



**HAL**  
open science

# Determination of cultural areas based on medieval pottery using an original divisive hierarchical clustering method with geographical constraint (MapClust)

Lise Bellanger, Arthur Coulon, Philippe Husi

## ► To cite this version:

Lise Bellanger, Arthur Coulon, Philippe Husi. Determination of cultural areas based on medieval pottery using an original divisive hierarchical clustering method with geographical constraint (MapClust). Journal of Archaeological Science, 2021, 132, pp.105431. 10.1016/j.jas.2021.105431 . hal-03279539

**HAL Id: hal-03279539**

**<https://hal.science/hal-03279539>**

Submitted on 2 Aug 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1

---

2 **Determination of cultural areas based on medieval pottery**  
3 **using an original divisive hierarchical clustering method with**  
4 **geographical constraint (MapClust)**

---

5

6 Lise Bellanger<sup>1</sup>, Arthur Coulon<sup>2</sup> and Philippe Husi<sup>2</sup>

7

8 **Abstract**

9

10 Spatial representation based on the distribution of artefact, especially pottery, is widely used  
11 in archaeology. The questions raised are often related for the construction of economic or  
12 cultural areas, based on fabrics, style or types of vessels. The middle Loire Valley with the  
13 study of an important corpus of medieval pottery constitutes an essential framework for  
14 understanding this mechanisms over long time. The corpus of data collected and studied for  
15 over twenty years includes archaeological sites with reliable chrono-stratigraphic sets whose  
16 pottery assemblages are sufficiently large to respond to the problem posed.

17 The permanent increase of the corpus of data, especially since the last publication on this  
18 subject in 2013 for which the approach was purely empirical, requires the development of  
19 statistical tools adapted here to the spatial clustering analysis of the data. This article deals  
20 with a divisive hierarchical clustering method with geographical constraints: MapClust. It is  
21 based on spatial indicators called spatial patches. This statistical tool allows the  
22 archaeologist to visualize and analyse geographical patterns easily. The comparison of the  
23 present results with those of 2013 is a way to see the relevance of the method.

24

25 **Keywords:** archaeology, pottery, cultural areas, middle Loire Valley, divisive hierarchical  
26 clustering, geographical constraints

27

28

29

---

<sup>1</sup> Université de Nantes Laboratoire de Mathématiques Jean Leray – UMR CNRS 6629, 2 rue de la Houssinière  
BP 92208, 44322 Nantes Cedex 03, France

<sup>2</sup> CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires, 40 rue James Watt,  
ActiCampus 1, 37200 Tours, France

## 30 1 Introduction

31  
32 The objective of an exhaustive study of pottery is to answer socio-economic and cultural  
33 questions from one of the major archaeological sources for the study of domestic life in pre-  
34 industrial periods. Spatialization of archaeological material like pottery is widely used in  
35 archaeology. The questions raised are often related to the distribution of products from the  
36 workshops, to the supply of consumer centres, to the spatial distribution of particular types of  
37 vessels for the construction of economic or cultural areas. These illustrative maps often  
38 remain very descriptive in the presentation of distribution by intensity or size of points, arrows  
39 linking sites, juxtaposed areas; drawn according to the knowledge and feeling of the  
40 archaeologist. This mode of representation has been frequently used since the 1970s, a  
41 period when data were still few in number and from a limited number of sites (Orton, 1969;  
42 Hodger, 1981). The development of preventive (rescue) archaeology since the 1990s and  
43 the mass of accumulated data has largely changed previous historical interpretations (Husi,  
44 2010). This calls for the development of reflection on new analytical tools at the interface  
45 between statistics and geomatics. Progress in particular in cartographic tools has made it  
46 possible to improve the quality of maps without abandoning an often largely descriptive  
47 approach to spatialized data (Vince, 1993; Delage, 1998; Husi, 2003, Legoux, 2010; Husi,  
48 2013; Chabert 2019; Allan, 2020; Henigfeld; 2021 to be published). The aim here is not to  
49 question the results and interpretations of such studies, but to attempt to overcome some of  
50 the limits of the spatial approach by adopting a statistical approach based on data analysis.  
51 At the same time, data analysis methods (clustering methods; factor analysis) have been  
52 widely used and for a long time in archaeometry, especially - but not only - in ceramology for  
53 the chemical and petrographic characterisation of production (Madsen, 1988; Tyers and  
54 Orton, 1991; Baxter, 2006; Vince, 2011; Holmqvist et al., 2018; Dervin and Bocquet, 2020;  
55 Dilmann et al. 2020).

56  
57 Cluster analysis allows the construction of groups or clusters of homogeneous observations  
58 in a data corpus. A wide range of numerical approaches exists to detect clusters (Kaufman  
59 and Rousseeuw, 1990; Everitt et al., 2001; Aggarwa et al., 2014). Cluster techniques have  
60 been employed in numerous disciplines such as archaeology for example, to investigate the  
61 relationship between various types of artefact: see for example Everitt et al. (2001) for a brief  
62 description or Baxter (2006). The two most widely used clustering algorithms are partitional  
63 and hierarchical clustering. In this paper, we concentrate on hierarchical clustering whose  
64 approach consists in developing a binary tree-based data structure called the dendrogram.  
65 Hierarchical clustering techniques may be subdivided into agglomerative methods that  
66 aggregate iteratively the individuals into clusters and divisive methods that split up each  
67 cluster into smaller ones. Agglomerative and divisive clustering are complementary. In the  
68 literature, divisive methods are largely ignored. The main reason is the computational aspect:  
69 divisive clustering is more expensive in terms of computing time than agglomerative.  
70 Nevertheless, it does not need to generate the complete divisive hierarchy to analyse a top-  
71 level partition. The scarcity of methods using divisive algorithms led us to focus on them. A  
72 number of splitting procedures were designed in the past, the oldest one being by Williams  
73 and Lambert (1959) is said to be monothetic in the sense that object sets are split according  
74 to the values of only one variable. Another approach using all variables simultaneously is  
75 said to be polythetic, it has the advantage to only depend on inter object distances or  
76 dissimilarities. The method of Macnaughton-Smith et al. (1964) combine the advantages to

77 be divisive, polythetic and computationally manageable. It is implemented in DIANA (Divisive  
78 ANALysis) clustering algorithm available on R free software environment.

79

80 In this work, we propose a new constrained Divisive Hierarchical Clustering (DHC) method  
81 named MapClust that takes into account geographical constraints and we apply it on an  
82 archaeological pottery dataset from middle Loire Valley (France). To study the contribution of  
83 the constraint, we also present the results obtained with DIANA. The MapClust clustering  
84 method was developed in a project on medieval and modern pottery from the middle Loire  
85 Valley (Husi dir., 2003, 2013). This project, which began more than twenty years ago, is one  
86 of the first research projects at the national level to have tackled the question of the study of  
87 pottery over a long period of time (6th to 17th c.), in a collective and systemic way,  
88 particularly in the structuring and analysis of data. This research has made possible the  
89 development of the european information network on medieval and modern ceramics named  
90 ICERAMM. The website associated with this network is built as an online database  
91 referencing typological tools and information on pottery assemblages for France and Belgium  
92 (<http://iceramm.univ-tours.fr/>). This work, which led to the creation of an important corpus of  
93 pottery, has opened up a new field of investigation in statistical data analysis, a collaboration  
94 between archaeologists from the Laboratoire Archéologie et Territoires (UMR CITERES,  
95 CNRS/Université de Tours) and statisticians from the Laboratoire de Mathématiques Jean  
96 Leray (UMR 6629, CNRS/Université de Nantes). This long-standing collaboration has  
97 contributed to the development of archaeo-statistical methods adapted to voluminous  
98 artefact data. The statistical tools developed cover both the estimation of dating and the  
99 clustering of archaeological contexts for the construction of periodization or spatial studies.  
100 This research has resulted in several publications in international journals (Bellanger et al.,  
101 2006, 2008, 2015; 2021; Bellanger and Husi, 2012).

102

103 As the pottery corpus continues to grow with the addition of new sites, a purely descriptive  
104 approach to the spatialization of the results is insufficient for a rigorous analysis of the  
105 pottery data. This is why we developed MapClust, a clustering method to deal with data that  
106 are not homogeneous in their spatial distribution and often patchy, for which traditional  
107 spatial statistics tools like kriging are inadequate. MapClust makes it possible to answer the  
108 problems of clustering and spatialization of archaeological contexts associating (i) pottery  
109 data that compose them, (ii) geographical distance between sites. This approach is essential  
110 for the construction of cultural areas, in this case pottery areas, in accordance with the  
111 distribution of data and the greater or lesser geographical proximity of the sites. The  
112 constructed clusters correspond to geographical areas formed by archaeological sites (i)  
113 close in distance and (ii) similar in terms of technical, typological and stylistic criteria of the  
114 pottery data.

115

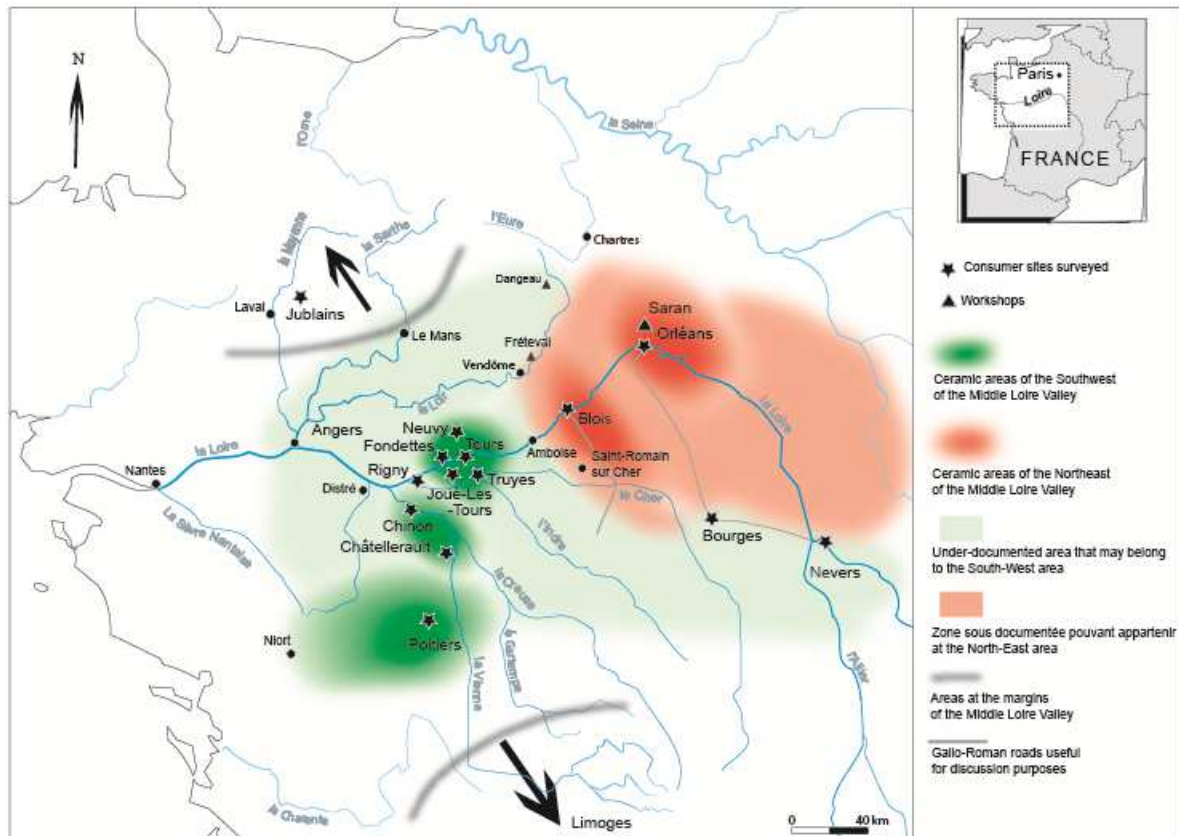
116 The paper is outlined as follows. Section 2 describes the archaeological background. Section  
117 3 presents the data corpus and the Mapclust clustering method. The results are shown in  
118 Section 4. Finally, we discuss the benefits of the novel clustering approach in Section 5.

