Defining iron-bearing rock sources in compositional data statistical analysis for provenance studies: A case study from South Africa

Laure Dayet

Defining iron-bearing rock sources in compositional data statistical analysis for provenance studies: a case study from South Africa

Laure Dayet

Laboratoire EDYTEM, UMR5204 Environnements, DYnamiques et TErritoires de la Montagne, CNRS-Université Savoir Mont Blanc, Bâtiment « Pôle Montagne », 5 bd de la Mer Caspienne, F-73376 Le Bourget du Lac cedex

laure.dayet@gmail.com

Abstract

In iron-bearing rock provenance researches the definition of what is a source vary from an author to another. Issues might occur when the sources are associated with complex weathering profiles. How may the definition of the source influence the results of provenance studies when different types of iron-bearing rocks are encountered within the same weathering profile? In order to deal with this research question, we used a geochemical dataset previously obtained from the study of the ochre geological collection collected around Diepkloof rock shelter, South Africa. We tested the efficiency of Linear Discriminant Analyses (LDA, supervised analysis) and Principal Component Analysis (PCA, non-supervised analysis) in source discrimination and in the attribution of archaeological pieces to the right source. Different assumptions were made: sources were defined by separating or grouping the different lithologies (shale and ferricrete) encountered at the same location. The results show that grouping the different lithologies allows a good attribution of the archaeological pieces to the source it comes from when elements positively correlated with iron are used in the analysis. Separating the different lithologies do not allow a rate of success

in archaeological pieces attribution as high as when the lithologies are grouped together. This result might be explained by a highest number of samples that define a source when the different lithologies are grouped together. Further research should be considered in this direction in order to define if this is a general trend.

Key words: ochre – provenance – multivariate statistical analyses – weathering

Acknowledgments

We would like to thank Pierre-Jean Texier, Guillaume Porraz and the Diepkloof archaeological team for their help and support in the geological surveys and the archaeological input. We thank the Service d'Analyse des Roches et Minéraux, CRPG, Nancy, for the ICP analyses. We are grateful to the laboratory Archéosciences Bordeaux, specifically to Floréal Daniel and Pierre Guibert for their support in this project. We thank Amandine Blin for her advice on the statistical analysis and Emilie Chalmin for her precious comments on the draft. Diepkloof excavations were funded by the French Ministry of Foreign Affairs (MAE), the Aquitaine region, the Provence-Alpes-Côte-d'Azur region, and the Centre National de la Recherche Scientifique (CNRS). The ochre characterisation project was funded by the Région Aquitaine (Project title "Ocres préhistoriques"). The work of Laure Dayet was supported by the Centre National de la Recherche Scientifique and by the French Agence National de la Recherche (project number ANR-21-CE27-0028).

1. Introduction

In archaeology, the definition of the sources is a pre-requisite of any provenance researches of geological materials. Iron-bearing geological materials, also called 'ochre' (*sensu lato*, see Popelka-Filcoff & Zipkin, 2022) due to their wide use in various archaeological contexts, makes no exception. Most of the research done on the provenance (or more specifically on the provenience) of iron-bearing rocks are based on the analysis of major and trace elements (Attard Montalto et al., 2012; Dayet et al., 2016; Eiselt et al., 2011, 2019; Kiehn et al., 2007; MacDonald et al., 2011; Macdonald et al., 2013; MacDonald et al., 2018; Mathis et al., 2014; Mauran et al., 2021; McGrath et al., 2022; Pierce et al., 2020; Popelka-Filcoff et al., 2007, 2008, 2012; Scadding et al., 2015; Tarbska et al., 2008; Velliky et al., 2019, 2021; Zarzycka et al., 2019; Zipkin et al., 2015, 2017, 2020). The geochemical data are then used in statistical analyses aiming at discriminating the different sources. The definition of the source is of great importance. In ochre provenance researches, the geological source can be submitted to different definition. It can be an outcrop of a certain raw material (Dayet et al., 2016; Scadding et al., 2015), a locality (a GPS point) whatever the raw materials encountered (Eiselt et al., 2011; Popelka-Filcoff et al., 2007, 2008; Zipkin et al., 2015, 2017, 2020), a geological formation (Mathis et al., 2014), or geological region (Mauran et al., 2021). Sometimes, samples from different localities are grouped together to form larger sources (MacDonald et al., 2011; Macdonald et al., 2013; Velliky et al., 2019, 2021). In other terms, the geochemical signature of iron-bearing rocks varies at different scales.

It has been shown that samples from the same locality may not present the same geochemical trends (Dayet et al., 2016). That's why at the same location two different sources were defined in a previous provenance study (Dayet et al., 2016). However, iron-bearing rock outcrops might be characterized by complex weathering profiles where many different

lithologies can be observed at the same place, each corresponding to a different horizon of alteration. Making different sources for each type of lithology might become complex and may result in biases in the statistical analyses of the geochemical data. The aim of this study is to evaluate how changes in the definition of iron-bearing rock sources influences the results of multivariate statistical analyses. In order to do so, we reinvestigated the dataset of geochemical analyses of the geological outcrops sampled around the archaeological site Diepkloof rock shelter, South Africa (Dayet et al., 2013, 2016). We decided to separate or group together the different types / lithologies of iron-bearing rocks observed at the same locality, namely within the same weathering profile. The chosen context is ideal as two of the locations sampled and studied are simplified weathering profiles, which means that different horizons of alteration are present at the same place (same GPS point but different lithologies). We used Principal Component Analysis (PCA), an unsupervised learning method, and Linear Discriminant Analysis (LDA), a supervised learning method, by considering the different lithologies from the same place grouped together or separated. We evaluated the results obtained in source discrimination and archaeological pieces attribution.

2. Background

Iron-bearing rocks' provenance researches based on the analysis of geochemical data

Several methods can be used to obtain geochemical data, the two most common being neutron activation analyses (NAA) and inductively coupled plasma mass spectrometry coupled or not with a laser (ICP or LA-ICP-MS) (see e.g. Dayet, 2021). Geochemical data are usually treated by using multivariate statistical analyses (Attard Montalto et al., 2012; Dayet et al., 2016; Eiselt et al., 2011, 2019; Kiehn et al., 2007; MacDonald et al., 2011; Macdonald et al., 2013; MacDonald et al., 2018; Mauran et al., 2021, 2022; McGrath et al., 2022; Pierce et al., 2020; Popelka-Filcoff et al., 2007, 2008, 2012; Scadding et al., 2015; Velliky et al., 2019, 2021;

Zarzycka et al., 2019; Zipkin et al., 2015, 2017, 2020). These analyses allow to determine whether the different sources can be considered as different groups with different compositional features. This is a particular efficient way to control if the sources respond to the provenance postulate (Weigand et al., 1977). The definition of what is considered as a group in the supervised statistical analyses (multivariate methods that require a classification with predefined groups), will conditioned the results. A group is composed of samples that are assumed to be representative of the source variability. But a source can have various definition as described above, which means that the way the groups are made in the supervised statistical analyses may vary from a study to another. The archaeological question in particular may have an influence on source definition. For instance, Mauran et al. (2021) aimed at questioning the existence of raw material circulation between different rock art massifs, while Velliky et al. (2021) mostly aimed at identifying non local material to question the "modernity" of past populations in Europe. The first defined a source as a geological region in relation with the rock art massifs. The second grouped together part of the local outcrops in order to separate them from distant outcrops. However, a major issue in group making is the importance of lithological features. Should lithological features be taken into account or not when the groups are defined? Or in other words, is the intra-source variability not too high when iron-bearing rocks with different composition in major elements are considered as coming from a single source? This is the issue we explore in this paper.

Statistical data treatment

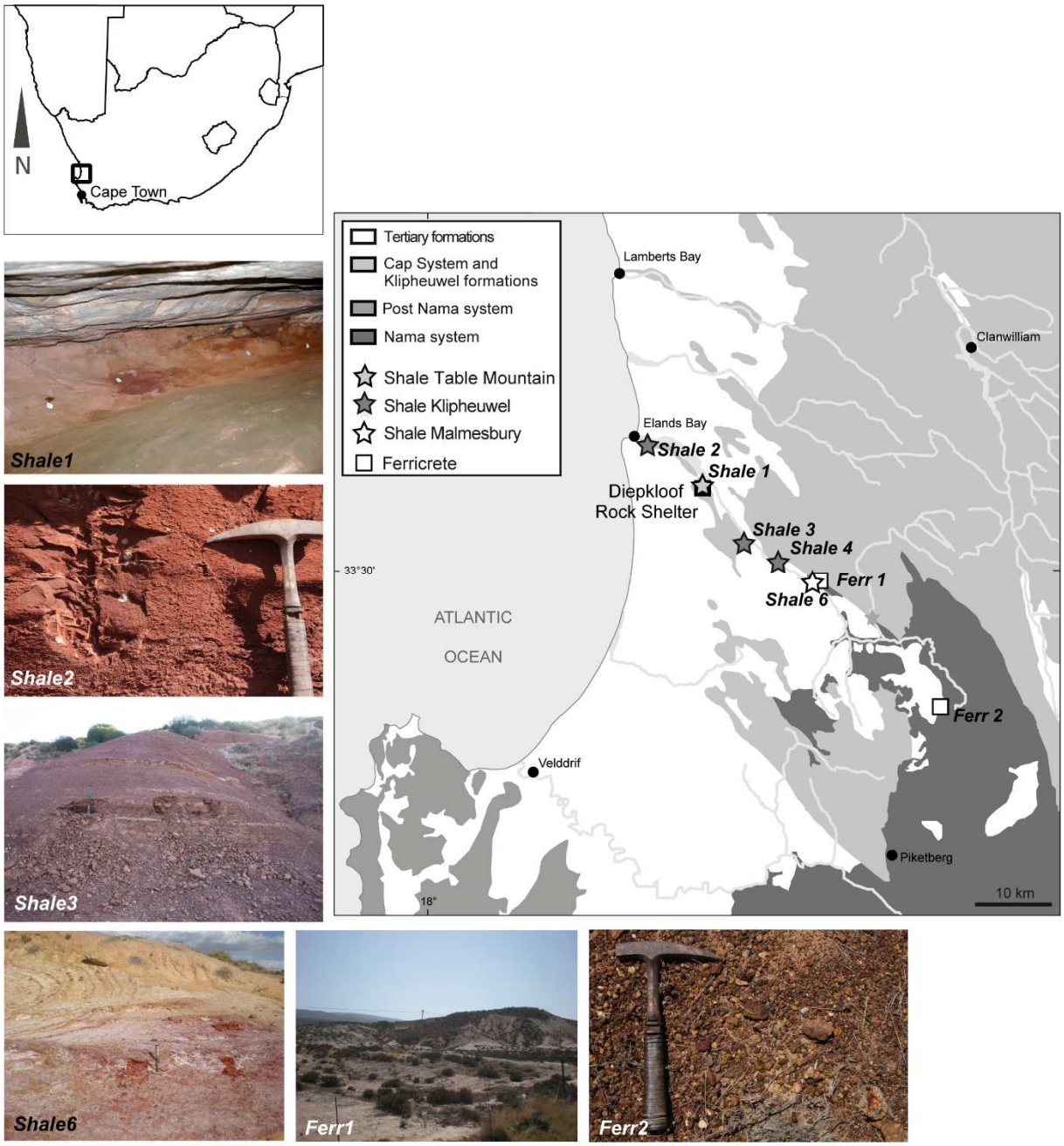
Two main multivariate analyses are used to treat the geochemical data in ochre provenance researches: the PCA (Dayet et al., 2016; Eiselt et al., 2011; MacDonald et al., 2011; Macdonald et al., 2013; MacDonald et al., 2018; Mauran et al., 2021; Pierce et al., 2020; Popelka-Filcoff et al., 2007, 2008, 2012; Tarbska et al., 2008; Zarzycka et al., 2019; Zipkin et

al., 2015, 2020) and the LDA, with a variant called canonical discriminant analysis, CDA (MacDonald et al., 2011; Macdonald et al., 2013; Mauran et al., 2021, 2022; McGrath et al., 2022; Pierce et al., 2020; Popelka-Filcoff et al., 2008; Scadding et al., 2015; Velliky et al., 2019, 2021; Zipkin et al., 2015, 2017). In a PCA, there is no assumption about the groups that should be made. The variability of the dataset is explored and synthesized in new variables that are called principal components. They represent the highest variability that can be observed amongst the geological samples. This variability can be in relation with the sources or can be in relation with other factors. For instance, the PCA performed by using a selection of trace elements based on their correlation with iron, on the geological sources sampled around Diepkloof rock shelter, South Africa, allow to group the raw materials depending on their lithology: shale pieces on one hand, ferricrete pieces on the other (Dayet et al., 2016). The sources of shale were not clearly separated. In this particular case the variability between the two categories of rocks is higher than the variability between the outcrops of one category of rock. A second selection of trace elements and the selection of the shale outcrops only in a second PCA allow to separate the outcrops of shale.

