, more crystallites are potentially brought into diffraction condition, i.e. the sample resembles more to a random powder (i.e. with a sufficient statistic for all possible orientation that is not obtained at the pixel size even for $< 2\,\mu m$ randomly orientated particles). This explains the much more ring-like XRD features on the XPAD images (see e.g. Figure 2). Clay mineral peaks are detected, however the 001 peaks, traditionally used to identify clays, are very small despite their important concentration in this fraction. Quartz is also clearly detected despite its small amount in this $< 2\,\mu m$ fraction. Small peaks of feldspars and goethite are also observed. This first analysis demonstrated the feasibility of our approach when randomly orientated particles are considered. In the case of thin sections, very few peaks are observed because of the discrete orientation of the crystals

in the illuminated sample volume (their volume is comparable with the illuminated sample volume).

B Background removal in the case of the first angular domain

Because of the steepness of all these patterns near the origin (see figure 5), it was impossible to obtain a satisfying fit with median regression in a standard way. Since COBS Ng and Maechler (2007) is very flexible, we could impose global shape conditions to $\widehat{L_i}$ (associated with the i^{th} pixel), as well as pointwise equality constraints: the solution must run through a local maximum of the diffractogram (among its 10 first values) and must have a "right" slope at the midpoint angle. More precisely, we considered the beginning $\mathcal{D}_{beg} := \{(a_1,\, d_1), \cdots, (a_{81},\, d_{81})\}$ of the diffractogram, and determined the angle a_{Max} such that $d_{Max} = \underset{1 \leq i \leq 10}{Max}\, d_i$. We first imposed to the output L of COBS the interpolation constraint:

$$L\left(a_{Max}\right) = d_{Max}. \tag{3}$$

Then, we considered the midpoint angle a_{41} of $\mathcal{D}_{beg}$ and the nonlinear fit of $\mathcal{D}_{beg}$ by

$$f\left(a; \lambda, \alpha, \beta, \gamma\right) := e^{-\lambda\, a}\left(\gamma + \alpha\, a + \beta\, a^2\right)$$

obtained for some optimal vector $\left(\tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}\right)$. We then imposed to L, in addition to constraint (3), the slope constraint:

$$\frac{d}{da} L\left(a_{41}\right) = \frac{d}{da} f\left(a_{41}; \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}\right). \tag{4}$$

C Estimation of σ

We drew from each zone a random sample of N pixel positions (we fixed $N \approx 1/16$ of the zone size) and obtained the required estimate of σ thanks to to the following algorithm:

1. perform stationary wavelet denoising (Mallat, 1999, Section 10.2.3) on each detrended signal $\mathcal{D}_i$:
$$\mathcal{D}_i \mapsto \mathcal{D}_i - \widehat{\varepsilon_i}$$
(the *SureShrink* procedure of Donoho and Johnstone (1995) was used), and obtain the **local** estimate $\widehat{\sigma}\left(\widehat{\varepsilon_i}\right)$ of the SD
2. check the normality of the residual series $\widehat{\varepsilon_i}$
3. robustly estimate the SD by the trimmed mean of order 0.25 of the local estimates $\{\widehat{\sigma}\left(\widehat{\varepsilon_1}\right), \cdots, \widehat{\sigma}\left(\widehat{\varepsilon_N}\right)\}$; in addition, verify that the trimmed mean of order 0.25 of the local estimates of the mean, $\{\widehat{\mu}\left(\widehat{\varepsilon_1}\right), \cdots, \widehat{\mu}\left(\widehat{\varepsilon_N}\right)\}$, is close to zero.

A typical illustration of the normality of residuals has been shown on figure 3.

D Stochastic binarization of detrended patterns

Remember first the well-known Bienaymé-Tchebychev theorem Renyi (1966).

Theorem 1 *(Bienaymé-Tchebychev inequality) Let D be a random variable with mean μ and finite standard deviation $\sigma > 0$ and $r > 1$. Then*

$$P\left(|D - \mu| > r\, \sigma\right) < 1/r^2.$$

The three sigma rule Pukelsheim (1994) came up from a kin inequality, suited for unimodal distributions and proved by Gauss.