119

## 120 2. Archaeological background: first results on cultural areas of the 121 Middle Loire Basin based on pottery

122 The first two parts of the research programme, completed in 2013, laid the foundations for a  
123 reflection on the transformation of cultural areas based on pottery over a long period of time  
124 and in an area with territorial coherence such as the Middle Loire basin (Husi dir., 2003,  
125 2013). The aim was already to better identify the permanent features and changes. Is the  
126 Middle Loire area composed of one or more economic or cultural entities? What are the  
127 pertinent spatial analysis scales to understanding changes in the area? How does this or  
128 these territorial networks change over time? What role does the Loire or its affluents play in  
129 the structuring of this area? If there is a middle Loire entity, what is its geographical limit and  
130 therefore its area of influence? These are all questions that pottery alone can only answer  
131 imperfectly, but which have the advantage of being an omnipresent source in archaeological  
132 sites, unlike other types of artefacts. Identifying economic and cultural spaces therefore  
133 requires a precise and quantified analysis of the data according to the technical, stylistical  
134 and functional traditions of pottery, including innovation, imitation and competition between  
135 potters. One way is to compare technical traditions, such glaze, painting, engobe, the  
136 presence of printing decorations or the choice of firing and therefore the colour of the wares.

137  
138 We present here the results acquired in 2013 which have guided the methodological  
139 developments of MapClust. We have therefore chosen to focus the presentation on the  
140 period between the 8th and 10th c. For this generally poorly documented period, pottery  
141 represent one of the best sources to understand the socio-economic and cultural  
142 mechanisms. The pottery corpus selected comes from excavations carried out in the main  
143 towns of the middle Loire Valley and some more minor sites. The sites are selected  
144 according to the typological quality of the pottery assemblages and the chronological quality  
145 of the stratigraphic set (domestic occupation, dumps...). Each stratigraphic set is included in  
146 an archaeological site located in a place (city, town...). The results of 2013 show a map at  
147 two scales (Fig. 1): (a) local economical areas around major consumption centres rarely  
148 exceeding 50 km around a production site or a major consumption centre ; (b) two larger  
149 cultural spaces (red and green on the map) with different manufacturing traditions. These  
150 two spaces are already distinguished by their pottery typology but also especially (i) in the  
151 southwest by a majority of white-beige ware with painted bands and glazed ware; (ii) in the  
152 northeast by a large presence of engobed ochre-red ware and the absence of glazed ware  
153 (Husi dir., 2013: 220-252). Although there may be small local evolutions between the 8th and  
154 10th centuries, these technical and decorative traditions are widely represented in North-  
155 West Europe during these three centuries. This observation justifies the choice of this time  
156 step to understand how the Loire area fits into a wider European cultural area (Husi, 2010).



157  
 158 Fig. 1: Map showing the distribution of areas based on pottery traditions on the scale of the  
 159 Middle Loire valley and local economic areas. (8th – 10th c.) (Husi dir. 2013 : 243).

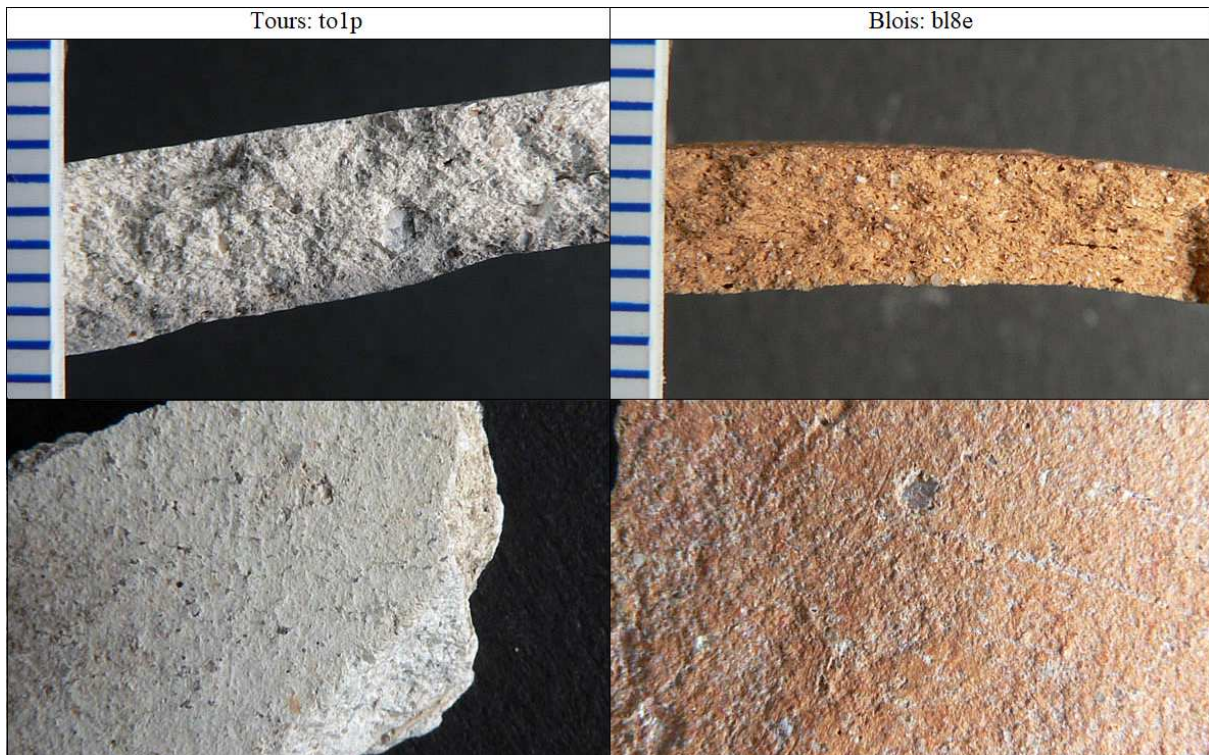
160  
 161 In 2013, the corpus studied for this period was 14 places, 47 sites, 214 stratigraphic sets and  
 162 about 13,000 vessels (Minimum Vessels Count); it is now 51 places, 87 sites, 333 sets or  
 163 about 25,000 vessels. The continuous augmentation of the corpus, especially the number of  
 164 sites and sets between 2013 and 2020, accentuates the spatial dispersion of the data at the  
 165 origin of the development of MapClust. This tool is a decision aid for the archaeologist in the  
 166 construction of cultural areas when the corpus becomes too large to interpret the data  
 167 without recourse to statistics.

### 168 3. Materials and methods

#### 169 3.1 Data collection

170 In this paper, which main aim is to illustrate the interest of the MapClust method, we have  
 171 chosen only two of the numerous technical and decorative traditions contained in the 2020  
 172 data corpus. We will present here the analyses of the spatial distribution of white-beige or  
 173 ochre-red pottery between the 8th and 10th c. mobilizing for this question 22 places out of  
 174 the 51 recorded for the Middle Loire basin (Fig. 2). This chronological choice is guided by the  
 175 changes that took place during this period, when light-coloured productions made in  
 176 oxidising firing replace those of dark black or grey colour in reducing firing.  
 177

178 We have chosen the technique of Minimum Vessels Count (MINVC) to quantify pottery by  
179 fabric (Orton and Tyers, 1992). It is the quantifying technique nearest to historical reality  
180 without, for example, problems of fragmentation rates such as the number of sherds.  
181



182  
183 Fig. 2. Example of fabrics : Tours white-beige (fabric to1p); Blois ochre-red (fabric bl8e)  
184

185 In this research work, we want to understand the transformation of cultural areas on the  
186 scale of the middle Loire Valley based on the evolution of pottery craft traditions. The  
187 manufacturing traditions corresponding to the pottery categories are defined as productions  
188 with identical or similar technical and stylistic characteristics. They come either from the  
189 same workshop, or from different workshops competing for the manufacture of fashionable  
190 pottery. In other words, the question is to identify spatially and over time the evolution of  
191 consumption habits, competition between products and potters' know-how.

### 192 3.2 Data analysis: a divisive hierarchical clustering approach with geographical 193 constraints

194 In this work, we focus on Divisive Hierarchical Clustering method (Kaufman and Rousseeuw,  
195 1990; Everitt et al., 2001; Aggarwa et al., 2014) to answer the problems of clustering and  
196 spatialization of archaeological contexts associating (i) pottery data that compose them, (ii)  
197 geographical distance between sites. The scarcity of methods using divisive algorithms led  
198 us to focus on them. Moreover, Macnaughton-Smith et al. (1964) argue that “divisive  
199 methods are safer than agglomerative ones, because “wrong” decisions in the early stages  
200 of an agglomerative analysis cannot be corrected later on, and one is mostly interested in the  
201 large clusters” (see Kaufman and Rousseeuw, 1990 page 273).  
202

203 We propose a new Divisive Hierarchical Clustering method named MapClust that takes into  
 204 account geographical constraints and compare it to the classical polythetic method DIANA  
 205 (Divisive ANALysis) (Macnaughton-Smith et al., 1964).

### 206 3.2.1 The proposed divisive hierarchical clustering method with geographical 207 constraints

208 MapClust is based on spatial indicators intended to capture spatial patterns of a set of  
 209 individuals (Woillez et al., 2009): the Centre of Gravity and the number of spatial patches.