Another multivariate analysis used in ochre provenance researches is the LDA. In this supervised analysis, there is an assumption about what the groups should be. The analysis allows finding the variables that account the most in the discrimination of the groups. It calculates new variables, the linear discriminant variables or canonical discriminant variables, that are responsible for the best separation between the groups. It also allows predicting the affiliation of an unknown sample to a group. This is the analysis that requires the highest attention in the definition of the sources as designated groups. For instance, the results will be different if samples from the same location are grouped together or if a difference is made in function of the rock category encountered at one particular location. In other terms, if we consider that the outcrops are consistent within the same location or if we consider that there

151 are two different outcrops of two different raw materials at the same location the results will
152 be different. After carrying out PCA in order to determine if the data are comparable, we will
153 evaluate the influence of the grouping or the separation of the samples of different lithologies
154 from the same locality on the results of the LDA.
155
156 *Attribution of archaeological pieces*
157 Both the PCA and LDA allow in a second step that consists in projecting the data of
158 archaeological pieces. The projection into the multivariate space of the archaeological data
159 will allow to attribute the pieces to one or the other source. We usually used graphic interfaces
160 to estimate the attribution to a source, for both multivariate analyses and biplots of elements.
161 For instance, the 11 archaeological pieces from the French Palaeolithic site Ormesson plot
162 into the ellipse with one sigma confidence of the samples from the Sparnacian geological
163 stage (Mathis et al., 2014). The interpretation is that the archaeological samples were
164 collected within this formation. In some cases, the samples plot aside the geological sources.
165 In this condition it is not possible to attribute the pieces to a source. In some other cases, a
166 cloud of points representing archaeological pieces fall at the frontier of a geological group. In
167 these conditions most authors considered that the archaeological pieces belong to this group
168 (MacDonald et al., 2018; Velliky et al., 2021). In LDA analyses, a prediction of attribution of
169 unknown samples to one or the other designated group can be done. Such predictions were
170 done by Mauran et al. (2021) in order to affiliate archaeological samples to different
171 geological massifs. The results allow them to attribute a provenience to part of the
172 archaeological samples. Velliky et al. (2021) also used Mahalanobis distance equations to
173 calculate group membership probability.
174 At Diepkloof rock shelter, there is a source right back in the shelter. This outcrop of shale can
175 be considered as a reference for most of the shale pieces found in the archaeological levels.

176 Here we will consider that the shale pieces attributed to this source according to a previous
177 work (Dayet et al., 2016) do come from this source. We will then rediscuss this conclusion in
178 the light of a the new PCA and LDA results.
179
180 *Archaeological background*
181 Diepkloof rock shelter is located near the west coast in the Western Cape in South Africa,
182 about 200 km north of Cape Town (32° 23′ 12″ S, 18° 27′ 10″ E; Fig. 1). The shelter
183 dominates the small Verlorenvlei River valley about 14 km from the present shoreline. From
184 1999 to 2013, it was investigated by a South African-French team led by J.-P. Rigaud, P.-J.
185 Texier, and C. Poggenpoel. Middle Stone Age (MSA) and Later Stone age (LSA) occupations
186 were recovered (Parkington et al., 2013). The MSA occupations are dated from about 110 to
187 55 ky (Tribolo et al., 2013). The MSA sequence is composed of major lithic industries in
188 South Africa such as the Still Bay and the Howiesons Poort (Porraz et al., 2013). Also
189 noticeable, hundreds of engraved ostrich egg shells were recovered in Howiesons Poort units
190 (Texier et al., 2010, 2013). The MSA occupations as a whole yielded several hundreds of
191 ochre pieces (Dayet et al., 2013). The samples studied in the present work come from the
192 'main sector' (squares M6 and N6). The small collection of 11 samples selected for analysis is
193 mainly composed of shale featuring high clay mineral contents and ferricrete featuring high
194 iron oxide contents (Dayet et al., 2016).
195



196

197 **Fig. 1 – Location of Diepkloof Rock Shelter, geological map of the surroundings of the**

198 **site and pictures of the outcrops of shale and ferricrete studied (modified after Dayet et**

199 **al. 2016).**

200

201

202 *Geological background*

In the vincinity of Diepkloof rock shelter two main categories of rocks containing iron oxides are described: coloured shale (laminated rocks rich in clay minerals) and ferricrete (iron-enriched nodules). Shale outcrops occur throughout the three main geological formations present around Diepkloof (Besaans et al., 1973; Fig. 1):

- The closest shale outcrop is located inside the rock shelter itself: it occurs as a thin dusky red bed in the back of the shelter, outcropping within quartzitic sandstones of the Table Mountain group (Paleozoic). This outcrop is called Shale1. It will be our reference for shale attribution.
- Klipheuphel beds (proto-Paleozoic) are located along the Verlorenvlei River which runs below the Diepkloof hill, the nearest being 10 km south of the site. Two outcrops were studied here, Shale2 and Shale3, that share similar lithological features.
- Malsmebury's red beds of shale (proto-Paleozoic, Nama System) were observed at more than 20 km south-west of the site. They are finely laminated phyllitic shale (Saggerson & Turner, 1995). One outcrop was studied, Shale6.

Ferricrete nodules are observed within tertiary to quaternary soils (Roberts, 2003; author's observations; Fig. 1). The reported ferricrete outcrops are located near Malmesbury formations, south-west from the site, and some of them are clearly associated with phyllitic shale beds. Localized lateritic weathering profiles explain such association. Two outcrops were studied, Ferr1 and Ferr2. The first is located just aside Shale6, they are from the same locality and belong to the same alteration profile although their petrological features and composition in major elements are different. Shale6 is the saprolite (the altered parent rock) and Ferr1 are the nodules of the dismantled ferricrust usually found above a first horizon of alteration. This horizon has disappeared. Ferr2 is partially composed of highly weathered shale pieces enriched in iron oxide. It can be separated in two subgroups, Ferr2a which is composed of indurated shale with a similar structure than the shale but with higher iron

contents, and Ferr2b which is composed of true ferricrete. They come from the same parent rock. Ferr2 is located at more than 20 km from the site. The sources described above can be either considered separately (sources Shale6 and Ferr2a considered as separate from sources Ferr1 and Ferr2b), or grouped together (source Shale6-Ferr1 and source Ferr2). The samples of each locality come from the same geological context but their lithological features are different.

3. Material and methods

3.1 Geological materials

At each outcrop, samples were taken at different position, horizontally and vertically. At least one sample per position were studied. From 10 to 19 samples were analysed per locality (Table 1) with a total of 64 samples. Pictures of some samples are presented in Fig. 2. Rock category and outcrop features were precisely defined by using Scanning electron microscopy coupled with energy dispersive spectrometry (SEM-EDS) and X-ray diffraction (XRD) (Dayet et al., 2016, 2013; Fig. 3).

Source	Number of samples	Locality	Rock category	Geological formation	Texture	Fabric
Shale 1	12	Redelinghuys1	Shale	Table montain series	Fine-grained	Laminated
Shale 2	10	Elands Bay	Shale	Klipheuwel formations	Fine-grained	Laminated
Shale 3	10	Redelinghuys2	Shale	Klipheuwel formations	Fine-grained	Slightly laminated
Shale 6	10	Redelinghuys3	Phyllitic shale	Malmesbury formations	Fine-grained	Finely laminated
Ferr 1	9	Redelinghuys3	Ferricrete	Tertiary/quater., associated with Malmesbury form.	Fine-grained	Massive
Ferr 2a	4	Eendekuil	Indurated shale	Tertiary/quater., associated with Malmesbury form.	Fine-grained	Slightly laminated, laminated
Ferr 2b	9	Eendekuil	Ferricrete	Tertiary/quater., associated with Malmesbury form.	Fine-grained, fine and coarse-grained	Massive, globular
TOTAL	**64**					

Table 1 – Macroscopic features of the different iron-bearing rock outcrops.

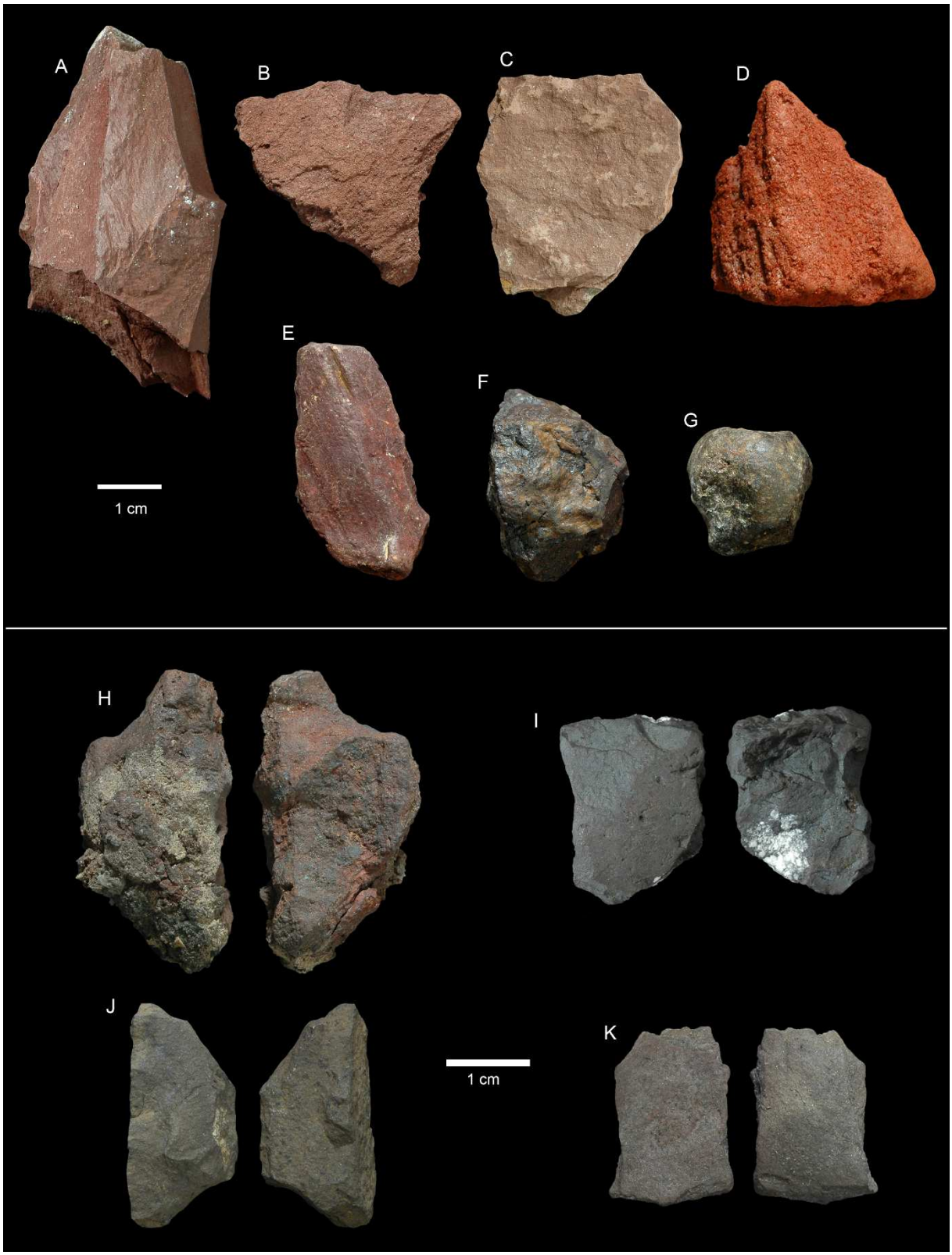


Fig. 2 – Pictures of geological (A to G) and archaeological (H to K) ochre samples. A: sample of shale from Shale1. B: sample of shale from Shale2; C: sample of shale from Shale3; D: sample of shale from Shale6; E: sample of indurated shale from Ferr2a. F:

sample of ferricrete from Ferr1. G: sample of ferricrete from Ferr2. H: sample of ferricrete Archeo1, unknown provenance. I: sample of ferricrete Archeo3, associated to Malsmebury formations. J: sample of shale Archeo7, attributed to the Shale1 source. K: sample Archeo8, attributed to Malmesbury formations.