Theorem 2 *(Gauss inequality) Let D be an absolutely continuous random variable with mode ν and $\tau := \sqrt{E\left((D-\nu)^2\right)}$. Then*

$$P\left(|D-\nu| > r\right) \leq \begin{cases} \frac{4}{9}\frac{\tau^2}{r^2} \; if \; r \geq \tau\sqrt{4/3} \\ 1 - \frac{r}{\tau\sqrt{3}} \; if \; r \leq \tau\sqrt{4/3} \end{cases} \quad . \tag{5}$$

If D is a symmetric distribution, we obtain this way $P\left(|D-\mu| > 3\,\sigma\right) \leq 4/81 \approx 0.05$: this is the three sigma rule, attached to the usual excess probability 0.05. Notice that in this case the Bienaymé-Tchebychev inequality only tells us that $P\left(|D-\mu| > 3\,\sigma\right) < 1/9$: this is approximately twice the excess probability given by the Gauss inequality, which is consequently more accurate than the Bienaymé-Tchebychev one.

Remark 3 *Of course, the accuracy is even better in the parametric setting: suppose $D \sim \mathcal{N}(0,\sigma)$. Then $P\left(|D| > 2\,\sigma\right) \approx 0.0455$, $P\left(|D| > 3\,\sigma\right) \approx 0.003$ and $P\left(|D| > 4\,\sigma\right) \approx 6.33\,10^{-5}$.*

Consider now a detrended diffractogram $\mathcal{D} = \{(a_1,\, d_1), \cdots, (a_J,\, d_J)\}$, where $J \in \{648, 624, 528\}$ is the length of the series and a_j is an angle. The second coordinate d_j is a realization of some random variable $D\left(a_j\right)$; if a_j corresponds indeed to pure noise, we may agree (see Section C) that $D\left(a_j\right)$ is centered Gaussian, while in other cases (when a_j is associated with some genuine signal) $D\left(a_j\right)$ should obey an entirely different distribution. Thus, we naturally postulated that D was a contaminated centered Gaussian distribution (that's the rationale of trimmed estimations in Section C). We only supposed D is symmetric and unimodal, with zero as mode; in this case $|D|$ can be controlled by Formula (5). Suppose now a risk η_0 has been fixed; thanks to inequality (5), we can associate to $\mathcal{D}$ the **binarized diffractogram** $\mathcal{B} := \{(a_1,\, b_1), \cdots, (a_J,\, b_J)\}$ defined by:

$$b_j = \begin{cases} 0 \; if \; \left(d_j \leq 0\right) \vee P\left(|D| > d_j\right) > \eta_0 \\ 1 \; if \; \left(d_j > 0\right) \wedge P\left(|D| > d_j\right) \leq \eta_0 \end{cases} \quad . \tag{6}$$

In other words, $b_j = 1$ only if d_j is big enough. We fixed $\eta_0 = 0.0026998$, corresponding to the probability of deviation from the mean $> 3\,\sigma$ in the Gaussian case.

E The generating family $\mathcal{T}$ of intervals, the associated tables, and presence indicators

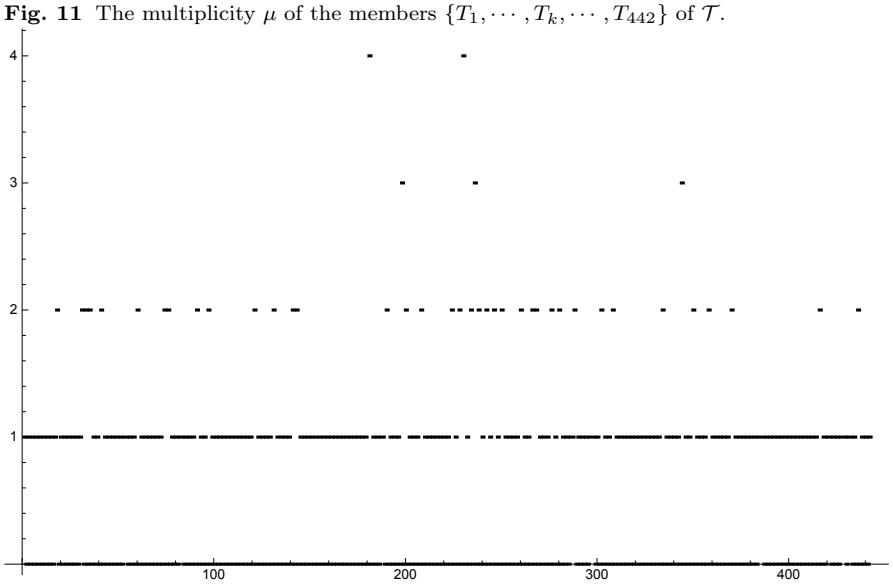
To each mineral of interest s was associated a whole original family $\left\{I_k^s : 1 \leq k \leq N^s\right\}$ of not necessarily disjointed intervals (resulting from expert knowledge) such that the angles associated with s belong to $E^s := \bigcup_{k=1}^{N^s} I_k^s$. Roughly speaking, each binarized diffractogram $\mathcal{B}$ of length J^m (associated with the m^{th} angular domain) defined in the previous subsection would give rise to S binary vectors, such that the one corresponding to s is $B^s = \left\{b_1^s, \cdots, b_{J^m}^s\right\}$, where $b_j^s = b_j$ if $a_j \in E^s$ and $b_j^s = 0$ if $a_j \notin E^s$. Thus, B^s measures the similarities between $\mathcal{B}$ and the "signature" E^s of the mineral s. But two problems are met:

- redundancy: because of the definition of the original intervals it is possible that for some j, $a_j \in I_k^s \bigcap I_{k+1}^s$; consequently, a single angle can "switch on" two intervals associated with a single s
- ambiguity: by construction, it is possible that $I_k^s \bigcap I_p^r \neq \emptyset$, or even $\left(I_k^s \bigcap I_{k+1}^s\right) \bigcap \left(I_p^r \bigcap I_{p+1}^r\right) \neq \emptyset$; consequently, a single angle can also switch on two (or more) intervals associated with different minerals.

Of course, these facts are not mutually exclusive. To control these problems, we built from the complete family of intervals $\left\{I_k^s,\, 1 \leq k \leq N^s,\, 1 \leq s \leq S\right\}$ another family, $\mathcal{T}_0$, of finer **disjoint generating** intervals, such that

$$\forall s,\, \forall k \;\; I_k^s = \coprod_{p \in P(s,k)} T_p$$

Fig. 11 The multiplicity μ of the members $\{T_1, \cdots, T_k, \cdots, T_{442}\}$ of $\mathcal{T}$.



(disjoint union), where $P(s,k)$ denotes the set of indices associated with the k^{th} original interval and the mineral s. We discarded afterward from $\mathcal{T}_0$ (538 intervals) all its "negligible" members, of width less than 0.02 degrees (angular resolution of the data), obtaining a family $\mathcal{T}$ of 442 disjoint intervals. Within this system, the redundancy phenomenon is eliminated, contrary to the ambiguity one. Nevertheless, notice that now an angle cannot switch on more than S intervals of $\mathcal{T}$.

E.1 The intervals of $\mathcal{T}$: their exclusiveness and exclusiveness

Consider again all the angles associated with the mineral s; they belong to $E^s = \bigcup_{k=1}^{N^s} I_k^s = \coprod_{p \in P(s)} T_p$, where $P(s)$ denotes the set of indices of members of $\mathcal{T}$ associated with s. We can thus represent some mineral s by its **binary reference spectrum** $\pi_s = \left(\pi_s^1, \cdots, \pi_s^{442}\right)$, where

$$\pi_s^k = \begin{cases} 1 \, if \, k \in P(s) \\ 0 \, if \, k \notin P(s) \end{cases}$$

and now, all the minerals are described in the common system $\mathcal{T}$! Let us now build from these binary spectra the matrix $\Pi := (\pi_1, \cdots, \pi_S)$ of size $442 \times S$, where S is the number of minerals of interest. Notice that Π only depends on the experimental apparatus and that $\mu := \Pi.1^S \in \mathbb{N}^{442}$ gives the number of minerals associated with each interval, its **multiplicity**. The multiplicity of the 442 members of $\mathcal{T}$ is displayed on figure 11. It can reach 4, while there are a number of intervals of multiplicity 1, which are really exclusive of some mineral. That is why we will define the **exclusiveness** of each interval by the inverse of its multiplicity, giving rise to the positive rational vector $\Omega := \left(\frac{1}{\mu(1)}, \cdots, \frac{1}{\mu(442)}\right)$.