#### 210 Centre of Gravity and number of spatial patches

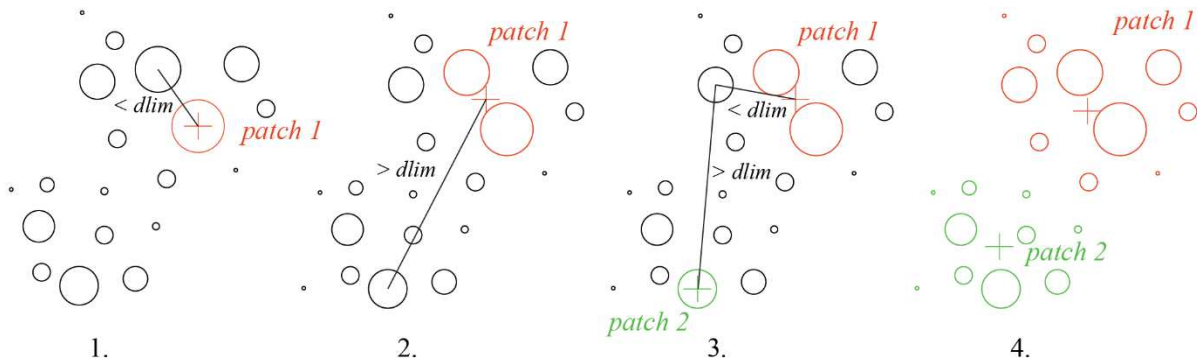
211 The centre of gravity  $CG$  is defined as the mean geographic location of the studied  
 212 individuals. Noting  $s_i = (x_i, y_i), i = 1, \dots, N$  the points in two-dimensional space and  $z(s_i)$  the  
 213 value of the regionalized variable measured in  $s_i$ ,  $CG$  is defined by:

$$214 \quad CG = \frac{\sum_{i=1}^N s_i z(s_i)}{\sum_{i=1}^N z(s_i)} \quad (Eq. 1)$$

215  
 216 To make sense, the regionalized variable of interest  $z$  in Eq. 1 must be positive such as for  
 217 example a frequency, a count or a probability density value. In this work,  $z$  will be associated  
 218 to the Minimum Vessel Counts of the studied stratigraphic set (see appendix B).

219  
 220 As the spatial distribution of the data may be heterogeneous and present local aggregations  
 221 or patches (e.g. archaeological pottery data), Woillez et al. (2007) proposed an algorithm to  
 222 identify them “by attributing each sample to the nearest patch, with respect to a maximal  
 223 threshold distance to its  $CG$ ” (see Eq. 1). A point  $s_i$  is assigned to a patch according to its  
 224 value  $z(s_i)$  and its distance to other existing patches. The position of a patch is then  
 225 determined by its  $CG$ . The algorithm starts with the largest value  $\max_i z(s_i)$  and then  
 226 considers each observation  $i$  in descending order of  $z(s_i)$  value. The highest value initiates  
 227 the first patch (Fig. 3.1). Then, the considered observation is assigned to the nearest patch,  
 228 provided that its distance to the patch's  $CG$  is smaller than a fixed threshold distance (noted  
 229  $dlim$ : limit distance) (Fig. 3.2). Otherwise, the observation forms a new patch (Fig. 3.3).

230



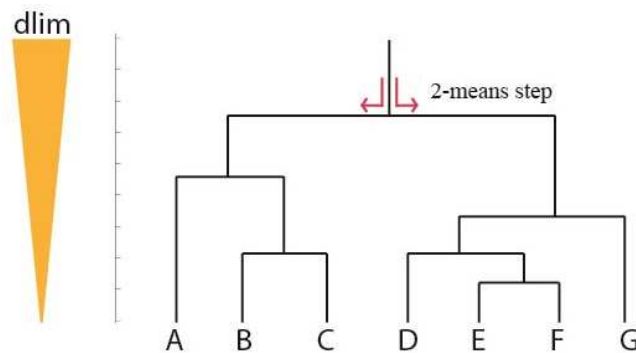
231  
 232 Fig. 3. Main steps of spatial patches detection algorithm for a fixed threshold distance  $dlim$ .  
 233 Simple case of 2 patches.  
 234



235 The number of detected spatial patches is of course influenced by the choice of the *dlim*  
236 threshold and the location of the highest *z* values: the smaller *dlim* is, the more patches will  
237 be built. Based on this observation, we have designed a divisive clustering procedure with  
238 geographical constraints by varying *dlim* in the algorithm written to identify patches (Woillez  
239 et al., 2007).

## 240 MapClust principle

241 In order to better understand the nesting of geographical spaces, the previous algorithm  
242 used to determine the number of spatial patches is repeated in a divisive hierarchical  
243 approach by decreasing the maximum acceptable threshold distance *dlim* (between  
244 observation points and the *CG* of a patch) until the first split in two. This approach leads to a  
245 divisive clustering approach (Everitt, 2001) based at each step on a k-means algorithm with  
246  $k = 2$ : each node (parent cluster) of the hierarchy is splitting to obtain a bipartition of two  
247 clusters (child node) until only one observation per cluster is obtained (see Appendix A for  
248 algorithm details). The result can be visualized as a dendrogram, which shows the  
249 hierarchical divisions performed at each step of the process. For MapClust clustering  
250 procedure, the lengths of the steams (heights) of the dendrogram represent the *dlim* values  
251 at which clusters are split up (Fig. 4).  
252



253  
254 Fig. 4. Dendrogram for MapClust clustering method.  
255

256 As we are not interested by the complete hierarchy but only in one selected partition,  
257 obtained from it, we must select one of the solutions in the nested sequences of clustering  
258 that comprise the hierarchy. This is equivalent to cut the dendrogram at a particular height to  
259 determine the number of clusters to be retained (sometimes called best cut). Large changes  
260 in division levels are taken to indicate the best cut. In this work, two methods based on the  
261 optimization of a criteria have been used to determination of the optimal number of clusters  
262 in the data: elbow method (with the total Within cluster Sums of Squares (WSS) criterion to  
263 minimize over a range of possible values for the number of clusters, see appendix C for  
264 details) and silhouette method (with the average silhouette width criterion to maximize over a  
265 range of possible values for the number of clusters, see appendix C for details). In addition,  
266 the archaeologist's expertise is called upon to select from the number of possible clusters the  
267 one that provides the most archaeologically interpretable clusters.  
268

### 269 3.2.2 MapClust clustering process for non positive or multivariate regionalized 270 variable

271 If the variable of regionalized interest  $z$  does not take positive values or is not univariate,  $CG$   
272 could not be defined as defined in (Eq. 1). However, these situations are not rare in practice.  
273 Therefore, to adapt and extend our method to these cases, we propose an approach inspired  
274 by the density-based clustering method proposed by Hartigan (1981). The  $N$  points  $z(s_i)$  are  
275 treated as a sample from a population with probability density function  $f$  which takes real  
276 values between 0 and 1.  $\hat{f}$ , a probability density estimate of  $f$  is obtained by the kernel  
277 method (Parzen, 1962), a non-parametric way to estimate  $f$ . Then  $z$  is replaced by  $\hat{f}(z)$  in  
278 Equation (Eq. 1) such that  $CG$  can be written:

$$279 \quad CG = \frac{\sum_{i=1}^N s_i \hat{f}(z(s_i))}{\sum_{i=1}^N \hat{f}(z(s_i))} \quad (Eq. 2)$$

280 The use of probability density estimation  $\hat{f}$  to compute  $CG$  in (Eq. 2) makes it possible to  
281 implement MapClust clustering algorithm (Appendix A) in non-positive or multidimensional  
282 variable of regionalized interest cases. It is also possible to (i) reduce the dimensionality of  
283 the data set using factorial analysis methods (Bellanger and Tomassone, 2014; Pagès,  
284 2014) like Principal Component Analysis (PCA) if variables are quantitative or  
285 Correspondence Analysis (CA) for non negative ratio scale data, (ii) apply the MapClust  
286 method using (Eq. 2) where  $\hat{f}$  is the estimated density of the chosen principal components.  
287 Factorial analysis allows to convert sets of higher-multidimensional feature into lower  
288 dimensional feature sets that preserve as much information as possible. The selection of the  
289 number of components depends on the data and must be chosen carefully. But it has no  
290 rigorous solution. This choice must allow the essential data to be extracted while eliminating  
291 less significant information during the clustering step.

292  
293 Finally, to study the contribution of the constraint, we also compare our results with MapClust  
294 to those obtained with DIANA clustering algorithm (Macnaughton-Smith et al., 1964). All  
295 statistical analyses are performed using the free software R (R Core Team, 2016).

## 296 4. Results

297 The aim is to compare with the use of MapClust the present results about culture areas with  
298 those of the 2013 map (Fig. 1), only partially here with the example of the white-beige and  
299 ochre-red pottery, all the criteria become part of a future larger publication on the subject in  
300 the Loire Valley (Husi dir., 2021). The difference with 2013 comes from a much larger site  
301 and pottery corpus and from a statistical and not simply descriptive approach of the data,  
302 central question of this paper.

### 303 4.1. Data description

304 The general corpus of 2020 is 51 places, 87 sites, 333 sets or about 25,000 vessel (§ 2). The  
305 reduced corpus of data mobilised to compare the distribution of white-beige and ochre-red  
306 pottery from the 8th to the 10th c. is structured by stratigraphic sets and pottery categories  
307 (see §3.1). It is defined in 22 places, 44 sites and 116 stratigraphic sets. The pottery data are  
308 quantified in Minimum Vessel Count (MINVC) by categories and represent a total corpus of

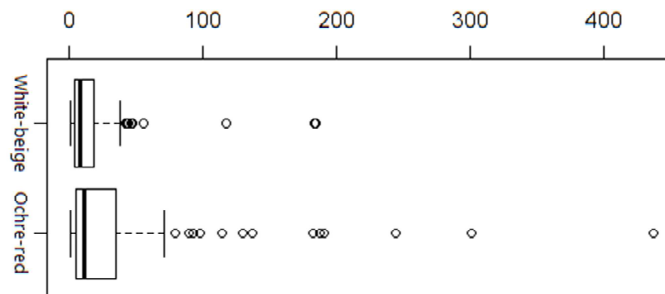
309 5668 vessels (Table 1). Among the 116 sets between 8th to the 10th centuries, there are 99  
 310 sets with white-beige production and 103 sets with ochre-red production. For the  
 311 multidimensional case 101 sets contain at least one of the productions in quantity satisfying  
 312 the condition of using a correspondence analysis (sum of all productions MINVC greater than  
 313 5 individuals).

314  
 315

Table 1. Description of white-beige and ochre-red categories of pottery.