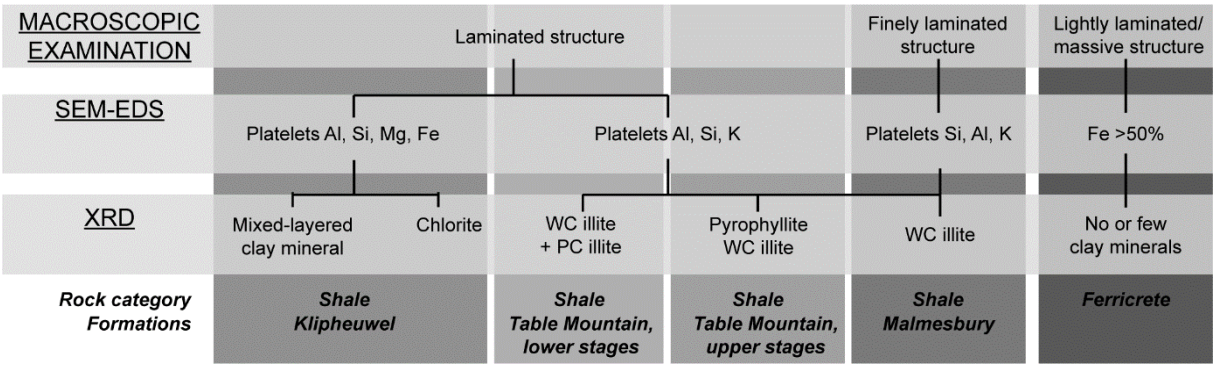


Fig. 3 – Specificity of each category of raw material and each geological formation and diagram of attribution of an unknown sample to a rock category or a geological formation. PC: Poorly crystallised; WC: well-cristallised.

3.2 Archaeological material

A total of 11 archaeological pieces were selected for ICP analyses. The petrological features of these samples were carefully analysed before their ICP analyses. Macroscopic, SEM-EDS and XRD analyses allow a first attribution of the pieces to one or the other source or geological formation based on their fabric and mineralogical composition (Dayet et al., 2016, 2013; Fig. 3). Finely laminated shale can be attributed to the phyllitic shale from the Malmsbury formations. Samples with a peak of well-crystallized illite can also be attributed to this formation. Samples with a mixture of poorly and well crystallysed illite, kaolinite and no other clay minerals can be attributed to the Shale1 source. Samples with a mixed-layered

271 clay mineral or with chlorite can be attributed to the shale sources from the Klipheuwel
272 formations. Samples with no or very low content of clay minerals can be attributed to the
273 ferricrete rock category.
274 The five first archaeological samples are ferricrete pieces (Table 2). Most ferricrete pieces
275 have an unknown provenience while two could be associated with the Malmesbury formation
276 according to the crystal shape of the iron oxides (Archeo3 and 4). The six last are shale
277 pieces. Amongst the shale pieces, four are attributed to Shale1 source according to
278 macroscopic, SEM-EDS, XRD and ICP analyses: Archeo6, Archeo7, Archeo10, Archeo11
279 (Table 2). The two other shale samples could come from the Malmesbury formations
280 (Archeo8 and 9).
281

Sample N°	Ref.	Stratigraphic unit	Square	Lithic industry	Mass	Texture	Fabric	Rock category	Provenience
Archeo1	**BDX13690**	EBEN	N6B	Late HP	11,2	CLAY-SILT	MASSIVE	FERRICRETE	Unknown
Archeo2	**BDX13722**	DEBBIE	M6A	Late HP	8,2	CLAY	FOLIATED	FERRICRETE	Unknown
Archeo3	**BDX13741**	ERIC	M6C	Late HP	6,4	CRYSTALLINE	MASSIVE, POROUS	FERRICRETE	Associated with Malmesbury formations
Archeo4	**BDX13749**	ERIC	M6C	Late HP	3,9	CRYSTALLINE	MASSIVE, POROUS	FERRICRETE	Associated with Malmesbury formations
Archeo5	**BDX13793**	JUDE	N6A	Und., above early HP	4,4	CLAY-SILT	MASSIVE	FERRICRETE	Unknown
Archeo6	**BDX13681**	CLAUDE	N6D	Post HP	6,7	CLAY-SILT	LAMINATED	SHALE	Shale1
Archeo7	**BDX13715**	DARRYL	M6A	Late HP	4,1	CLAY-SILT	LAMINATED	SHALE	Shale1
Archeo8	**BDX13718**	DEBBIE	M6A	Late HP	1,9	CLAY-SILT	FINELY LAMINATED	SHALE	Malmesbury formations
Archeo9	**BDX13752**	ESTER	M6A	Late HP	2,4	CLAY-SILT	LIGHTLY LAMINATED	SHALE	Malmesbury formations
Archeo10	**BDX13758**	GOVERNOR	N6D	Intermediate HP	7,3	CLAY-SILT	LAMINATED	SHALE	Shale1
Archeo11	**BDX13777**	KIM	N6B	SB	3,7	CLAY-SILT	LAMINATED	SHALE	Shale1

282 **Table 2 – Macroscopic features of the archaeological pieces and attribution to a geological formation or an outcrop.**

283

3.3 Sample preparation

Geological and archaeological pieces were reduced into powder before being analysed by ICP methods. Their surface was cleaned with demineralised water, and then samples were prepared using a diamond blade: a fragment was cut and the external surface was removed by abrasion against the diamond blade. Each fragment was subsequently crushed into powder with an agate milling bowl and then finely ground with a micro-grinder equipped with an agate bowl.

3.4 ICP-MS and ICP-OES analyses

Elemental analyses were carried out by the *Service d'Analyse des Roches et Minéraux* (SARM) at Nancy, France. Major and minor elements were determined by Inductive Coupled Plasma Optical Emission Spectroscopy (ICP-OES). The instrument used is ICP-OES iCAP Pro (Thermo Scientific). Trace elements were analysed by Inductive Coupled Plasma Mass Spectrometry (ICP-MS), with an iCAP Q spectrometer (Thermo Scientific). After being ground into powder, 200 mg of each sample was melted with $LiBO_2$ and then dissolved in nitric acid. A total of 43 trace elements (dosed by ICP-MS, expressed as ppm) and 10 major or minor elements (dosed by ICP-OES, percentage calculated as oxides) were quantified. The raw data are given in Table S1 and S2.

3.5 Statistical data treatment

Data pre-treatment

Variables with missing values, which means, elements with values lower than detection limits of the ICP analyses, were not taken into account in the statistical analyses, as advocated by Popelka-Filcoff & Zipkin (2022). An exception was made for Sb with only 3 % of missing values (geological samples alone), because of its interest as an element positively correlated

with Fe (Dayet et al., 2016). In this case missing values were replaced by 0.6 x the minimum value (Comas Cufí & Thió i Fernández de Henestrosa, 2011). The element Pb was not taken into account due to its presence in the fuel of cars and the proximity of roads near some geological outcrops. In order to decrease the number of correlated variables (for LDA analyses the variables must not be collinear), a PCA of all the trace and major elements were carried out. Then, by using the variable coordinates' biplot (Dim1-Dim2), elements with collinear vectors were removed. A first set of elements results from this selection: MgO, K_2O, Al_2O_3, V, Cr, Fe_2O_3, Co, Ni, Ga, Ge, As, Rb, Zr, In, Cs, Hf, Th, U (set1). A second PCA with the trace elements only were carried out. Again, we used the variable coordinates' biplot (Dim1-Dim2) to graphically select the variables that are not highly correlated. A second set of elements results from this treatment: V, Cr, Co, Ni, Ga, Ge, As, Rb, In, Cs, Hf, Th, U (set2). Biplots of the loadings of the two PCA can be found in Fig. S1. A third set of elements were used: elements that are positively correlated with Fe in shale outcrops, according to the procedure proposed by Popelka-Filcoff et al. (2007). This allow selecting the Fe oxide element signature. To be more accurate, in this study we kept only elements whose correlation with Fe is higher than with Al for the shale outcrops. This is the set of elements that were used in a PCA in a previous study, allowing to separate shale and ferricrete geological pieces (Dayet et al., 2016). This set of elements (Pb excepted) was reinvestigated in the present work: V, Cr, Co, As, Nb, Sb, Ta, Ba, U (set3).

The elemental composition cannot be used as percentages in a multivariate analysis. The data must be transformed. Centered logarithm ratios (clr) of the elemental percentages were used for set1 and set2 (Aitchison, 1986). Clr transformation normalizes the compositional data by the geometric mean of the variables (transformed into logarithm). Clr transformation was used in two previous ochre provenance studies (Mauran et al., 2021; Pierce et al., 2020). Additive logarithm ratios (alr) of the elemental percentages were used for set1, set2 and set3. Alr

transformation normalizes the data by the concentration of one element (transformed into logarithm). In this work, ratios to Fe were used as already described in many ochre provenance studies (Dayet et al., 2016; Eiselt et al., 2011; MacDonald et al., 2011; Macdonald et al., 2013; MacDonald et al., 2018; Pierce et al., 2020; Popelka-Filcoff et al., 2007, 2008, 2012; Velliky et al., 2019, 2021). We used ratios to Fe_2O_3 considering that data are given in mass oxide, using the following formula: log ([element]/[Fe_2O_3]). These data were calculated by using the software CODAPAK 2.03.01 (Comas Cufí & Thió i Fernández de Henestrosa, 2011).

Mulivariate analyses

PCA were carried out by using the FactoMineR package (Lê et al., 2008) of R software version 4.1.2. LDA were carried out by using the MASS package (Venables & Ripley, 2002) of R software version 4.1.2. Kfold cross-validation were done by using the caret package (Kuhn, 2008). The sample is divided into k smaller sets (folds) and the model is trained on k-1 folds and validated with the last one. A number of three folds were used for the kfold cross validation considering the small number of samples included in the analysis. The closer from 1 the accuracy and the kappa statistic are, the better the LDA (Kuhn, 2008). The Kappa statistic is a measure of concordance for categorical data (the groups) that measures agreement relative to what would be expected by chance. All the graphics were made by using the ggplot2 R package. Codes are given in Supplementary Information. The loadings of the PCA and the coefficients of the linear discriminants of the LDA are given in Supplementary Information (Table S3 to S14). They will not be discussed in the text as they are not the main aspect this paper deals with.

4. Results

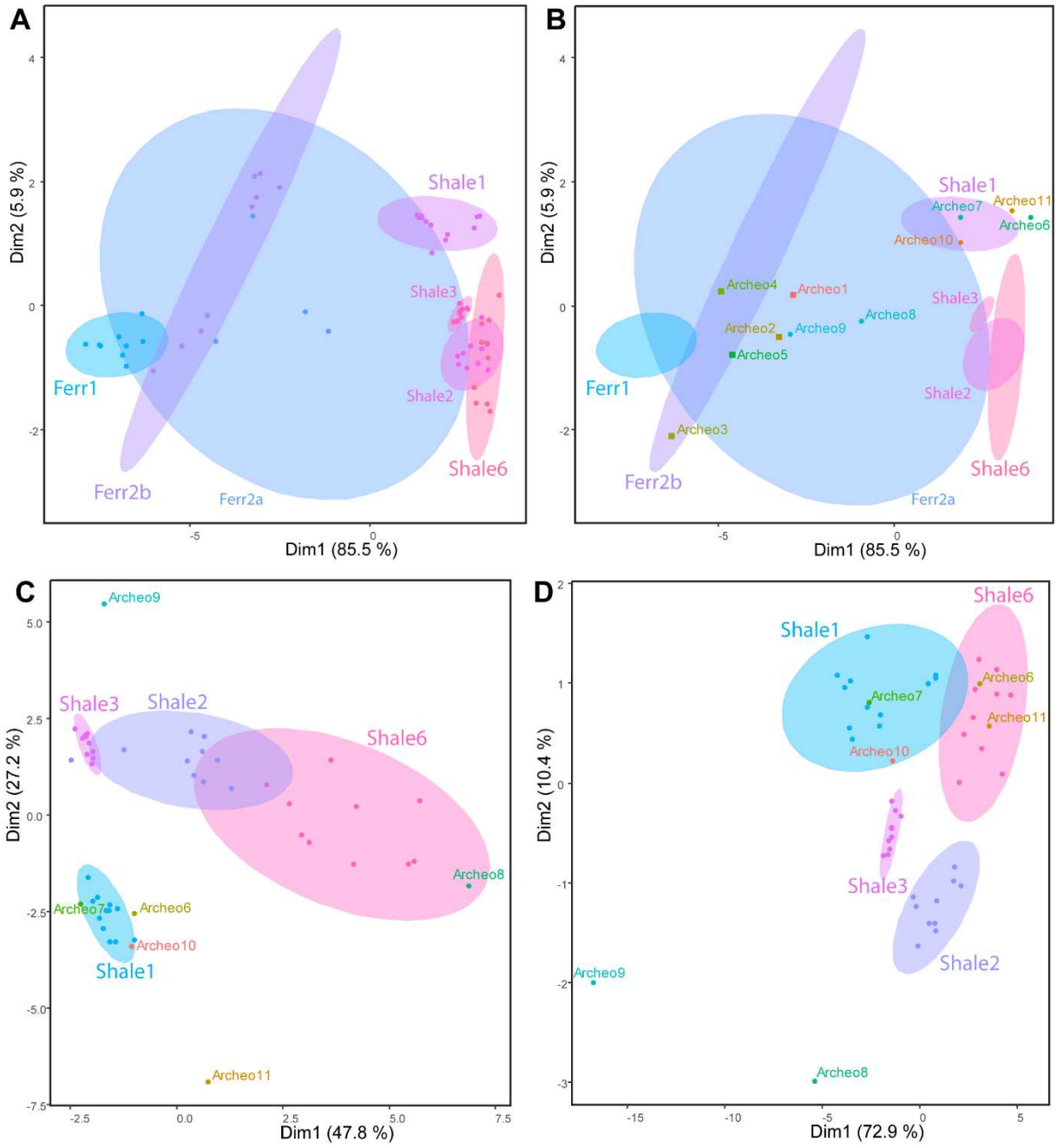
4.1 Unsupervised learning method, PCA

4.1.1 All the sources

We will explore the results of PCA analyses done without any assumption about what the sources should be. When all the sources are considered, the alr ratios to Fe_2O_3 allow a better discrimination between the sources than the clr, whatever the set of elements used. The attribution of the shale pieces to the Shale1 source is similar between the PCAs done on alr and clr coordinates, whatever if the set1, set2 or set3 is considered. We will focus on the alr results for set1.