E.2 The operator Λ^m associating $\mathcal{T}$ with angles of the m^{th} angular domain

Consider now the vector $A^m := \left(a_1^m, \cdots, a_{\#\tau}^m\right)$, of length $\#\mathcal{T} = 442$, of angles associated with this angular domain. Like Π, it is totally independent of the data, as well as the operator Λ^m of dimensions $J^m \times 442$ whose j^{th} row is the binary vector assigning to the sampled angle a_j^m the **unique** index $p(m,j) \leq \#\mathcal{T}$ such that $a_j^m \in T_{p(m,j)}$. We displayed on figure 12 the operator Λ^1 associated with the first angular domain. Notice the vertical patterns on this figure, which correspond to sequences of sampled angles belonging to a common interval of $\mathcal{T}$. Consider finally some vector B^i of binarized responses in the angular domain m (issued from the processing of some pixel, described in D), of length J^m. We will associate to B^i the vector of the spectra space $\Lambda^m\left(B^i\right) \in \mathbb{N}^{442}$ given by the matrix product $\Lambda^m\left(B^i\right) := B^i \cdot \Lambda^m$.

E.3 The presence indicators associated with minerals

The binarization procedure detailed in D gives rise, for each pixel and the m^{th} angular domain (with J^m angles), to some **detrended binary signal** $\mathcal{B}^m := \left\{\left(a_1^m,\, b_1^m\right), \cdots, \left(a_{J^m}^m,\, b_{J^m}^m\right)\right\}$. We then translate $\mathcal{B}^m$ into the $\mathcal{T}$ language, in order to evaluate the **possibility of presence** of each mineral at each pixel. Notice that the vector of first coordinates of $\mathcal{B}^m$ (angles), denoted A^m, only depends on the protocol while the vector of second coordinates, B^m, actually depends on the data. The operator $\Lambda^m \in J^m \times 442$ defined in E.2 associates to B^m the vector $\Lambda^m\left(B^m\right) := B^m.\Lambda^m \in \mathbb{N}^{\#\mathcal{T}} = \mathbb{N}^{442}$ whose p^{th} coordinate is the count of all the **significant angles** (i.e. a_j^m such that $b_j^m = 1$) of the detrended pattern which belong to the interval $T_p \in \mathcal{T}$ (the dot denotes the scalar product in the spectra space $\mathbb{R}^{\#\mathcal{T}} = \mathbb{R}^{442}$).

Balancing the information bore by intervals of $\mathcal{T}$

Since some XRD-peaks may be associated to several minerals while other may be exclusive of a single mineral, a ponderation was applied to give more weight to exclusive peaks than to peaks shared by several minerals. We defined an indicator assigning to each peak a weight decreasing with the number of minerals that shares it. Consider some mineral species s, whose **binary reference spectrum** is π_s, and the vector of **exclusiveness** Ω defined too in E.1; both these vectors belong to $\mathbb{Q}^{\#\mathcal{T}} = \mathbb{Q}^{442}$. While the "raw weight" of s is naturally $\overline{\pi_s} := \sum_{j=1}^{\#\mathcal{T}} \pi_s(j)$, we will define similarly its "exclusive weight" by $\omega_s := \sum_{j=1}^{\#\mathcal{T}} \left(\Omega \circ \pi_s\right)(j)$; of course, $\omega_s \leq \overline{\pi_s}$. Notice that these weights only depend on the matrices Π (defined in E.1) and $\{\Lambda^m,\, 1 \leq m \leq 3\}$ (defined in E.2), which are associated with the experimental setting, not with data.

Presence indices

Definition 3 *Considering the i^{th} "pixel" of a zone, we associate to the corresponding position $(x_i,\, y_i)$ the **Presence Index** of the mineral s at this position:*

$$PI\left(s; x_i,\, y_i\right) := \frac{1}{\overline{\pi_s}} \pi_s \cdot \Lambda^\bullet\left(B^\bullet\right)\left(x_i, y_i\right) \in \mathbb{Q} \tag{7}$$

where $\Lambda^\bullet\left(B^\bullet\right)\left(x_i, y_i\right) := \sum_{m=1}^{3} \Lambda^m\left(B^m\left(x_i, y_i\right)\right) \in \mathbb{N}^{442}$.

Definition 4 *Consider the i^{th} pixel of a zone and the vector of exclusiveness Ω. We associate to its position the **Weighted Presence Index** of s:*

$$WPI\left(s; x_i,\, y_i\right) := \frac{1}{\omega_s}\left(\Omega \circ \pi_s\right).\left(\Omega \circ \Lambda^\bullet\left(B^\bullet\right)\left(x_i, y_i\right)\right) \in \mathbb{Q} \tag{8}$$

with the same conventions as in Definition 3.

Fig. 12 The operator Λ^1; vertical patterns correspond to sampled angles belonging to a common interval of $\mathcal{T}$ (the arrows mark such cases).