	Sites	Stratigraphic sets	Total	Vessels (MINVC)		
				Median (per set)	Mean (per set)	Standard deviation (per set)
<b>White-beige</b>	30	99	1713	8	17.30	29.19
<b>Ochre-red</b>	35	103	3955	11	38.39	67.80

316



317  
 318  
 319  
 320

Fig. 5. MINVC for white-beige (top) and ochre-red (bottom) in the stratigraphic sets (99 for white-beige, 103 for ochre-red).

321 For both white-beige and ochre-red, half of the stratigraphic sets have less than 10 vessels  
 322 (Fig. 5, Table 1). The studied dataset also contains some stratigraphic sets with large  
 323 quantities of ochre-red pottery (10 sets with more than 100 vessels) and of white-beige  
 324 pottery (3 sets with more than 100 vessels). These sets explain the difference between the  
 325 median and the mean and the relatively high value of standard deviation (Table 1). It can be  
 326 noted that for a relatively similar number of sets (of which 86 are common) the number of  
 327 ochre-red pottery is twice as important as that of the white-beige production. The median is  
 328 close between the two types of production but not the mean value. The number of sets with  
 329 large quantities of ochre-red pottery is greater than those with large quantities of white-beige  
 330 production. As the total quantity of pottery is very different from one set to another, the raw  
 331 MINVC data for white-beige and ochre-red are unusable without transformation. Therefore in  
 332 all the following, for the unidimensional cases, we use variable of interest derived from  
 333 MINVC (Appendix B).

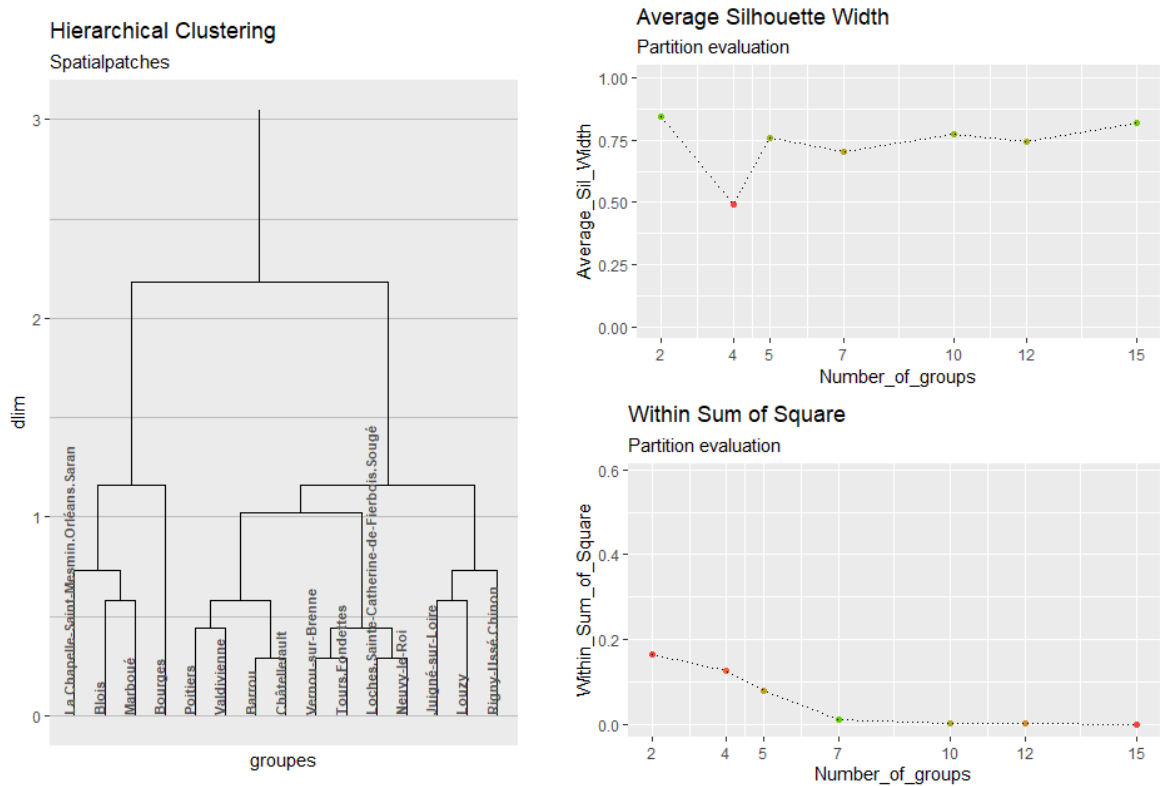
#### 334 4.2. From the data to the determination of homogeneous areas using 335 unidimensional information

336 The colours of the vessels depend on the preferences of the consumers but also on the  
 337 choice and the technical constraints of the potters. These constraints depend on the  
 338 characteristics of the clays, in our case on the importance of iron oxide for the red colour, but

339 also on the firing atmosphere. As the same clays are present all over this part of the middle  
 340 Loire Valley, their extraction is done locally according to the colour chosen by the potters.  
 341 Consequently, the spatial distribution based on the colour of the recipients is rather cultural,  
 342 linked to a stylistic choice, than dictated by potential technical constraints.  
 343 The objective is to first look at the ochre-red and white-beige data separately and secondly  
 344 performing a multidimensional analysis by integrating the two data sets.

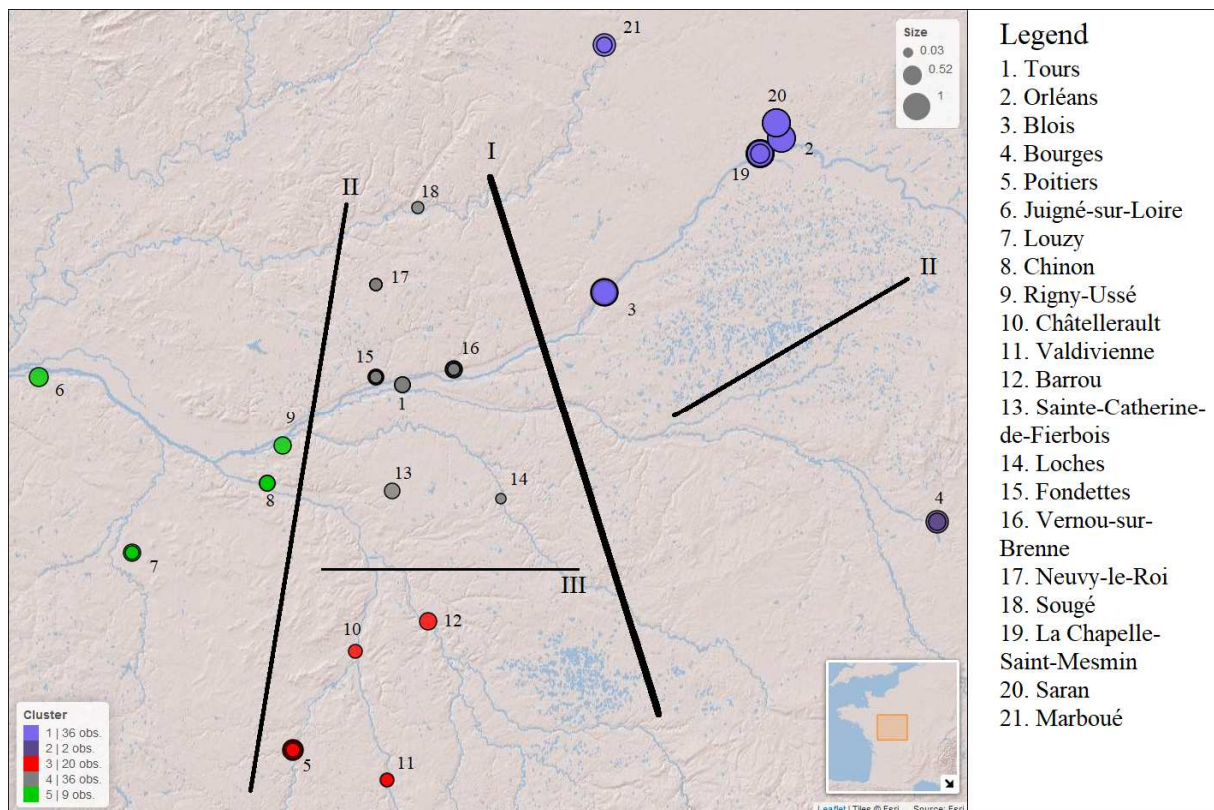
#### 345 4.2.1 Ochre-red productions between the 8th and the end of the 10th century

346 The aim here is to analyse the distribution of ochre-red pottery (see Fig. 1) by stratigraphic  
 347 set taking account of the geographical distance between places. Each place being  
 348 composed of several sets sharing the same coordinates, we have chosen to present the  
 349 results of MapClust clustering on places, the best scale for the definition of cultural areas.  
 350



351 Fig. 6. Ochre-red - Dendrogram obtained with MapClust (left).  
 352 Average silhouette width (top right) and WSS (bottom right).  
 353  
 354

355 Referring to the average silhouette criterion, calculate on weighted Euclidean distances  
 356 (Appendix C), the MapClust partitions are correct except for the one corresponding to a  
 357 partition into 4 clusters (Fig. 6). The WSS indicates that the optimal number of clusters can  
 358 be chosen between 5 and 7. The choice was made to retain the partition with 5 clusters, a  
 359 finer partition with more clusters having historically little meaning.