PCA1: PCA done on elements' set1, alr ratios to Fe_2O_3 coordinates

The PCA done by using the elements' set1 and the alr ratios to Fe_2O_3 coordinates do not allow a good discrimination between the sources (Fig. 4A). Amongst the shale sources, only Shale1 and Shale3 are clearly distinguished on the Dim1-Dim2 biplot. Nonetheless, the ferricrete sources are all found on the left of the biplot while shale sources are all found on the right (discrimination along the Dim1 axis). Ferr2a is partly distributed with Ferr2b and in between shale and ferricrete pieces. Archeo 7 and 10 are attributed to the Shale1 source while Archeo6 and Archeo11 are very close to its distribution (Fig. 4B). The ferricrete pieces are all found nearby the distribution of Ferr2b or Ferr2a. The archaeological samples are close enough to the sources to allow using predictions in the LDA.



381

382 **Fig. 4 – Results of the PCAs. A: PCA1 carried out by considering all the sources with the**
383 **trace elements' set3 (V, Cr, Co, As, Nb, Sb, Ta, Ba, U) and alr ratios to Fe_2O_3**
384 **coordinates. B: PCA1 with archaeological samples. C: PCA2 carried out by considering**
385 **the sources of shale alone with the trace elements' set2 (V, Cr, Co, Ni, Ga, Ge, As, Rb,**
386 **In, Cs, Hf, Pb, Th, U) and clr coordinates, with archaeological pieces. D: PCA3 carried**
387 **out by considering the sources of shale alone with trace elements' set3 and alr ratios to**

Fe_2O_3 coordinates, with archaeological pieces. Squares: ferricrete pieces; Circles: shale pieces. Confidence ellipses with a 95 % of probability.

4.1.2 Outcrops of shale alone

When only the sources of shale are taken into account, the sources are graphically better separated when using the clr coordinates than the alr ratios to Fe_2O_3 for set 1 and set 2. The results for set1 and set2 are very close whatever if the clr or alr coordinates are used. We will focus on the results of the clr coordinates for the set2 and then on the alr coordinates of the set3.

PCA2: PCA done on elements' set2, clr coordinates

The separation between the shale sources when reducing the geological collection to shale sources and doing a PCA on trace elements' set2 is not very good (Fig. 4C). Archeo7 and Archeo10 fall in the distribution of the Shale1 source in the Dim1-Dim2 space. Archeo6 is very close to the distribution but Archeo11 is significantly distant.

PCA3: PCA done on elements' set3, alr ratios to Fe_2O_3 coordinates

Apart for a small overlap between Shale1 and Shale6 sources, the sources of shale are well separated in the Dim1-Dim2 space of the PCA done on the shale sources only by using the trace elements' set3 (Fig.4D). Archeo7 and Archeo10 fall in the distribution of Shale1 in the Dim1-Dim2 biplot while Archeo6 and Archeo11 fall in the distribution of Shale6, which is very close to the one of Shale1.

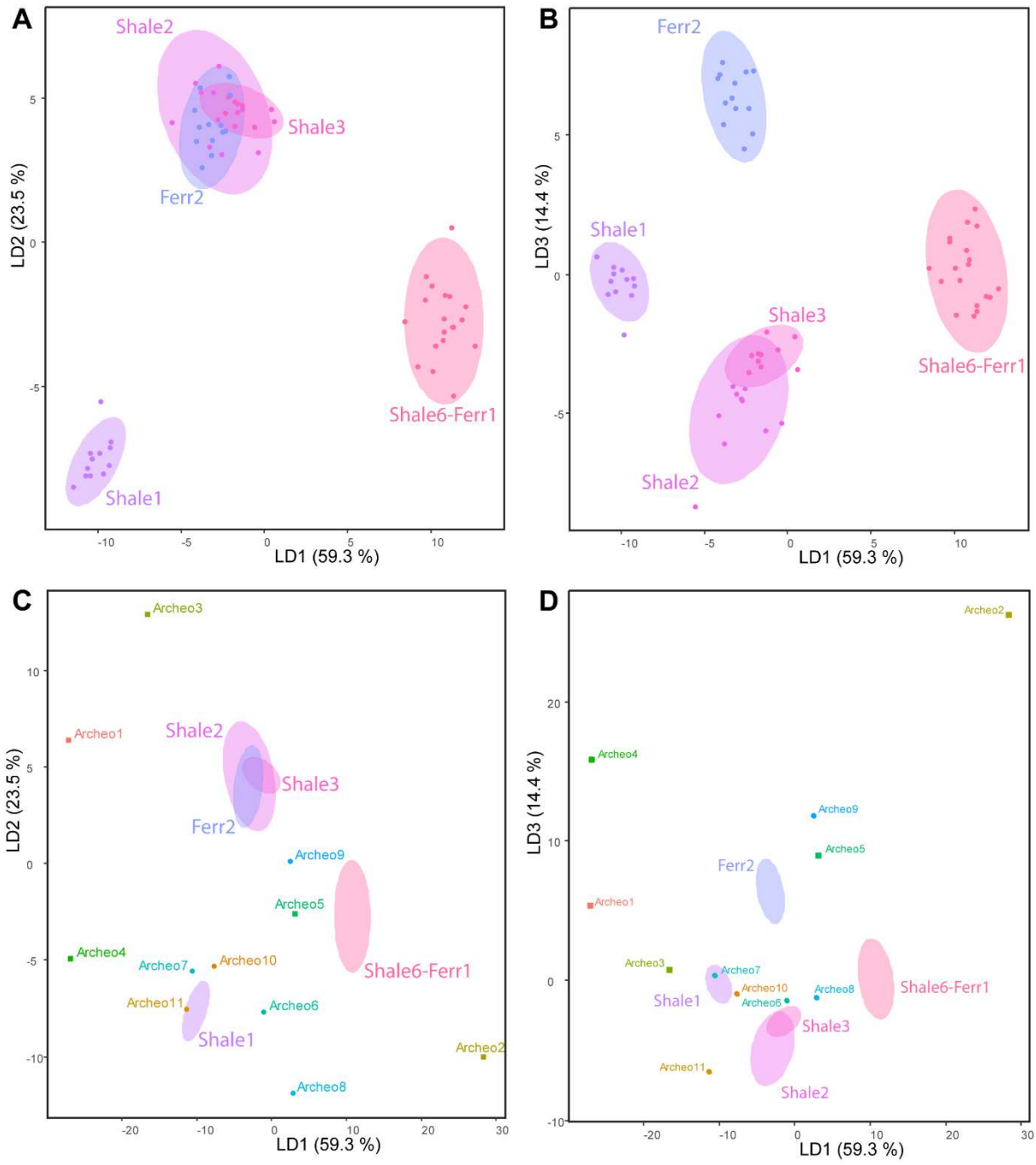
4.2 Supervised learning method, LDA

4.2.1 Sources defined as localities (shale and ferricrete grouped together)

First, we considered that a source is a locality, a GPS point where several types of iron-bearing rocks are encountered. Following this view, we grouped together Shale6 and Ferr1 and we made no separation between the indurated shale Ferr2a and the true ferricrete Ferr2b.

LDA1: LDA done on elements' set1, clr coordinates and alr ratios to Fe_2O_3 coordinates

When using clr coordinates, some elements from set1 are collinear despite our selection. A LDA cannot be applied. There is no more collinearity when using the alr ratios to Fe_2O_3 coordinates. The LDA carried out on the elements' set1 with the alr coordinates (LDA1) allows a good separation between Shale6-Ferr1, Shale1 and the other sources (Fig. 5A). Along LD1 and LD2 biplot, however, Shale2, Shale3 and Ferr2 are not separated. Shale2 and Shale3 are still not separated along LD1 and LD3 biplot (Fig. 5B). This might be due to shared features given that they come from the same geological formation. Nonetheless, the attribution of each sample to each source is good, there is no mistake done in the model. Moreover, when performing a kfold cross validation, the probability and kappa equal 0.98. The statistical model is strong. When we plot the archaeological samples in the LD1-LD2 space, we observe that only Archeo11 is attributed to the Shale1 group (Fig. 5C) while in the LD1-LD3 space only Archeo7 plot in the Shale1 group (Fig. 5D). The ferricrete pieces do not follow a clear trend, most are not closer to ferricrete sources than to shale sources. When we consider the predictions of the statistical model, the four shale pieces from Shale1 are attributed to this group but some ferricrete samples are also attributed to it.



436

437 **Fig. 5 – Results of LDA1 carried out by considering the sources of shale and ferricrete**
438 **from the same locality grouped together with the elements' set1 (MgO, K_2O, Al_2O_3, V,**
439 **Cr, Fe_2O_3, Co, Ni, Ga, Ge, As, Rb, Zr, In, Cs, Hf, Th, U) and alr ratios to Fe_2O_3**
440 **coordinates. A: LD1 and LD2 biplot. B. LD1 and LD3 biplot. C: LD1 and LD2 biplot**
441 **with the archaeological pieces. D: LD1 and LD3 biplot with the archaeological pieces.**

Squares: ferricrete pieces; Circles: shale pieces. Confidence ellipses with a 95 % of probability.

LDA2: LDA done on trace elements' set2, clr coordinates and alr ratios to Fe_2O_3 coordinates

Again, when using clr coordinates some elements appear to be collinear. This is not the case when using alr ratios to Fe_2O_3 coordinates. The results of the LDA done on the trace elements' set2 with alr coordinates (LDA2) show a good separation between the five sources (Fig. 6A). The results are good, all the geological samples are attributed to the right source. As regard to the kfold cross-validation, the accuracy is of 0.95 and the kappa statistic is of 0.94. Sample Archeo6 is attributed to the Shale1 group when considering the LD1-LD2 space (Fig. 6B). Archeo10 and Archeo7 are very close to the distribution but Archeo11 departs significantly. Most ferricrete pieces plot far away from the ferricrete sources. When we look at the predictions of the statistical model, the four pieces from Shale1 are correctly attributed to this group but another sample of shale and a sample of ferricrete are also attributed to it. The other ferricrete pieces are attributed to Ferr2.

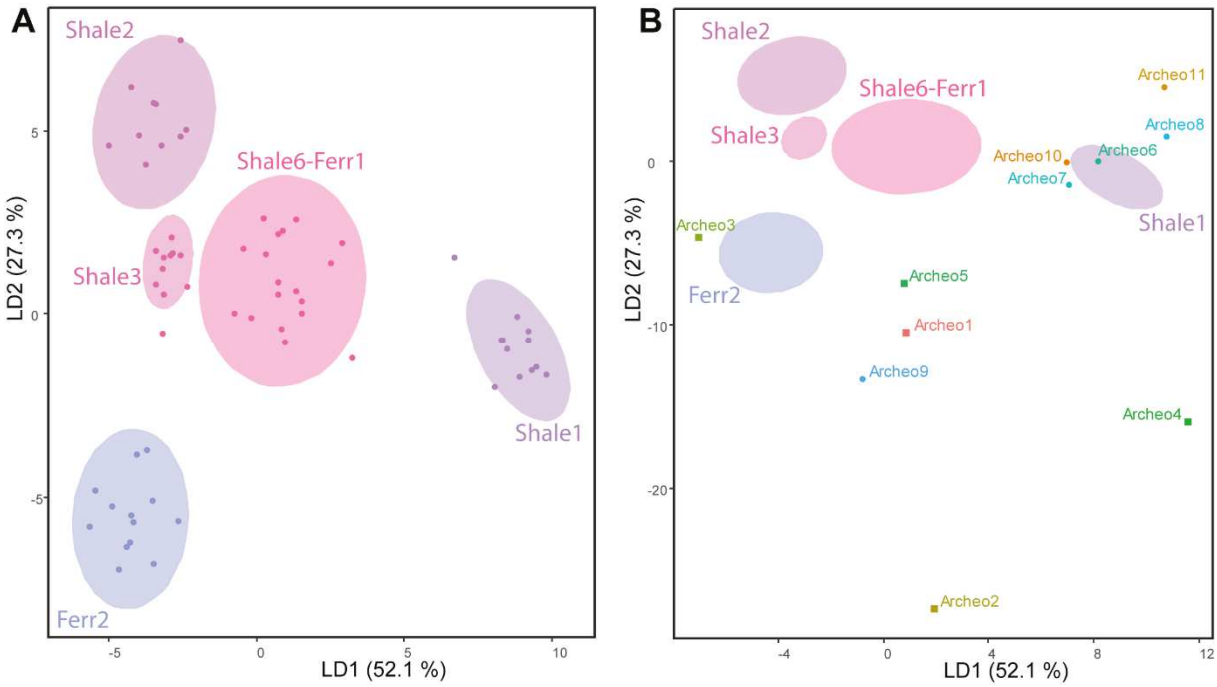


Fig. 6 – Results of LDA2 carried out by considering the sources of shale and ferricrete from the same locality grouped together with the trace elements' set2 (V, Cr, Co, Ni, Ga, Ge, As, Rb, In, Cs, Hf, Th, U) and alr ratios to Fe_2O_3 coordinates. A: LD1 and LD2 biplot. B: LD1 and LD2 biplot with archaeological pieces. Squares: ferricrete pieces; Circles: shale pieces. Confidence ellipses with a 95 % of probability.