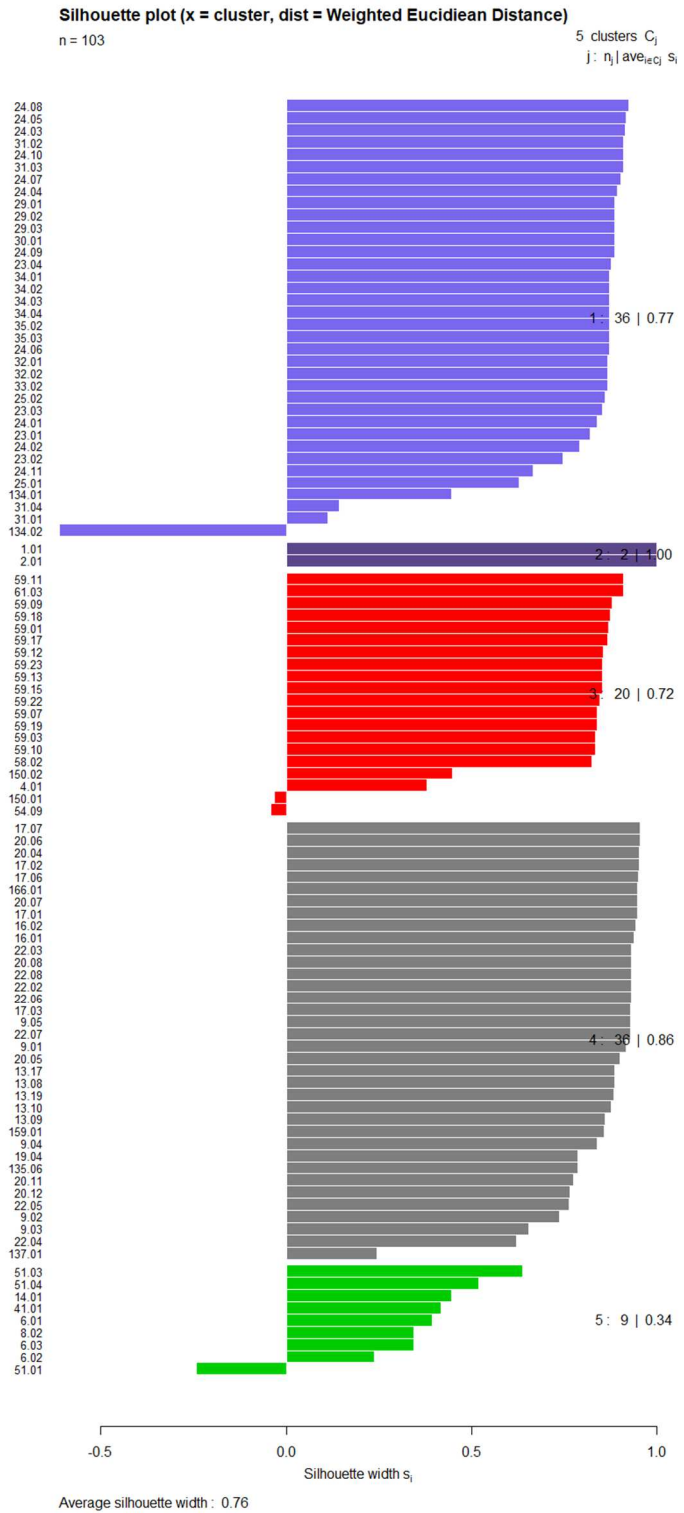


360  
361  
362  
363  
364  
365

Fig. 7. Ochre-red - 5 clusters partition (spaces) resulting from the MapClust algorithm on 103 sets. The lines give the hierarchy of partitions (I to III, because level 2 occurs in both sub cluster in the same time). The size of the circle increases with the frequency of the productions.

366  
367  
368  
369  
370  
371

The ochre-red colour productions between the end of the 8th c. and the 10th c. are divided into 5 clusters representing five areas, (Fig. 7). The off-centre geographical position of the Marboué and Louzy sites in relation to the core of the middle Loire Valley is related to their poor ranking according to the silhouette index. The first partition in two clusters separates a North-Eastern area of the middle Loire Valley, which includes the majority of ochre-red production, from a south-eastern area where they are generally less represented.



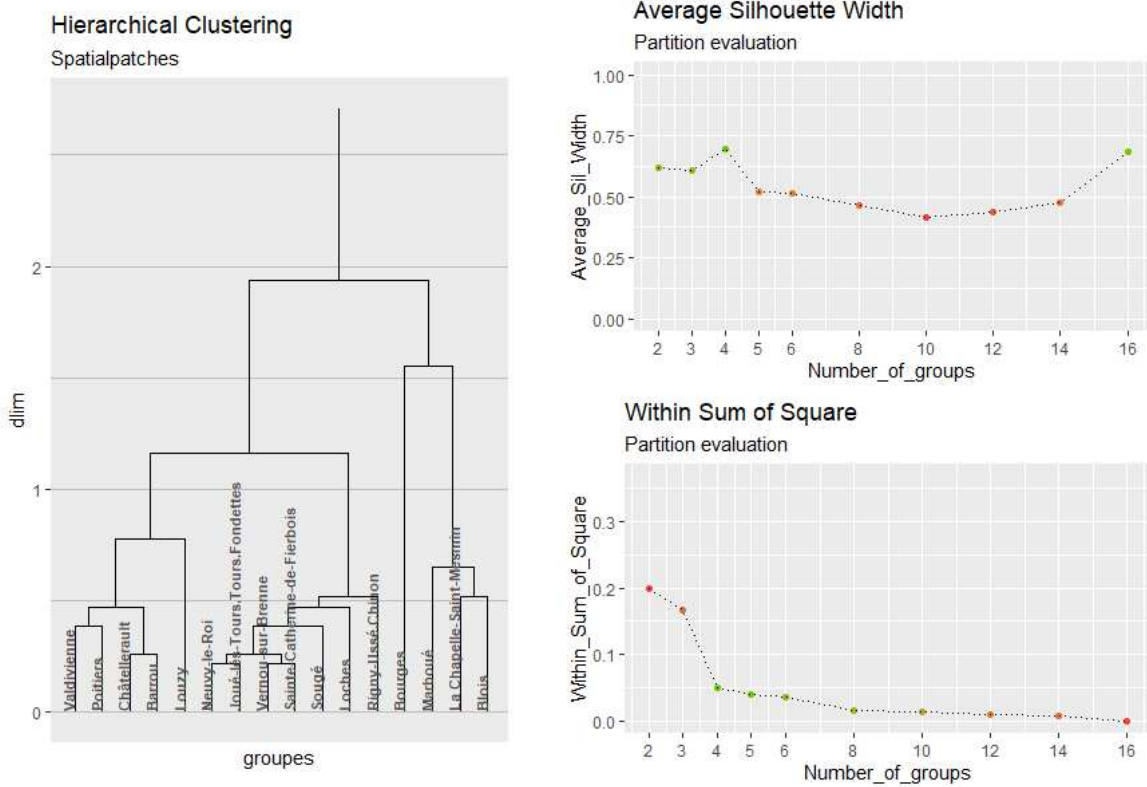
372  
 373  
 374  
 375  
 376  
 377  
 378  
 379  
 380

Fig. 8. Ochre-red - Silhouette index by MapClust cluster.

Two sites, from Marboué (134.02) and Louzy (51.01) are misclassified (Appendix C, silhouette index) (Fig. 8). The ochre-red productions are less significant at Marboué than at the other sites in the cluster 2, and the opposite is true for the Louzy site in the cluster 5. However, because of their geographical position, they are included in these clusters despite the fact that their pottery production is not similar. These sets are located in places at the margins of our space and are influenced by other production areas.

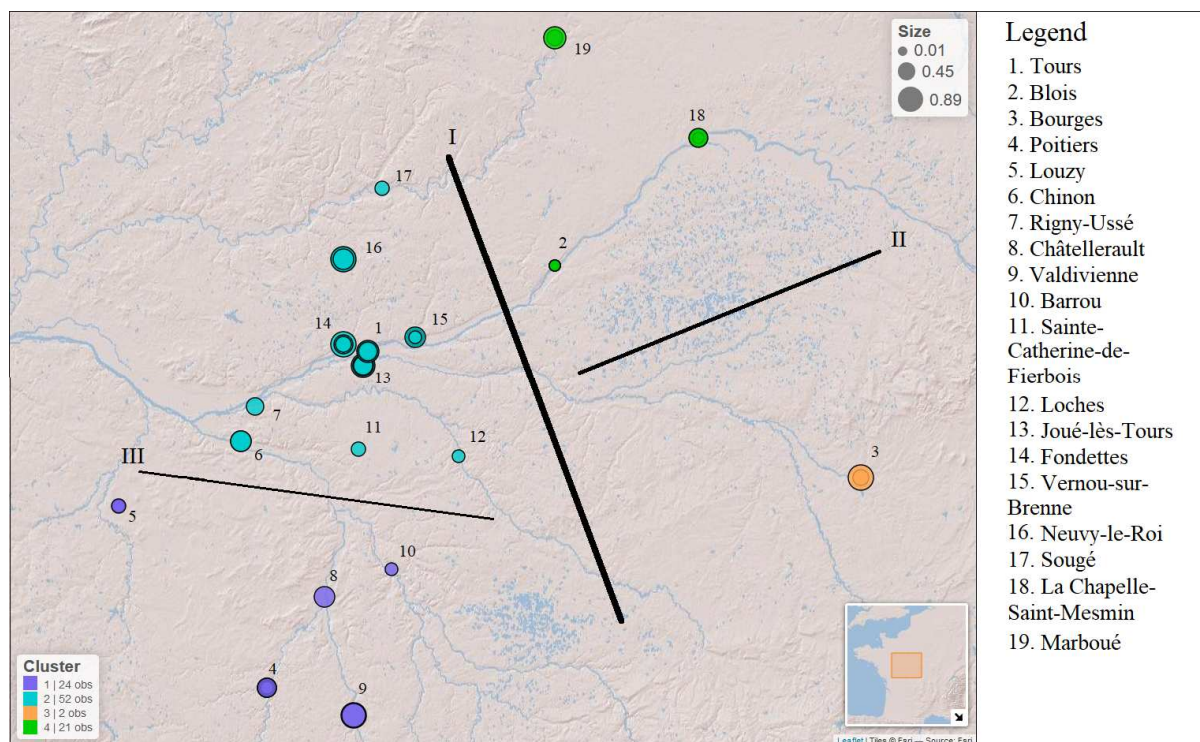
381 4.2.2 White-beige productions from the 8th to the end of the 10th century

382 As for ochre-red pottery, the objective is to analyse the distribution of white-beige pottery  
383 based on their quantified in Minimum Vessel Count by stratigraphic set and taking account of  
384 geographical distance between places. Each place is composed of several sets that share  
385 the same coordinates.  
386



387  
388 Fig. 9. White-beige - Dendrogram obtained with MapClust (Left).  
389 Average silhouette width (top right) and WSS (bottom right).

390 Referring to the average silhouette and WSS criteria, the best value corresponds to a  
391 partition into 4 clusters (Fig. 9).  
392  
393



394  
395  
396  
397  
398

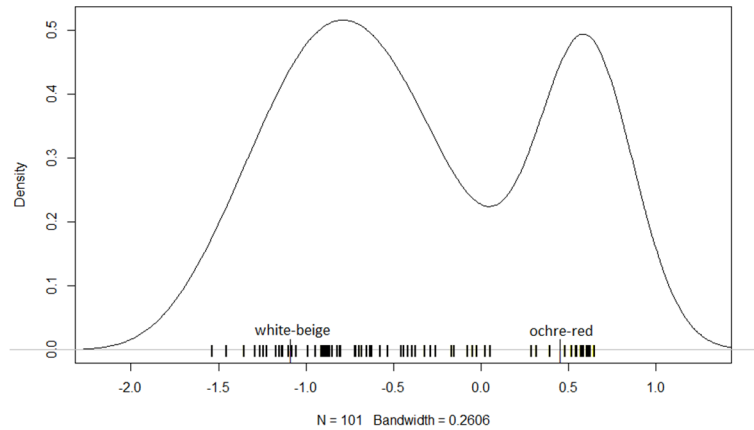
Fig. 10. White-beige - 4 clusters partition (spaces) resulting from the MapClust algorithm on 99 sets. The lines give the hierarchy of partitions (I to III). The size of the circle increases with the frequency of the productions.