LDA3: LDA done on trace elements' set3, alr ratios to Fe_2O_3 coordinates

The LDA carried out by using trace elements' set3 with alr ratios to Fe_2O_3 (LDA3) allows a good separation between all the sources except for the two simplified weathering profiles (Fig. 7A and 7B). The results are good, all the geological samples are attributed to the right source. Moreover, when performing a kfold cross validation, the probability is of 0.92 and the kappa statistic of 0.90 which is still good. The ferricrete sources are separated from the shale sources. Samples Archeo6, Archeo10 and Archeo11 are attributed to the Shale1 group in the LD1-LD2 space (Fig. 7C). Archeo7 is very close to the Shale1 distribution. Most of the ferricrete pieces fall on the right part of the biplot where the ferricrete groups are represented. The attribution of the shale archaeological pieces is less conclusive on the LD1-LD3 biplot

(Fig. 7D) although the shale sources are better separated. When the predictions of the statistical model are considered, we observe that all the pieces from Shale1 are attributed to this group. Three of the ferricrete pieces are attributed to Ferr2 group while the two last are attributed to Shale2 group.

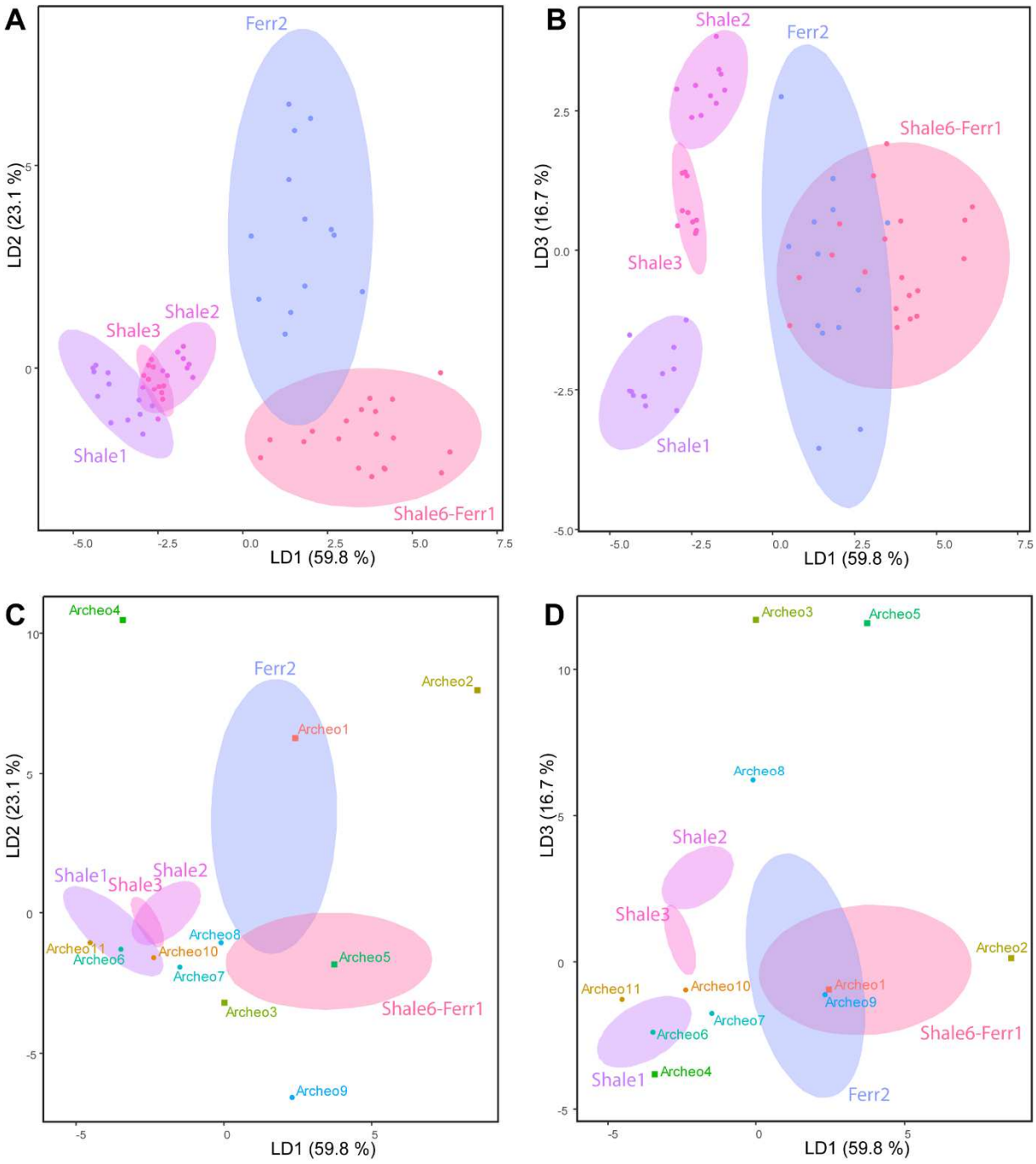


Fig. 7 – Results of LDA3 carried out by considering the sources of shale and ferricrete from the same locality grouped together with the trace elements' set3 (V, Cr, Co, As, Nb, Sb, Ta, Ba, U) and alr ratios to Fe_2O_3 coordinates. A: LD1 and LD2 biplot. B. LD1 and LD3 biplot. C: LD1 and LD2 biplot with the archaeological pieces. D: LD1 and LD3 biplot with the archaeological pieces. Squares: ferricrete pieces; Circles: shale pieces. Confidence ellipses with a 95 % of probability.

4.2.2 Sources defined from lithological features (shale and ferricrete separated)

In a second step, we can consider that shale are different raw materials than ferricrete. Their composition in major elements is different. As we can separate them from various features according to macroscopic examination, SEM-EDS and XRD analyses, we can also separate them in the statistical treatment. We considered here that Shale6 and Ferr1 as well as Ferr2a and Ferr2b are single separated sources of raw materials.

LDA4: LDA done on elements' set1, alr ratios to Fe_2O_3 coordinates

Some variables being collinear when using the clr coordinates from set1, we will focus on the results obtained with alr ratios to Fe_2O_3 coordinates. The results of the LDA carried out on the variables from the set1 with alr coordinates, when considering Shale6 and Ferr2a as single sources (LDA4), show a good separation between the sources except for Shale2 and Shale3 (Fig. 8A). This might be due to the fact that they come from the same geological formation. The results are good, all the geological samples are attributed to the right source. The accuracy of the kfold cross-validation is equal to 0.99 while the kappa statistic equal 0.98. The statistical model is strong. The ferricrete sources are on the left part of the LD1-LD2 biplot, the shale sources are on the right part while the indurated shale source (Ferr2a) is in

the middle (separation following the LD1 axis; Fig. 8A). This is consistent with their major element composition. None of the archaeological pieces is attributed to the Shale1 group on the biplot LD1-LD2 (Fig. 8B). Three of the five archaeological ferricrete pieces are close to ferricrete groups. When we look at the predictions of the statistical model, the four samples from Shale1 are correctly attributed, but another sample of shale and two samples of ferricrete are also attributed to this group.