399 The white-beige productions between the end of the 7th century and the 10th century are  
400 divided into four clusters representing four local economic areas (Fig. 10). The geographical  
401 position, off-centre of the Marboué site, revealing an economic area with little connection to  
402 the Loire, is certainly the cause of a poor allocation to clusters 4. The first partition in two  
403 clusters, which is also interesting, separates two areas, North-East and South-West,  
404 according to this manufacturing tradition: (i) the first with a marginal representation of these  
405 products in Blois and Orléans, (ii) the second which concentrates the majority of them  
406 between Haut-Poitou and Touraine. A set of the site of Marboué is badly clustered (Appendix  
407 D: 134.02). While white-beige products are well represented in this set, it belongs to the  
408 cluster 4, which is characterised by a low presence of such products. These set is located in  
409 a place at the margins of our space and is influenced by other production areas.

#### 410 4.3. Multidimensional analysis of the productions of white-beige and ochre-red 411 colour between the 8th and 10th century

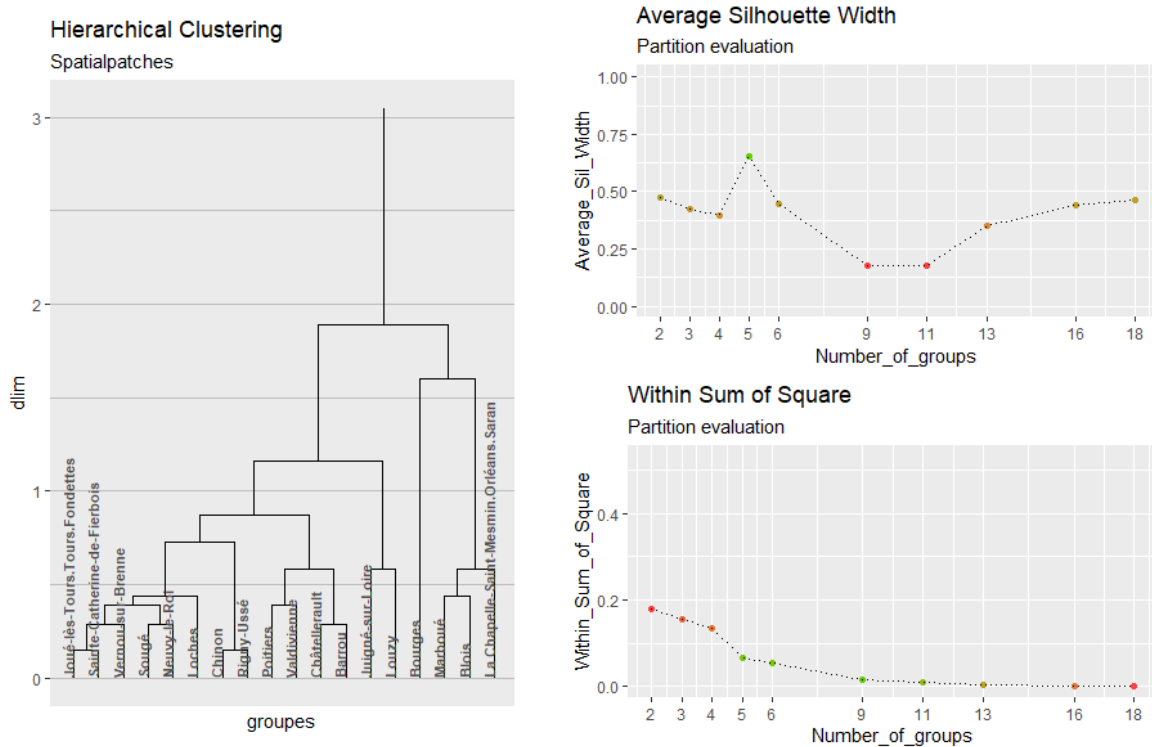
412 We consider here the same question of defining pottery areas based on the colour of the  
413 productions, but analysing simultaneously the two productions, white-beige and ochre-red.  
414 Correspondence Analysis (CA) (Benzécri, 1973; Greenacre, 2007) was performed on both  
415 variables to describe the structure of the dataset. We use  $\hat{f}$ , the probability density  
416 estimation of the unique principal component of the CA as the variable of interest to compute  
417  $CG$  in (Eq. 2) and obtain a partition with MapClust clustering method.





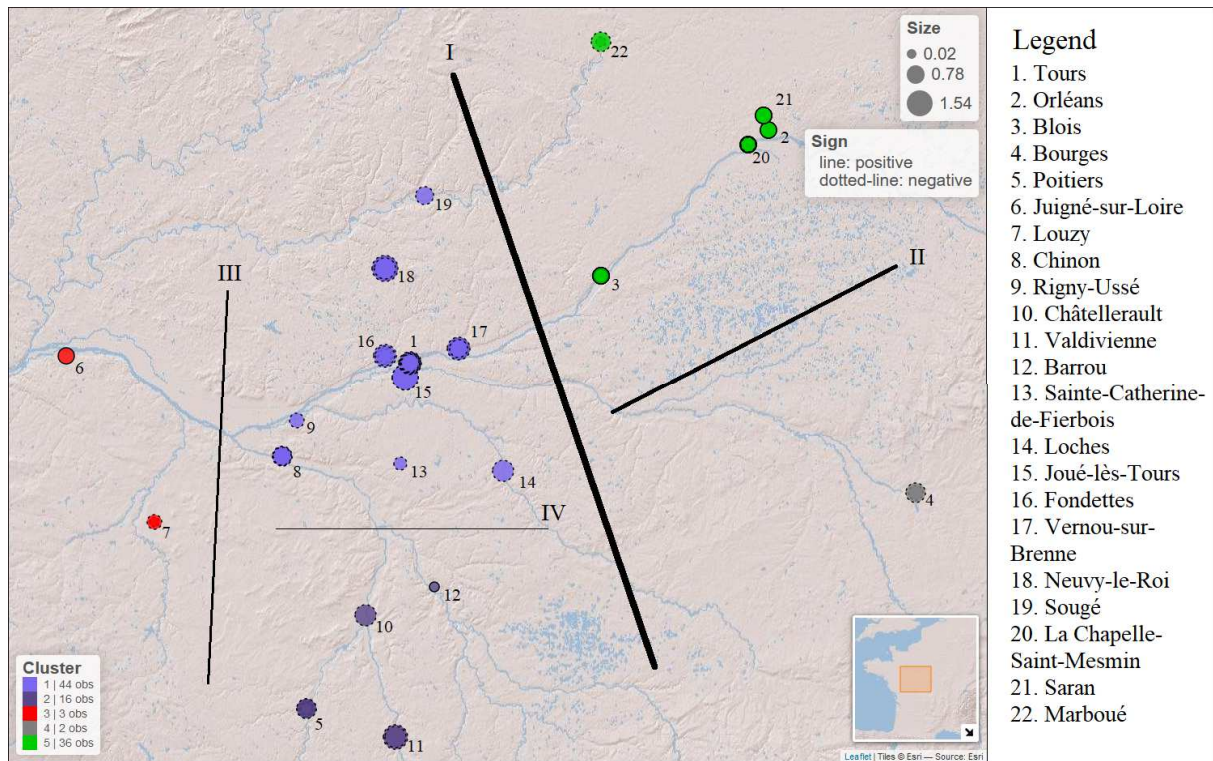
418  
 419 Fig. 11. Ochre-red & white-beige - Kernel density estimate of the only CA principal  
 420 component. Contributions of the two variables and the sets are plotted on the abscissa.  
 421

422 Only those sets and pottery categories with more than 5 individuals have been selected for  
 423 the CA. The ochre-red (negative values) and white-beige (positive values) colour productions  
 424 are opposed on the first axis of the CA, thus characterising two different profiles (Fig. 11).  
 425  
 426



427  
 428 Fig. 12. Ochre-red & white-beige - Dendrogram obtained MapClust (left).  
 429 Average silhouette widths (top right) and WSS (bottom right).  
 430

431 Referring to the average silhouette widths criterion, the best value correspond to the partition  
 432 into 5 clusters (Fig. 12). The WSS criterion confirms the choice of a 5-clusters partition.  
 433



434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459

Fig. 13. Ochre-red & white-beige - 5 clusters partition (spaces) resulting from the MapClust algorithm on 101 sets. The lines give the hierarchy of partitions (I to IV). The size of the circle increases with the frequency of the productions.

Dotted circles: places with sets containing a majority of white-beige productions; solid lines for ochre-red productions.

Such a partition highlights five areas based - when they are sufficiently informed - on the main consumption centres of the middle Loire Valley, of which the best attested are Touraine, Haut-Poitou and Blésois-Orléanais. Two areas in the far west and Berry, on the steps of the middle Loire Valley, are still insufficiently informed to be associated with one or other of the traditions (Fig. 13). The three sets, of Juigné, Chinon and Vernou-sur-Brenne are misclassified (Appendix E: respectively sets 41.01; 6.01; 22.06). White-beige production is less important at these sites of Chinon and Vernou-sur-Brenne than at the other sites of cluster 1, where this production dominates, and ochre-red production is in the majority at the Juigné site, in contrast to the other elements of cluster 3 which clearly lacks data. The first partition (I) of the dendrogram very quickly separates a South-Western space of white-beige tradition (dashed circles) and a North-Eastern space of ochre-red tradition (solid circles), the South-Western/North-Eastern.

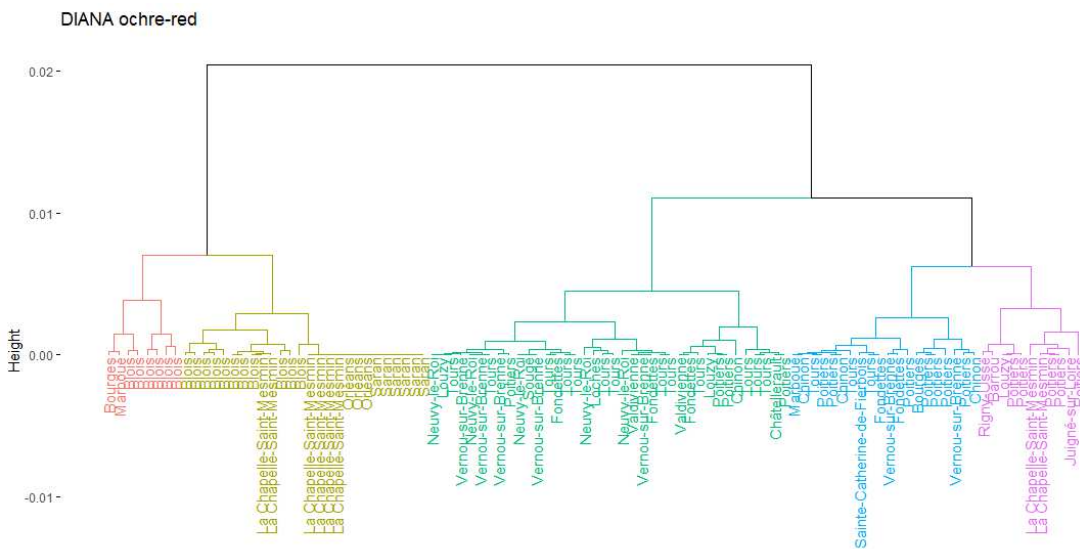
The multidimensional case has a slightly smaller average silhouette width (0.65, Appendix D) but is still good. We find the same overall clusters than previously. The multivariate analysis, together with the univariate analyses on each colour separately, allow to highlight the robustness of the results found on all these analyses (e.g. the first separation between North-East and South-West). Some partitions differences may be observed due to (i) particularity in the studied datasets, (ii) small amounts of MINVC for some stratigraphic sets

460 4.4 Comparison of MapClust and divisive hierarchical clustering DIANA results

461 In order to analyse the impact of the geographical constraint, we compared MapClust to the  
 462 polythetic Divisive hierarchical clustering method DIANA. To compare the agreement of the  
 463 partitions and the number of clusters fixed between MapClust and DIANA, we use: (i) the  
 464 confusion matrix (e.g. contingency table) that summarizes the overlap between the two  
 465 possible partitions, (ii) the Adjusted Rand Index (ARI) (Rand 1971) that measures the  
 466 similarity of objects present in the clusters: the closer the value is to 1, the better the  
 467 accuracy between the partitions is.

468 4.4.1 Ochre-red productions

469 The DIANA partition in 5 clusters shows a first separation between two geographical spaces  
 470 of the study area (South-Western/North-Eastern) (Fig. 14).



471 Fig. 14. Ochre-red - Dendrogram obtained with DIANA (5-clusters partition).  
 472  
 473

474 This result can be explained by a high production of ochre-red pottery in the North-Eastern  
 475 part. The cities of Blois and Bourges are separated in two clusters (the clusters 1 and 2 and  
 476 respectively the clusters 1 and 4) and it is globally difficult to identify geographical area with  
 477 DIANA method (Table 2).  
 478

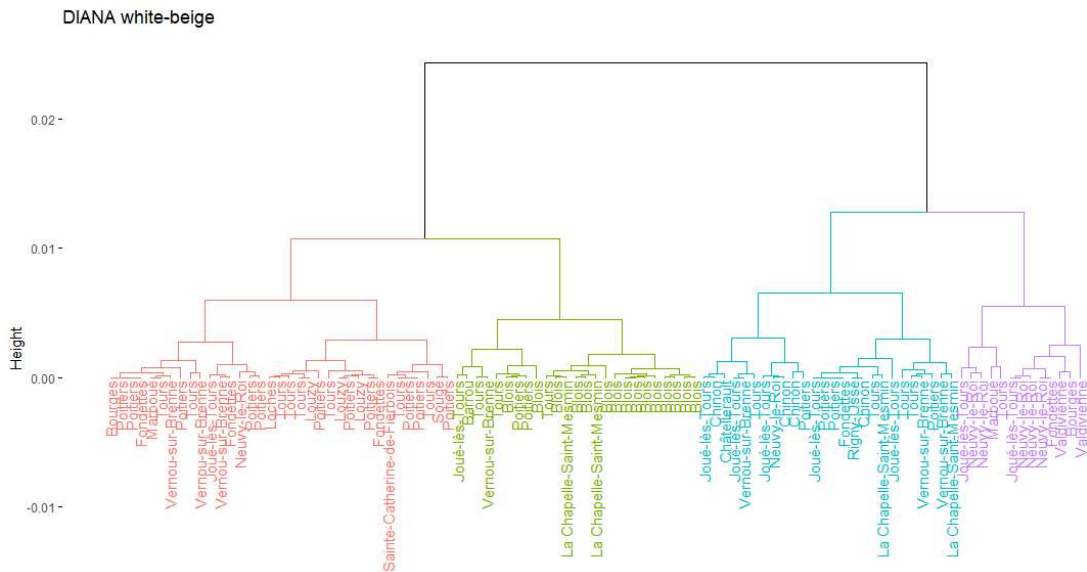
479 Table 2: Ochre-red - Confusion matrix between DIANA clusters and MapClust clusters.

Clusters		DIANA				
		1	2	3	4	5
MapClust	1	7	26	0	1	2
	2	1	0	0	1	0
	3	0	0	7	7	6
	4	0	0	28	8	1
	5	0	0	3	3	2

480 An ARI value of 0.43 indicates a mediocre agreement between the two partitions.

481 4.4.2 White-beige productions

482 DIANA results on white-beige dataset do not show a geographical separation of the study  
 483 area (Fig. 15) as for the ochre-red dataset.



484  
 485 Fig. 15. White-beige - Dendrogram obtained DIANA (4-clusters partition).  
 486

487 The clusters are spread over large areas. It seems that DIANA method does not work  
 488 properly for this type of data (Table 3).  
 489

490 Table 3: White-beige - Confusion matrix between DIANA clusters and MapClust.

Clusters		DIANA			
		1	2	3	4
MapClust	1	14	3	5	2
	2	19	5	19	9
	3	1	0	0	1
	4	1	17	2	1

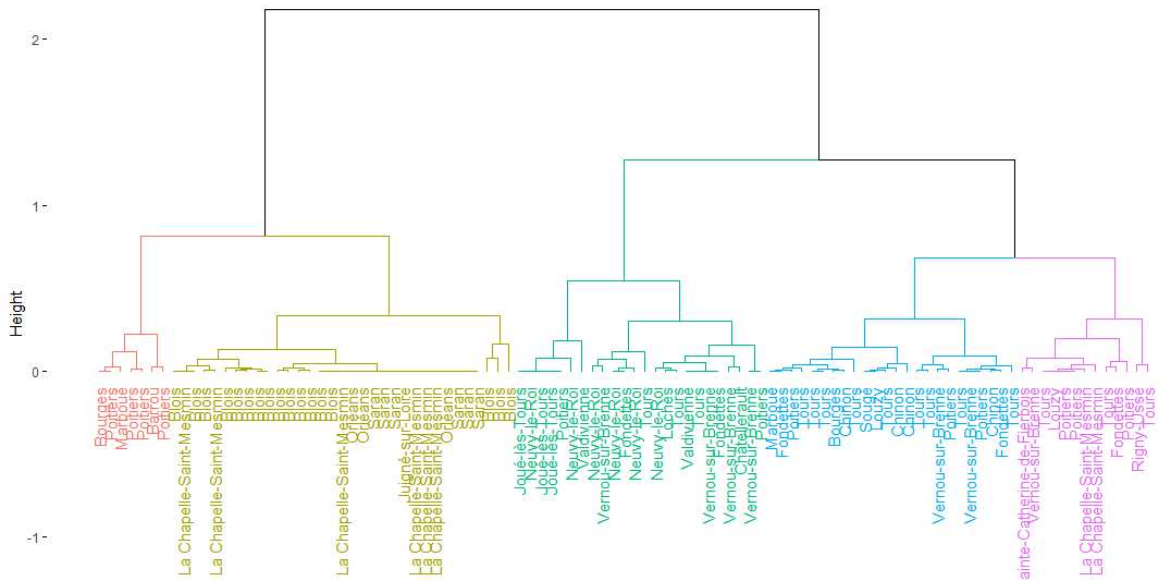
491 It yields ARI value as 0.14 that indicates a very bad agreement between the two partitions.  
 492

493 The ochre-red dataset show a result with the first separation on the dendrogram. Despite  
 494 this, the DIANA algorithm do not help for data with strong geographical context. This and the  
 495 very few results with the white-beige dataset show the necessary need of geographical  
 496 constraint in the context of univariate analysis.

497 4.4.3 White-beige and ochre-red productions

498 Starting from the contingency table of the two productions, we perform a correspondence  
 499 analysis and use the coordinates of the unique principal component to generate the distance  
 500 matrix. We then applied the DIANA clustering method (Fig. 16).

501



502  
503

Fig. 16. Ochre-red & white-beige - Dendrogram obtained DIANA (5-clusters partition).

504

The results are very similar to those we obtained with DIANA with the ochre-red dataset. The agreement between DIANA and Mapclust partition is very poor (Table 4; ARI 0.45). DIANA gives also the first separation between two geographical areas (South-West/North-East).

505

506

507

508

509

510

Table 4: Ochre-red & white-beige - Confusion matrix between DIANA clusters and MapClust.

Clusters		DIANA				
		1	2	3	4	5
MapClust	1	0	0	19	18	7
	2	5	0	5	3	3
	3	0	1	0	1	1
	4	1	0	0	1	0
	5	1	32	0	1	2

511

An ARI value of 0.45 indicates a very bad agreement between the two partitions

512

513

The contribution of geographical constraints is essential for certain types of data in order to help determine clusters. The case of archaeology, characterised by the distribution of data in semi points, requires a review of how to integrate these constraints.