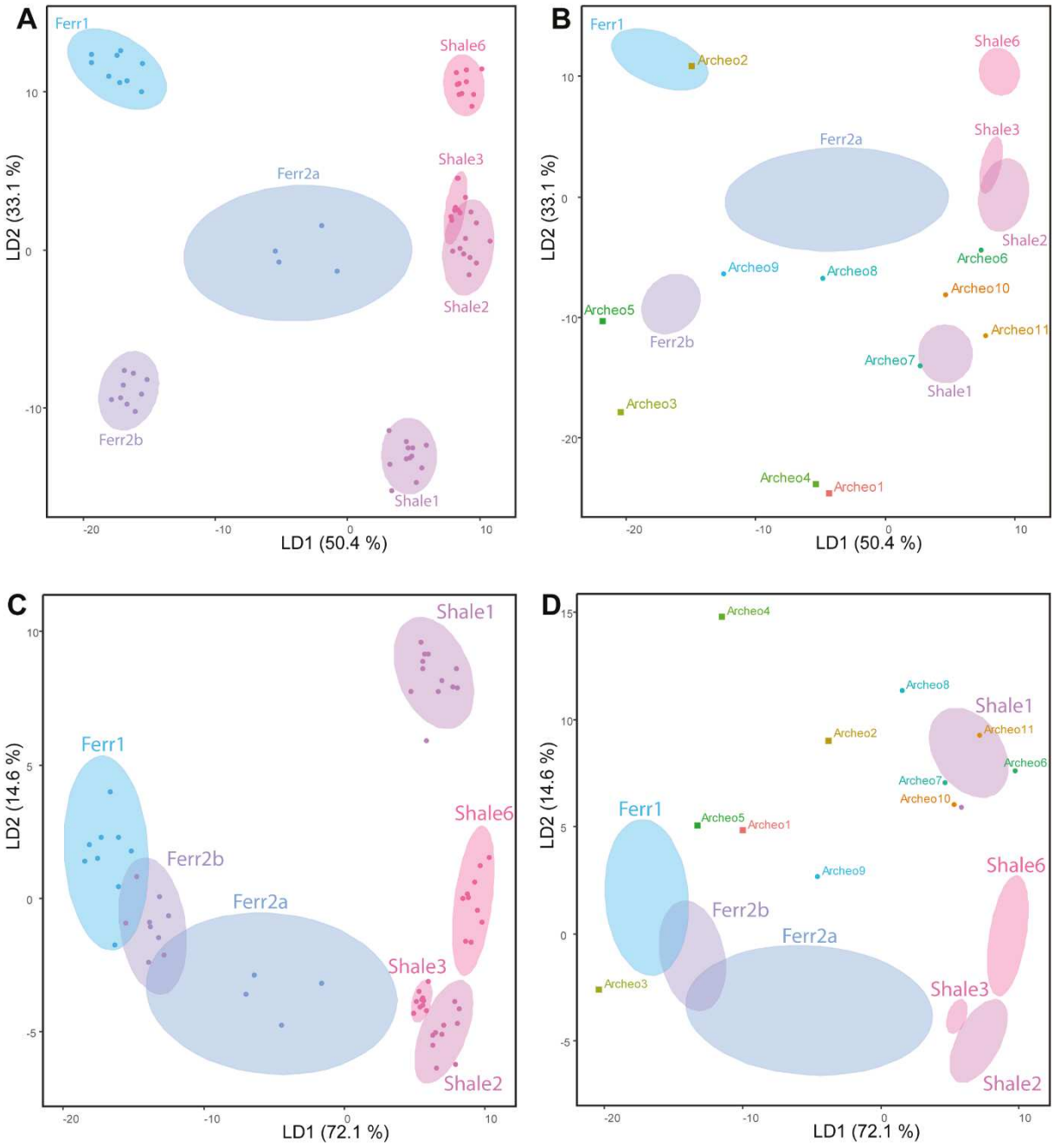


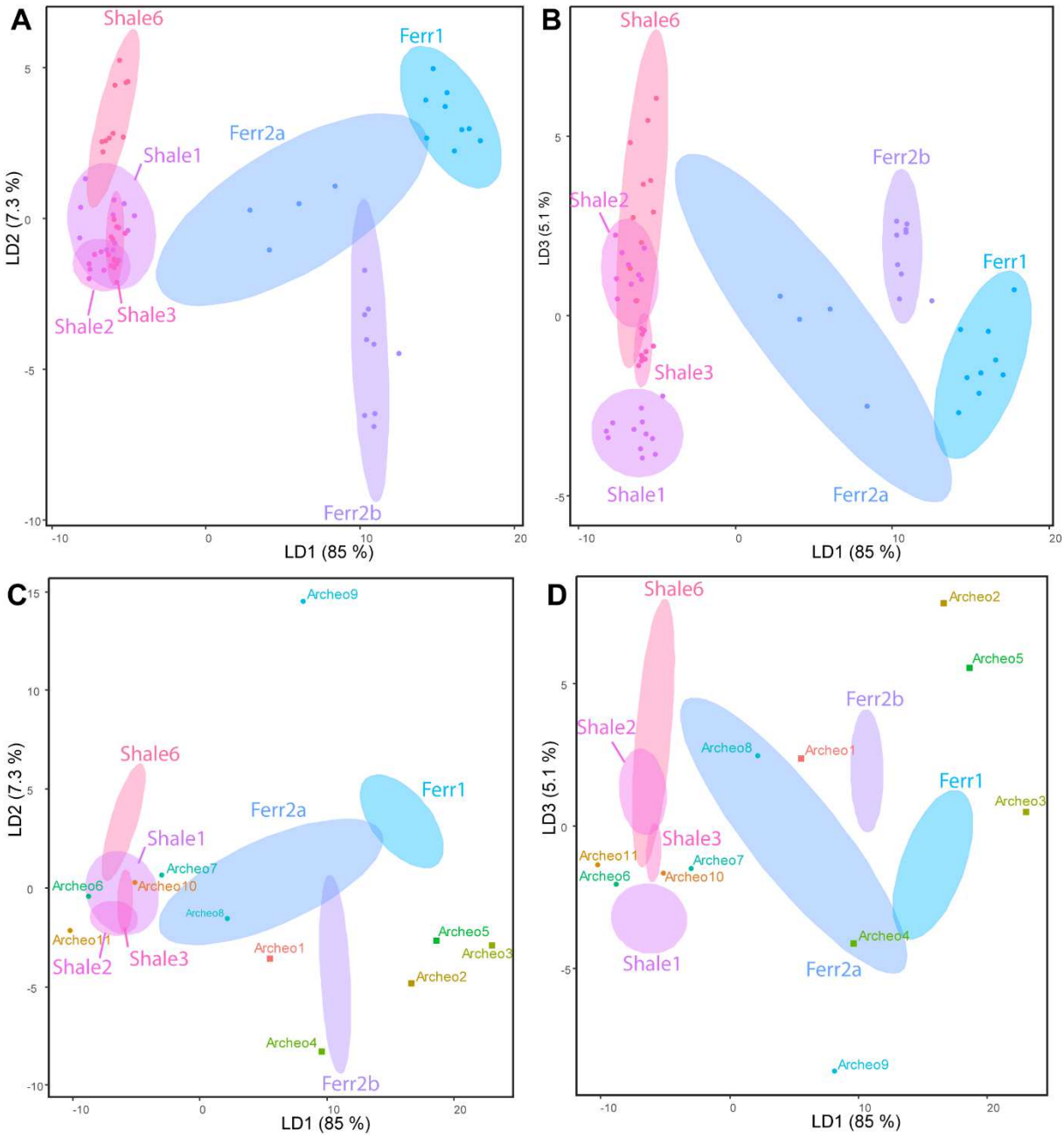
Fig. 8 – Results of the LDA carried out by considering the sources of shale and ferricrete from the same locality separated. A: LDA4 with the trace elements' set1 (MgO, K_2O, Al_2O_3, V, Cr, Fe_2O_3, Co, Ni, Ga, Ge, As, Rb, Zr, In, Cs, Hf, Th, U) and alr ratios to Fe_2O_3 coordinates, LD1 and LD2 biplot. B: LDA4 with archaeological pieces. C: LDA5 with the trace elements' set2 (V, Cr, Co, Ni, Ga, Ge, As, Rb, In, Cs, Hf, Th, U) and alr ratios to Fe_2O_3 coordinates, LD1 and LD2 biplot. B: LDA5 with archaeological pieces. Squares: ferricrete pieces; Circles: shale pieces. Confidence ellipses with a 95 % of probability.

LDA5: LDA done on trace elements' set2, clr coordinates and alr ratios to Fe_2O_3 coordinates

Some elements being collinear when using clr coordinates, we investigated the results of the LDA done with alr ratios ot Fe_2O_3 coordinates only. The LDA done on the trace elements' set2 variables with alr coordinates (LDA5) allows an incomplete separation of the ferricrete sources along the LD1-LD2 axes (Fig. 8C). Nonetheless, the attribution of each geological piece to a source is correct. As regard to the kfold cross validation, the accuracy of the statistical model is of 0.95 and the kappa statistic of 0.94. Again, ferricrete and shale sources are separated on the LD1-LD2 biplot. Archeo11 belongs to the distribution of Shale1 in the LD1-LD2 space (Fig. 8D). Archeo6, Archeo7 and Archeo10 could also be attributed to this group when considering their proximity to its distribution. All the ferricrete archaeological pieces fall in the left part of the LD1-LD2 biplot, where the ferricrete sources are grouped together. The predictions of the statistical model for shale samples indicate that five of them are attributed to the Shale1 group instead of four. The five ferricrete pieces are attributed to Ferr2b.

LDA6: LDA done on trace elements' set3, alr ratios to Fe_2O_3 coordinates

The LDA done on the trace elements' set3 variables (LDA6) does not allow a clear separation of the shale sources along the LD1-LD2 axes (Fig. 9A). Along the LD1-LD3 axes, the source Shale1 is better separated but not the others (Fig. 9B). Despite this graphical negative result, the attribution of each geological sample to each geological source is correct. Moreover, when doing a kfold cross validation, the accuracy of the statistical model is of 0.97 and the kappa statistic of 0.96. The model is strong enough. As for the two latter analysis, ferricrete and shale sources are separated along the LD1-LD2 axes. Archeo6 and Archeo10 are correctly attributed to the Shale1 group in the LD1-LD2 space (Fig. 9C). However, the proximity of the Shale2 and Shale3 groups makes difficult to be sure that this should be the right attribution. Four of the five ferricrete archaeological pieces are closer to ferricrete sources than to shale sources. The last one is close to the indurated shale group (Ferr2a). The results in the LD1-LD3 space are not good for the attribution of archaeological shale pieces: none of the shale fall in the distribution of the Shale1 group (Fig. 9D). When considering the predictions of the statistical model, the four shale samples from Shale1 are attributed to this group. The ferricrete pieces are attributed to one or the other ferricrete source except for one piece.



556

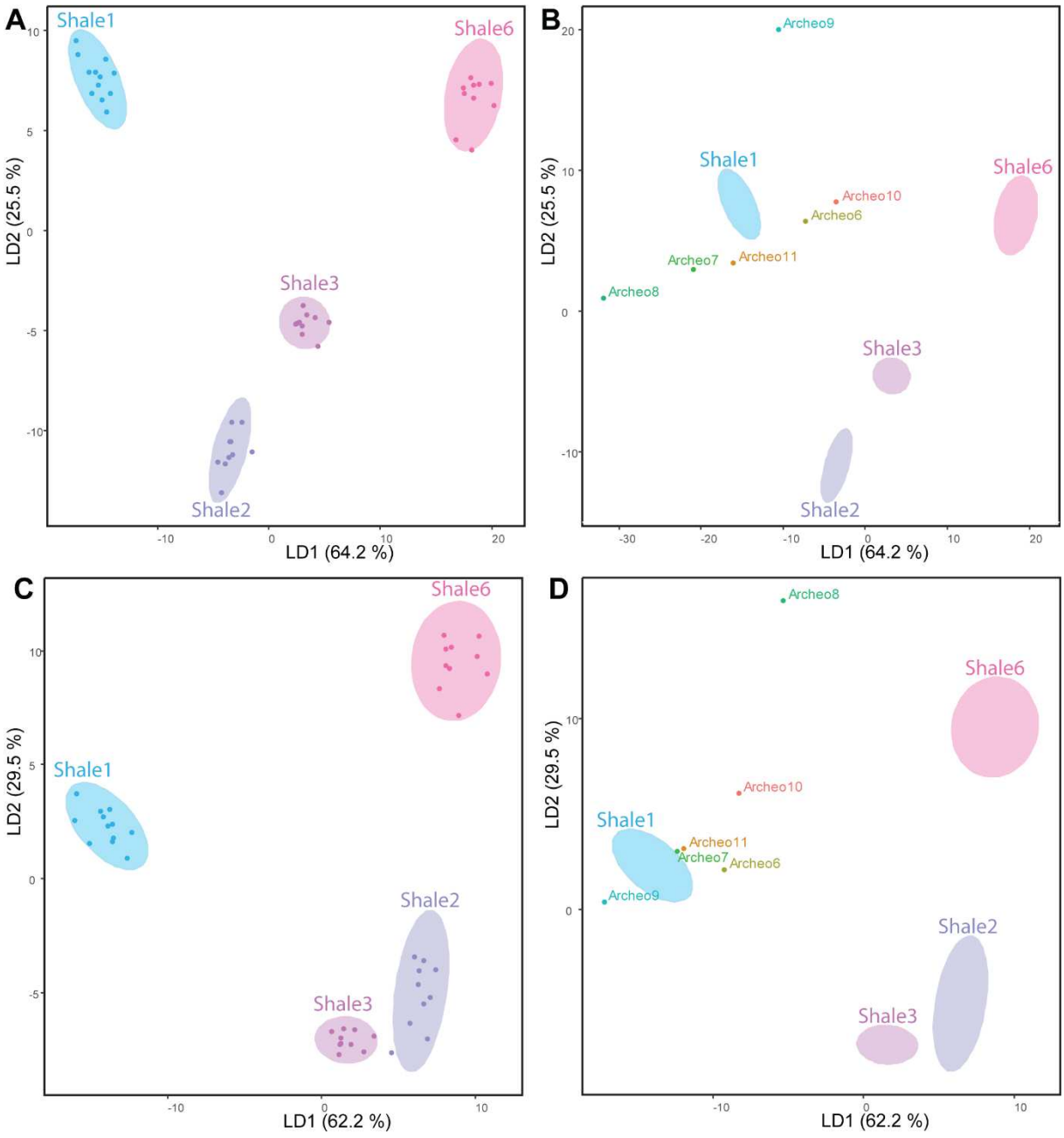
557 **Fig. 9 – Results of LDA6 carried out by considering the sources of shale and ferricrete**
558 **from the same locality separated with the trace elements' set3 (V, Cr, Co, As, Nb, Sb,**
559 **Ta, Ba, U) and alr ratios to Fe_2O_3 coordinates. A: LD1 and LD2 biplot. B. LD1 and LD3**
560 **biplot. C: LD1 and LD2 biplot with the archaeological pieces. D: LD1 and LD3 biplot**
561 **with the archaeological pieces. Squares: ferricrete pieces; Circles: shale pieces.**
562 **Confidence ellipses with a 95 % of probability.**

563

564

565 **4.2.3 Outcrops of shale alone**

566 In a last step, if we consider that shale and ferricrete have distinct lithologies, it makes sense
567 to separate them in the statistical analysis. We can compare archaeological shale pieces to
568 archaeological shale sources only. This is the last statistical treatment based on LDA we can
569 make with our dataset. The results are very similar whatever the set of elements chosen
570 (LDA7 to 9). The sources are well separated on the LD1-LD2 biplots (Fig. 10 and 11). The
571 accuracy and kappa statistic fall between 0.97 and 1. However, none of the shale pieces
572 affiliated to Shale1 fall into the distribution of this group. As regard to the predictions, for
573 LD7 and 8 the six shale pieces are attributed to Shale1 instead of four. For the last analysis
574 (LDA9), only two shale pieces are attributed to Shale1.



575

576 **Fig. 10 – Results of the LDA carried out by considering the sources of shale alone. A:**
577 **LDA7 with the trace elements' set1 (MgO, K_2O, Al_2O_3, V, Cr, Fe_2O_3, Co, Ni, Ga, Ge, As,**
578 **Rb, Zr, In, Cs, Hf, Th, U) and alr ratios to Fe_2O_3 coordinates, LD1 and LD2 biplot. B:**
579 **LDA7 with archaeological pieces. C: LDA8 with the trace elements' set2 (V, Cr, Co, Ni,**
580 **Ga, Ge, As, Rb, In, Cs, Hf, Pb, Th, U) and alr ratios to Fe_2O_3 coordinates, LD1 and LD2**
581 **biplot. D: LDA8 with archaeological pieces. Confidence ellipses with a 95 % of**
582 **probability.**

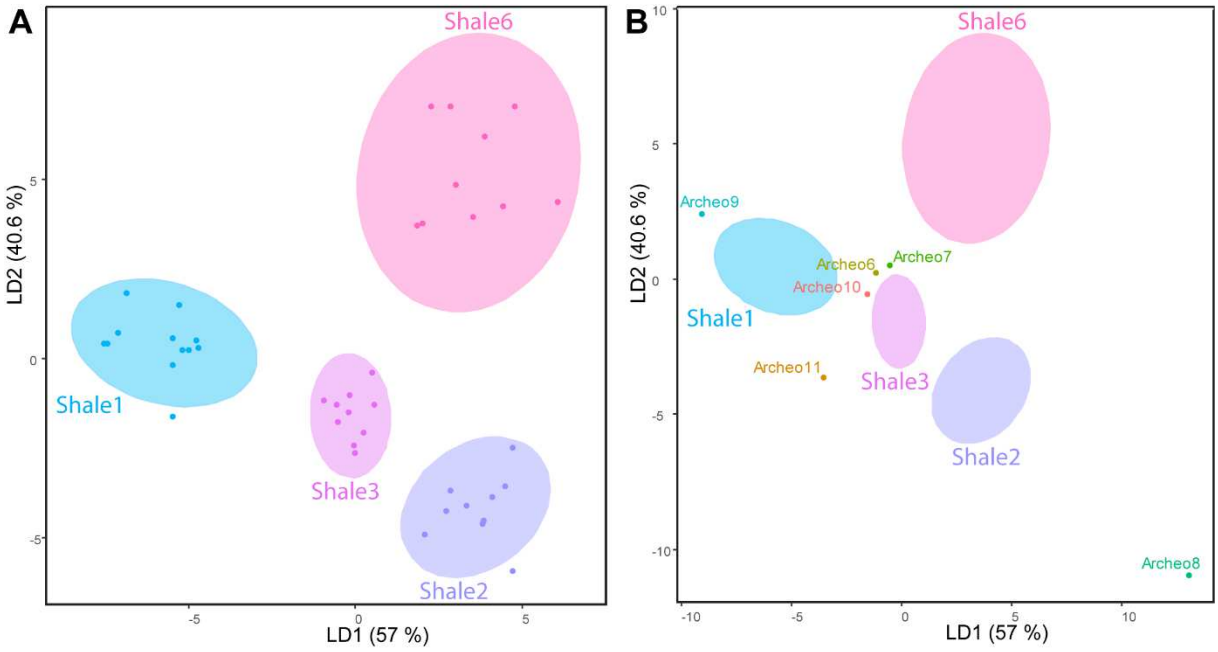


Fig. 11 – Results of LDA9 carried out by considering the sources of shale alone with the trace elements' set3 (V, Cr, Co, As, Nb, Sb, Ta, Ba, U) and alr ratios to Fe_2O_3 coordinates. A: LD1 and LD2 biplot. B. LD1 and LD2 biplot with archaeological samples. Confidence ellipses with a 95 % of probability.

5. Discussion

In our case study alr ratios to Fe_2O_3 coordinates appear to be more suitable for LDA statistical analyses than clr coordinates. Indeed, with the latter some variables were collinear despite our selection while they were not anymore with alr coordinates. Although graphically, the separation between the sources doesn't seem to be effective in some LDA treatments when considering LD1 and LD2, the accuracy and kappa statistic of all LDA models are good enough to consider that it allows a good discrimination amongst the sources, whatever the

599 sources are. The main difference between the treatments will be their aptitude to allow a
600 correct attribution of archaeological pieces.
601
602 As regard to the attribution of shale pieces to the Shale1 source, graphically a maximum of
603 three of the four samples that were assumed to come from this source do plot within the
604 distribution of Shale1 in the LDA: Archeo6, Archeo10, Archeo 11 (LDA3). The predictions
605 attribute the four pieces to the Shale1 group in all LDA models except for one (LDA9).
606 However, in some models, other shale or ferricrete samples are also attributed to it. The
607 predictions are not systematically the best way to assess the provenience of an archaeological
608 piece. It gives an indication but both graphical attribution and PCA results must be considered
609 too before reaching conclusions. In the PCA, sample Archeo7 and Archeo11 are always
610 attributed to Shale1. According to this result, along with the results of LDA3 and the
611 predictions, we can securely assess that all the four shale pieces previously attributed to
612 Shale1 are closely related to this source.
613
614

Statistical analysis	Samples	Data transformation	Groups *a priori*	Set of elements	Samples graphically attributed to Shale1	Separation of the ferricrete and the shale	Attribution of the ferricretes
PCA1	All the sources	alr	None	Set1	Archeo7, Archeo10	Separated	Near Ferr2a and Ferr2b
PCA2	Shale only	clr	None	Set2	Archeo7, Archeo10	-	-
PCA3	Shale only	alr	None	Set3	Archeo7, Archeo10	-	-
LD1	All the sources	alr	Shale6-Ferr1 and Ferr2a-Ferr2b grouped together	Set1	Archeo11	-	Dispersed
LD2	All the sources	alr	Shale6-Ferr1 and Ferr2a-Ferr2b grouped together	Set2	Archeo6	-	Dispersed
LD3	All the sources	alr	Shale6-Ferr1 and Ferr2a-Ferr2b grouped together	Set3	Archeo6, Archeo10, Archeo11	-	Near Ferr2 and Shale6-Ferr1
LD4	All the sources	alr	Shale6 Ferr1 and Ferr2a Ferr2b separated	Set1	None	Separated	Dispersed
LD5	All the sources	alr	Shale6 Ferr1 and Ferr2a Ferr2b separated	Set2	Archeo11	Separated	Ferricrete area
LD6	All the sources	alr	Shale6 Ferr1 and Ferr2a Ferr2b separated	Set3	Archeo6, Archeo10	Separated	Near Ferr2b
LD7	Shale only	alr	Shale separated	Set1	None	-	-
LD8	Shale only	alr	Shale separated	Set2	None	-	-
LD9	Shale only	alr	Shale separated	Set3	None	-	-

Table 3 – Synthesis of the results obtained with the different statistical analyses.

As regard to the separation of shale and ferricrete sources, the results of the LDA done when separating Shale6 and Ferr2a and the PCA done with all the sources are conclusive. The treatment allows a clear distinction between the ferricrete sources on one hand and the shale sources on the other (LDA4 to 6, PCA1). The archaeological ferricrete pieces are correctly attributed to the ferricrete area in the LDA done on set2 and set3 (LDA5 and 6) and on PCA1 done on set 1. However, the predictions of the LDA indicate an attribution to a ferricrete source (Ferr2b) for all archaeological ferricrete pieces in only one LDA, the LDA done when separating the sources of shale and ferricrete with set2 (LDA5). It is likely that we did not sample the outcrops the ferricrete samples come from. These sources may no longer be available nowadays. However, there seems to be a relation between the archaeological ferricrete pieces and the sources sampled, especially Ferr2b. The archaeological pieces may come from the same formations, namely tertiary to quaternary soils associated with Malmesbury formations.

The LDA done on trace elements' set3 by considering the shale and ferricrete grouped together (LDA3) gives the best results in graphical attribution of the archaeological samples. Three of the four archaeological pieces whose lithology and mineralogy indicate a provenance from Shale1 fall into the distribution of this source (Table 3). Despite some differences in geochemical trends, the elemental fingerprint of the shale and ferricrete from the same weathering profile is partially similar. This similitude is enhanced by the LDA. It is, however, worth noticing that the LDA3 is the one with the lowest accuracy and kappa statistic in the kfold validation. The statistical model is less robust than for other LDA. It might be less robust because it increases the variability encountered at each source, decreasing the inter-source variability. This is not a bad thing in provenance researches, since we are interested in a realistic statistical model, that takes into account all the intra-source variability and that

641 allows samples with a slightly different composition than those collected from an outcrop (or
642 a region, or a geological formation) to be attributed to this group because it comes from this
643 outcrop (region, geological formation) in reality.
644
645 The results of the LDA done with the shale alone are very bad regarding archaeological pieces
646 attribution: no archaeological pieces fall into the distribution of Shale1. The predictions are
647 not good either, the two shale pieces with a structure and a mineralogy that is not compatible
648 with the ones of Shale1 are attributed to this source in two cases, while in the last case only
649 two of the four shale samples affiliated to Shale 1 are attributed to this source. When shale
650 sources are separated from ferricrete sources, maximum two of the archaeological shale
651 pieces fall in the distribution of Shale1. This is lower than when the ferricrete and shale
652 sources are grouped together. Separating the lithologies from a weathering profile when doing
653 a LDA is not the best option if we consider this case study. For PCA analyses, whatever the
654 treatment made, the use of alr or clr, the use of different sets of elements, always two
655 archaeological pieces are attributed to Shale 1, Archeo7 and Archeo11 (Table 3). This
656 consistency suggests this is a necessary preliminary step when using multivariate statistical
657 analyses in iron-beating rocks' provenance researches.
658
659 The results of the LDA are contrasted. If we are interested in raw material category,
660 separating the different lithologies of iron-bearing rocks makes more sense. The sources must
661 be defined based on lithological features and the rock types of the geological samples must be
662 the same than the ones of archaeological artefacts. The different horizon of a weathering
663 profile should be distinguished during the field work. But from a provenience point of view,
664 the separation is not an obligation. According to the LDA, there is a certain similarity in
665 composition between the shale and ferricrete from the same weathering profile. A weathering

profile can be defined as a source as a whole, a separation between the different lithologies from a particular locality is not a necessity. In this case study this might be due to the fact that the elements that characterize the weathering profiles are those that show the best consistency within Shale1. They may be immobile ions that remain with constant concentration or with equal proportion to Fe wherever the samples are taken within an outcrop or within a weathering profile. This might explain why samples from Shale1 are more often attributed to this source with the LDA model built with weathering profiles as a whole defined as sources. This could also be explained by the higher number of samples per source: by increasing the number of samples per source, we may have a better representation of their compositional variability, which allow a better attribution of unknown archaeological pieces. When defining a source as a weathering profile as a whole, nonetheless, the best results are obtained when using trace elements that are positively correlated with Fe (LDA3). This selection might guarantee that the trace elements used are consistent with the location of the samples. It is important to keep this in mind when doing LDA analyses for provenance researches without separating the different lithologies of iron-bearing rocks encountered at a weathering profile. This would also allow to address provenance issues of ochre deposited on archaeological artefacts or found in paintings by selecting the iron oxide fingerprint and thus removing the influence of the substrate.

Issues might occur in the definition of this set of elements positively correlated with Fe. When there are no elements positively correlated with Fe when all the sources are taken into account, a source-by-source approach should be privileged (MacDonald et al., 2018; Mauran et al., 2021). MacDonald and colleagues (2018) show that the elements positively correlated with Fe vary from a source to another. We did confront to such kind of issue and we decided to choose the set of elements associated to one category of raw material, the shale pieces.

691 Indeed, a higher number of elements were positively correlated with Fe when considering the
692 shale sources instead of the ferricrete sources. The results we obtain prove that we can
693 successfully select elements that are useful for provenance researches with this method. The
694 inconvenience of such approach is that it is dependent on the operator. In order to increase the
695 chances of success, different sets of trace elements can be tested as already proposed by some
696 authors (Mauran et al., 2021; Zipkin et al., 2017). In view of the results obtained in the
697 present work, the elements selected in these sets could be those positively correlated with Fe
698 when considering each source and each rock type individually.
699
700 This study highlights another trend for the definition of the sources. The source Shale2 and
701 Shale3 that come from the same geological formation have close distribution in most of the
702 multivariate analyses. This is in favour of a definition of sources as geological contexts and
703 not only as geographic locations. This is what we did with mineralogical data, sources were
704 grouped per geological formations. This could also be done with the geochemical data for
705 sedimentary and metamorphic formations of iron-bearing rocks such as shale and phyllitic
706 shale.
707

708 **6. Conclusion**

709 The definition of the sources in ochre provenance researches is of high interest. Weathering
710 profiles that often account for iron-bearing rock sources might be difficult to treat as single
711 sources given that different lithologies of ochre are represented. We have shown that source
712 definition influences the LDA in a significant way. The results of the LDA are more accurate
713 in archaeological pieces attribution when considering the different lithologies grouped
714 together instead of being separated. We demonstrated that most archaeological shale pieces
715 plot into the distribution of our source of reference, the source in the back of the shelter. In

our case study, nonetheless, this source considered as a reference for archaeological pieces attribution is not affected by the separation or grouping of different lithologies. We may wonder if the results would have been the same if we considered as a reference a source where different lithologies of iron-bearing rocks are present. Moreover, the attribution to a category of rock cannot be made when grouping the different lithologies. This study poses new guidelines for ochre provenance researches. The use of LDA is confirmed has an advantageous choice to deal with simplified weathering profiles although it must be considered with caution. The use of trace elements positively correlated to iron is preferable to any other sets of elements. We also showed that PCA results are more consistent than LDA results whatever the set of elements used. This is in favour of starting with PCAs before doing LDAs. These considerations will allow to draw a better picture of ochre supply strategies and routes of transport in the future. Further research should be undertaken in order to define to which extend the trends highlighted in this study are global trends in iron-bearing rocks' provenance studies.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability* (Chapman&Hall Ltd.).

Attard Montalto, N., Shortland, A., & Rogers, K. (2012). The provenancing of ochres from the Neolithic Temple Period in Malta. *Journal of Archaeological Science*, *39*(4), 1094-1102. https://doi.org/10.1016/j.jas.2011.12.010

Besaans, A. J., Visser, H. N., & Theron, J. N. (1973). *Geological Survey [South Africa] 1:250 000 geological series. 3218. Clanwilliam* (Government Printer) [Map].

740 Comas Cufí, M., & Thió i Fernández de Henestrosa, S. (2011). *CoDaPack 2.0 : A stand-*
741 *alone, Multi-platform Compositional Software* (J. J. Egozcue, R. Tolosana-Delgado, &
742 M. I. Ortego, Éds.). Universitat Politècnica de Catalunya. Centre Internacional de
743 Mètodes Numèrics en Enginyeria (CIMNE). https://dugi-
744 doc.udg.edu/handle/10256/13645
745 Dayet, L. (2021). Invasive and Non-Invasive Analyses of Ochre and Iron-Based Pigment Raw
746 Materials : A Methodological Perspective. *Minerals*, *11*(2), Article 2.
747 https://doi.org/10.3390/min11020210
748 Dayet, L., Le Bourdonnec, F.-X., Daniel, F., Porraz, G., & Texier, P.-J. (2016). Ochre
749 Provenance and Procurement Strategies During The Middle Stone Age at Diepkloof
750 Rock Shelter, South Africa. *Archaeometry*, *58*(5), 807-829.
751 https://doi.org/10.1111/arcm.12202
752 Dayet, L., Texier, P.-J., Daniel, F., & Porraz, G. (2013). Ochre resources from the Middle
753 Stone Age sequence of Diepkloof Rock Shelter, Western Cape, South Africa. *Journal*
754 *of Archaeological Science*, *40*(9), 3492-3505.
755 https://doi.org/10.1016/j.jas.2013.01.025
756 Eiselt, B. S., Dudgeon, J., Darling, J. A., Paucar, E. N., Glascock, M. D., & Woodson, M. K.
757 (2019). In-situ Sourcing of Hematite Paints on the Surface of Hohokam Red-on-Buff
758 Ceramics Using Laser Ablation – Inductively Coupled Plasma – Mass Spectrometry
759 (LA–ICP–MS) and Instrumental Neutron Activation Analysis. *Archaeometry*, *61*(2),
760 423-441. https://doi.org/10.1111/arcm.12427
761 Eiselt, B. S., Popelka-Filcoff, R. S., Darling, J. A., & Glascock, M. D. (2011). Hematite
762 sources and archaeological ochres from Hohokam and O'odham sites in central
763 Arizona : An experiment in type identification and characterization. *Journal of*
764 *Archaeological Science*, *38*(11), 3019-3028. https://doi.org/10.1016/j.jas.2011.06.030

765 Kiehn, A. V., Brook, G. A., Glascock, M. D., Dake, J. Z., Robbins, L. H., Campbell, A. C., &
766 Murphy, M. L. (2007). Fingerprinting Specular Hematite from Mines in Botswana,
767 Southern Africa. In *Archaeological Chemistry* (Vol. 968, p. 460-479). American
768 Chemical Society. https://doi.org/10.1021/bk-2007-0968.ch025
769 Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of*
770 *Statistical Software*, *28*, 1-26. https://doi.org/10.18637/jss.v028.i05
771 Lê, S., Josse, J., & Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis.
772 *Journal of Statistical Software*, *25*, 1-18. https://doi.org/10.18637/jss.v025.i01
773 MacDonald, B. L., Fox, W., Dubreuil, L., Beddard, J., & Pidruczny, A. (2018). Iron oxide
774 geochemistry in the Great Lakes Region (North America) : Implications for ochre
775 provenance studies. *Journal of Archaeological Science: Reports*, *19*, 476-490.
776 https://doi.org/10.1016/j.jasrep.2018.02.040
777 Macdonald, B. L., Hancock, R. G. V., Cannon, A., McNeill, F., Reimer, R., & Pidruczny, A.
778 (2013). Elemental Analysis of Ochre Outcrops in Southern British Columbia, Canada.
779 *Archaeometry*, *55*(6), 1020-1033. https://doi.org/10.1111/j.1475-4754.2012.00719.x
780 MacDonald, B. L., Hancock, R. G. V., Cannon, A., & Pidruczny, A. (2011). Geochemical
781 characterization of ochre from central coastal British Columbia, Canada. *Journal of*
782 *Archaeological Science*, *38*(12), 3620-3630. https://doi.org/10.1016/j.jas.2011.08.032
783 Mathis, F., Bodu, P., Dubreuil, O., & Salomon, H. (2014). PIXE identification of the
784 provenance of ferruginous rocks used by Neanderthals. *Nuclear Instruments and*
785 *Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*,
786 *331*, 275-279. https://doi.org/10.1016/j.nimb.2013.11.028
787 Mauran, G., Caron, B., Beck, L., Détroit, F., Noûs, C., Tombret, O., Pleurdeau, D., Bahain, J.-
788 J., & Lebon, M. (2022). Standardization procedure to provide a unified multi-method
789 elemental compositional dataset, application to ferruginous colouring matters from

Namibia. *Journal of Archaeological Science: Reports*, *43*, 103454. https://doi.org/10.1016/j.jasrep.2022.103454

Mauran, G., Caron, B., Détroit, F., Nankela, A., Bahain, J.-J., Pleurdeau, D., & Lebon, M. (2021). Data pretreatment and multivariate analyses for ochre sourcing : Application to Leopard Cave (Erongo, Namibia). *Journal of Archaeological Science: Reports*, *35*, 102757. https://doi.org/10.1016/j.jasrep.2020.102757

McGrath, J. R., MacDonald, B. L., & Stalla, D. (2022). Middle Stone Age mineral pigment procurement at Pinnacle Point 5–6 North, Western Cape province, South Africa. *Archaeometry*, *64*(1), 193-217. https://doi.org/10.1111/arcm.12694

Parkington, J. E., Rigaud, J.-Ph., Poggenpoel, C., Porraz, G., & Texier, P.-J. (2013). Introduction to the project and excavation of Diepkloof Rock Shelter (Western Cape, South Africa) : A view on the Middle Stone Age. *Journal of Archaeological Science*, *40*(9), 3369-3375. https://doi.org/10.1016/j.jas.2013.02.017

Pierce, D. E., Wright, P. J., & Popelka-Filcoff, R. S. (2020). Seeing red : An analysis of archeological hematite in east central Missouri. *Archaeological and Anthropological Sciences*, *12*(1), 23. https://doi.org/10.1007/s12520-019-00984-4

Popelka-Filcoff, R. S., Lenehan, C. E., Glascock, M. D., Bennett, J. W., Stopic, A., Quinton, J. S., Pring, A., & Walshe, K. (2012). Evaluation of relative comparator and k0-NAA for characterization of Aboriginal Australian ochre. *Journal of Radioanalytical and Nuclear Chemistry*, *291*(1), 19-24. https://doi.org/10.1007/s10967-011-1236-2

Popelka-Filcoff, R. S., Miksa, E. J., Robertson, J. D., Glascock, M. D., & Wallace, H. (2008). Elemental analysis and characterization of ochre sources from Southern Arizona. *Journal of Archaeological Science*, *35*(3), 752-762. https://doi.org/10.1016/j.jas.2007.05.018

814 Popelka-Filcoff, R. S., Robertson, J. D., Glascock, M. D., & Descantes, Ch. (2007). Trace
815 element characterization of ochre from geological sources. *Journal of Radioanalytical*
816 *and Nuclear Chemistry*, *272*(1), 17-27. https://doi.org/10.1007/s10967-006-6836-x
817 Popelka-Filcoff, R. S., & Zipkin, A. M. (2022). The archaeometry of ochre sensu lato : A
818 review. *Journal of Archaeological Science*, *137*, 105530.
819 https://doi.org/10.1016/j.jas.2021.105530
820 Porraz, G., Texier, P.-J., Archer, W., Piboule, M., Rigaud, J.-P., & Tribolo, C. (2013).
821 Technological successions in the Middle Stone Age sequence of Diepkloof Rock
822 Shelter, Western Cape, South Africa. *Journal of Archaeological Science*, *40*(9),
823 3376-3400. https://doi.org/10.1016/j.jas.2013.02.012
824 Roberts, D. L. (2003). *Age, genesis and significance of South African coastal belt silcretes:*
825 *Vol. Memoir 95* (Council for Geosciences, South Africa).
826 Saggerson, E. P., & Turner, L. M. (1995). *A review of metamorphism in the Republic of South*
827 *Africa and the kingdoms of Lesotho and Swaziland* (Council for Geoscience,
828 Geological Survey of South Africa).
829 Scadding, R., Winton, V., & Brown, V. (2015). An LA-ICP-MS trace element classification
830 of ochres in the Weld Range environ, Mid West region, Western Australia. *Journal of*
831 *Archaeological Science*, *54*, 300-312. https://doi.org/10.1016/j.jas.2014.11.017
832 Tarbska, J., Walanus, A., Ciesielczuk, J., Samek, L., & Dutkiewicz, E. (2008). *Ferruginous*
833 *Raw Material Sources for Palaeolithic in Poland (Central Europe) ? Provenance*
834 *Studies : Occurrence, Litostratigraphy and Application*. 9th Int. Conference on NDT
835 of Art, May 2008 (Art 2008), Jerusalem, Israel.
836 Texier, P.-J., Porraz, G., Parkington, J., Rigaud, J.-P., Poggenpoel, C., Miller, C., Tribolo, C.,
837 Cartwright, C., Coudenneau, A., Klein, R., Steele, T., & Verna, C. (2010). A
838 Howiesons Poort tradition of engraving ostrich eggshell containers dated to 60,000

years ago at Diepkloof Rock Shelter, South Africa. *Proceedings of the National Academy of Sciences*, *107*(14), 6180-6185. https://doi.org/10.1073/pnas.0913047107

Texier, P.-J., Porraz, G., Parkington, J., Rigaud, J.-P., Poggenpoel, C., & Tribolo, C. (2013). The context, form and significance of the MSA engraved ostrich eggshell collection from Diepkloof Rock Shelter, Western Cape, South Africa. *Journal of Archaeological Science*, *40*(9), 3412-3431. https://doi.org/10.1016/j.jas.2013.02.021

Tribolo, C., Mercier, N., Douville, E., Joron, J.-L., Reyss, J.-L., Rufer, D., Cantin, N., Lefrais, Y., Miller, C. E., Porraz, G., Parkington, J., Rigaud, J.-P., & Texier, P.-J. (2013). OSL and TL dating of the Middle Stone Age sequence at Diepkloof Rock Shelter (South Africa) : A clarification. *Journal of Archaeological Science*, *40*(9), 3401-3411. https://doi.org/10.1016/j.jas.2012.12.001

Velliky, E. C., Barbieri, A., Porr, M., Conard, N. J., & MacDonald, B. L. (2019). A preliminary study on ochre sources in Southwestern Germany and its potential for ochre provenance during the Upper Paleolithic. *Journal of Archaeological Science: Reports*, *27*, 101977. https://doi.org/10.1016/j.jasrep.2019.101977

Velliky, E. C., MacDonald, B. L., Porr, M., & Conard, N. J. (2021). First large-scale provenance study of pigments reveals new complex behavioural patterns during the Upper Palaeolithic of south-western Germany. *Archaeometry*, *63*(1), 173-193. https://doi.org/10.1111/arcm.12611

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S, Fourth edition* (Springer).

Weigand, P. C., Harbottle, G., & Sayre, E. V. (1977). Chapter 2 - Turquoise sources and source analysis : Mesoamerica and the Southwestern U.S.A. In T. K. Earle & J. E. Ericson (Éds.), *Exchange Systems in Prehistory* (p. 15-34). Academic Press. https://doi.org/10.1016/B978-0-12-227650-7.50008-0

Zarzycka, S. E., Surovell, T. A., Mackie, M. E., Pelton, S. R., Kelly, R. L., Goldberg, P., Dewey, J., & Kent, M. (2019). Long-distance transport of red ocher by Clovis foragers. *Journal of Archaeological Science: Reports*, *25*, 519-529. https://doi.org/10.1016/j.jasrep.2019.05.001

Zipkin, A. M., Ambrose, S. H., Hanchar, J. M., Piccoli, P. M., Brooks, A. S., & Anthony, E. Y. (2017). Elemental fingerprinting of Kenya Rift Valley ochre deposits for provenance studies of rock art and archaeological pigments. *Quaternary International*, *430*, 42-59. https://doi.org/10.1016/j.quaint.2016.08.032

Zipkin, A. M., Ambrose, S. H., Lundstrom, C. C., Bartov, G., Dwyer, A., & Taylor, A. H. (2020). Red Earth, Green Glass, and Compositional Data : A New Procedure for Solid-State Elemental Characterization, Source Discrimination, and Provenience Analysis of Ochres. *Journal of Archaeological Method and Theory*, *27*(4), 930-970. https://doi.org/10.1007/s10816-020-09448-9

Zipkin, A. M., Hanchar, J. M., Brooks, A. S., Grabowski, M. W., Thompson, J. C., & Gomani-Chindebvu, E. (2015). Ochre fingerprints : Distinguishing among Malawian mineral pigment sources with Homogenized Ochre Chip LA–ICPMS. *Archaeometry*, *57*(2), 297-317. https://doi.org/10.1111/arcm.12090