514  
515

## 516 5. Discussion and conclusion

517

518

519

520

521

522

523

524

The spatial analysis of artefacts, particularly pottery, is generally realized only according to the intensity of the data, with a blurred boundary from one space to another, constructed without any real statistical processing. Moreover, in our case, archaeological data are measured at a limited number of locations. As a consequence, it is not possible to use geostatistical methods. Indeed, we choose to use simple spatial indicators without additional statistical assumptions to build our proposed clustering method. MapClust brings a new perspective to the archaeological problem of determination of cultural areas using pottery

525 since it takes into account the data to build the clusters but also the distance between sites.  
526 The application of MapClust clustering method on pottery data leads to results near similar to  
527 those already observed on an empirical study in 2013, confirming the relevance of a  
528 statistical approach applied to a dataset too large for a simple descriptive analysis. Overall,  
529 there is great stability in the local economic areas located around the main consumption  
530 centres in the Loire area. Whether MapClust is used separately on white-beige or ochre-red  
531 dataset or with both of them, we can observe: (i) three main areas, Haut-Poitou, Touraine  
532 and Blésois-Orléanais, (ii) one or two secondary areas more on the periphery to the extreme  
533 west with Anjou and to the east with Berry. There is a strong tradition of common ochre-red  
534 manufacturing for the Blésois and Orléanais, always in the same cluster at a low level of  
535 partition (IV) of the dendrogram (Fig. 6, 9 and 12). In the same vein, it is necessary to wait  
536 for the third partition (III) to differentiate Touraine from Haut-Poitou, this statistical proximity  
537 reflecting a strong common anchoring in the white-beige tradition. Whatever the analysis, the  
538 first division with MapClust results in a similar partition of Loire space in two pottery area  
539 opposing South-West and Nord-East. By looking more globally at the structure of the  
540 partition, the first MapClust partition split this Loire region into two distinct cultural areas. The  
541 white-beige productions corresponding mainly to the south-west of the middle Loire Valley  
542 (Touraine and Haut-Poitou). The ochre-red productions are concentrated in the North-  
543 Eastern part of the middle Loire Valley with Orléanais and Blésois. More generally, the use of  
544 MapClust with the other criteria chosen for the definition of cultural areas, but not presented  
545 here, still reveals this same South-West and Nord-East partition (Husi dir. 2021 forthcoming).  
546

547 From a statistical point of view, MapClust corresponds to the original divisive hierarchical  
548 clustering method with the addition of geographical constraints, which is new! Our choice of  
549 a hierarchical divisive clustering approach has many advantages over other approaches: (i) it  
550 is more efficient because it is often not necessary to generate the complete hierarchy, (ii) it is  
551 also more accurate than the agglomerative version because it takes into consideration the  
552 global distribution of data when making top-level partitioning decisions. Although clustering is  
553 regularly used in archaeology, in this work we present a new one, MapClust that is an  
554 original divisive hierarchical clustering method with geographical constraints providing a  
555 good way to answer the question of how to construct cultural areas from a large amount of  
556 data. The interest also comes from the fact that MapClust can be used for pottery but also for  
557 many other artefact. This methodological research is currently being pursued in a project of  
558 the French National Research Agency (ANR) entitled ModAThOM for "Explanatory Model of  
559 Urban Construction of Angkor Thom". This project includes the study of an important pottery  
560 corpus, but also other artefact resulting from many excavations made in the capital of the  
561 Khmer empire in Cambodia (<https://anr.fr/Projet-ANR-17-CE27-0008>). In addition, MapClust  
562 has easy-to-use tools to choose the number of clusters and to characterize geographically  
563 the created clusters. It does not require a lot of practice of the R language to run, thanks in  
564 particular to a web application Shiny more intuitive for users. The choice was made to  
565 develop both approaches, classical R package and web application in order to enlarge the  
566 target audience. This statistical method is integrated in a larger interdisciplinary project  
567 dedicated to the processing of data in archaeology, whose tools will be implemented an  
568 update of the R package: **Statistical P**attern **R**ecognition and **daT**ing using **A**rchaeological  
569 **A**rtifacts assemblage**S** (SPARTAAS, Coulon et al., 2021) available on the CRAN. In this  
570 study we present a new method, MapClust, designed to cluster archaeological data  
571 considering geographical constraints. It allows to keep a geographical consistency in the

572 cluster analysis. Moreover, in the future, this method could find applications in many other  
573 fields.

574

575

576

577

578

579 **Authors' contributions statement** P.H. and L.B. designed the research, P.H. collected the  
580 data, P.H and A.C. checked the data quality, A.C. developed the R code, L.B. and A.C.  
581 performed the data analyses, and P.H., L.B. and A.C. wrote the paper. All authors reviewed  
582 the manuscript.

583

584 **Declaration of competing interest** the authors have no competing interests.

585

586 *Acknowledgements*

587 This paper was supported by the project ANR ModAThom ([https://anr.fr/Projet-ANR-17-](https://anr.fr/Projet-ANR-17-CE27-0008)  
588 CE27-0008)

**Algorithm 1:** MapClust**1. Initialization**

- Root group initialization  $C = s_1, \dots, s_N$
- Hierarchy Initialization  $H_0 = C$
- Initialization of the first  $dlim$ : maximum distance from the point cloud
- Initialization of the step between two  $dlim$

**2. Preparation of variable z**

**if** several variables are available **then**

Dimensional reduction with correspondence analysis (use of the first axes of the results)

Or use an multivariate kernel density estimation (PARZEN, 1962).

**end if**

**if** the variable of interest  $z$  is not equivalent to a count **then**

The variable  $z$  is replaced by the kernel density estimates.

Transformation from  $z$  to  $f(z)$

**end if**

**3. Construction of the hierarchy**

**while**  $\langle dlim \rangle > 0$  **and** groups have more than 1 observation **> do**

Retrieving groups from the hierarchy at the previous level

Create  $H_{tmp}$ : the current hierarchy level

**for**  $\langle$  each group at the previous level  $\rangle$  **do**

Sort the observations in decreasing order of the variable of interest

Create group  $C_x^* = s_i$  where  $s_i = \operatorname{argmax} z = s_{(N)}$   $s_i \in C_x$   $s_i$  corresponds to the observation with the maximal value of  $z$

**for**  $\langle$  each observation  $\rangle$  **do**

Recalculate the Centre of Gravity of  $C_x^*$ ,  $CG = \frac{\sum_{i=1}^N s_i z(s_i)}{\sum_{i=1}^N z(s_i)}$

Calculate the distance  $D_i = d(s_i, CG)$  where  $d$  is the weighted eucliden distance defined as

$$d(i, j) = \sqrt{\left( (x_i - x_j)^2 + (y_i - y_j)^2 \right) * (z_i - z_j)^2}$$

**if**  $D_i < dlim$  **then**

$C_x^* = C_x^* \cup s_i$

$C_x = C_x \setminus C_x^*$

**end if**

**end for**

**for**  $\langle$  each observations  $s_i \rangle$  **do** try every observations without recalculate the centre of gravity

Calculate the distance  $D_i = d(s_i, CG)$  without recalculate  $CG$  or  $CG'$

Calculate the distance  $D'_i = d(s_i, CG')$  where  $CG'$  is the center of gravity of the group  $C_x$

**if**  $D_i > D'_i$  **then**

$C_x = C_x \cup s_i$

if the observation is closer to the centre of gravity of  $C_x$  then

$C_x^* = C_x^* \setminus C_x$

the observation returns to the group  $C_x$

**end if**

**end for**

Add  $C_x^*$  et  $C_x$  to the current hierarchy level  $H_{tmp}$

**end for**

Add  $H_{tmp}$  to the hierarchy  $H$

Delete  $H_{tmp}$

Reduce  $dlim$ : new  $dlim = dlim - step$

**end while**



591 Appendix B: MapClust regionalized variable of interest derived from  
 592 Minimum Vessel Count

593 Unidimensional variable of interest

594 For each unidimensional case, ochre-red and white-beige, we start with a contingency table  
 595 with all pottery categories as columns and all the sets concerned by the case as rows. Only  
 596 sets with a non-zero MINVC of the focus production are kept. MINVC information on pottery  
 597 is therefore contained in a contingency table  $\mathbf{N}$  of size  $103 \times J$  for the ochre-red case (resp.  
 598  $99 \times J$  for the white-beige case).  
 599

$$600 \quad \mathbf{N} = \begin{bmatrix} n_{1;1} & \cdots & n_{1;OR} & \cdots & n_{1;J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{103;1} & \cdots & n_{103;OR} & \cdots & n_{103;J} \end{bmatrix}$$

601  
 602 Thus  $n_{i;j}$  ( $i = 1, \dots, n_{OR} = 103$  (resp.  $n_{WB} = 99$ ) and  $j = 1, \dots, J$ ) represents the MINVC of  
 603 pottery category  $j$  found in set (e.g. context)  $i$  for the ochre-red case (resp. white-beige  
 604 case).  
 605

606 For the unidimensional analysis of ochre-red (resp. white-beige) productions, the variable of  
 607 interest  $z_{OR}$  (resp.  $z_{WB}$ ) in (Eq. 1) is defined for each set, as the MINVC of ochre-red pottery  
 608 (resp. white-beige) divided by the total MINVC found in the set  $i$  ( $\frac{n_{i;OR}}{\sum_{j=1}^J n_{i;j}}$ ), also divided by

609 the sum over all sets.  $z_{OR}$  (resp.  $z_{WB}$ ) is consequently given by

$$610 \quad 611 \quad z_{OR}(S_i) = \frac{n_{i;OR}}{\sum_{j=1}^J n_{i;j}} \bigg/ \sum_i \frac{n_{i;OR}}{\sum_{j=1}^J n_{i;j}} \quad (\text{resp. } z_{WB}(S_i) = \frac{n_{i;WB}}{\sum_{j=1}^J n_{i;j}} \bigg/ \sum_i \frac{n_{i;WB}}{\sum_{j=1}^J n_{i;j}})$$

613 MapClust algorithm (Appendix A) can then be applied on  $z_{OR}$  (resp.  $z_{WB}$ ).  
 614

615 Multidimensional variable of interest

616 For the multidimensional analysis of the productions of white-beige and ochre-red colour, A

617 correspondence analysis (CA) of the two columns matrix  $\mathbf{N} = \begin{bmatrix} n_{1;OR} & n_{1;WB} \\ \vdots & \vdots \\ n_{101;OR} & n_{101;WB} \end{bmatrix}$  (only sets

618 with more than 5 individuals in both pottery categories taken together are kept) makes it  
 619 possible to reduce the number of variables by constructing a unique synthetic variable (e.g.  
 620 principal component). Then,  $\hat{f}$ , the probability density estimation of the unique principal  
 621 component of the CA is used as the variable of interest to compute  $CG$  in (Eq. 2) and obtain  
 622 a partition with MapClust clustering method.  
 623  
 624

625 Appendix C: Number of cluster selection and quality evaluation of the  
626 partitions

627 Several criteria exist to determine the number of clusters to be selected. Two indices have  
628 been selected: the average of the silhouettes widths and the Within Sum of Square (WSS).  
629 The expertise of the archaeologist, derived from his knowledge of the nature of the context  
630 and the data, was also taken into account in determining the number of clusters.

631  
632 *The silhouette index* (Rousseeuw, 1987) is a widely used index for assessing the fit of  
633 individual observations in the clustering, as well as the quality of clusters and the entire  
634 partition. It is an index calculated for each observation. Let  $i$  be an observation belonging to  
635 cluster  $A$ , denote by  $C$  a cluster not containing  $i$ , the silhouette index is defined as:

636 
$$s_{sil}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

637 where  $a(i)$  is defined as the average dissimilarity between  $i$  and all other observations in  $A$ ,  
638 and  $b(i) = \min_{C \neq A} d(i, C)$  with  $d(i, C)$  is the average dissimilarity between  $i$  and all observations  
639 in  $C$ .  $b(i)$  is the mean dissimilarity to the observations of the nearest cluster.

640  
641 In our case, we use a weighted dissimilarity defined as:

642 
$$d(i, j) = \sqrt{\left( (x_i - x_j)^2 + (y_i - y_j)^2 \right) * (z_i - z_j)^2}$$

643 where  $x$  is the x-coordinate,  $y$  the y-coordinate and  $z$  the variable of interest.

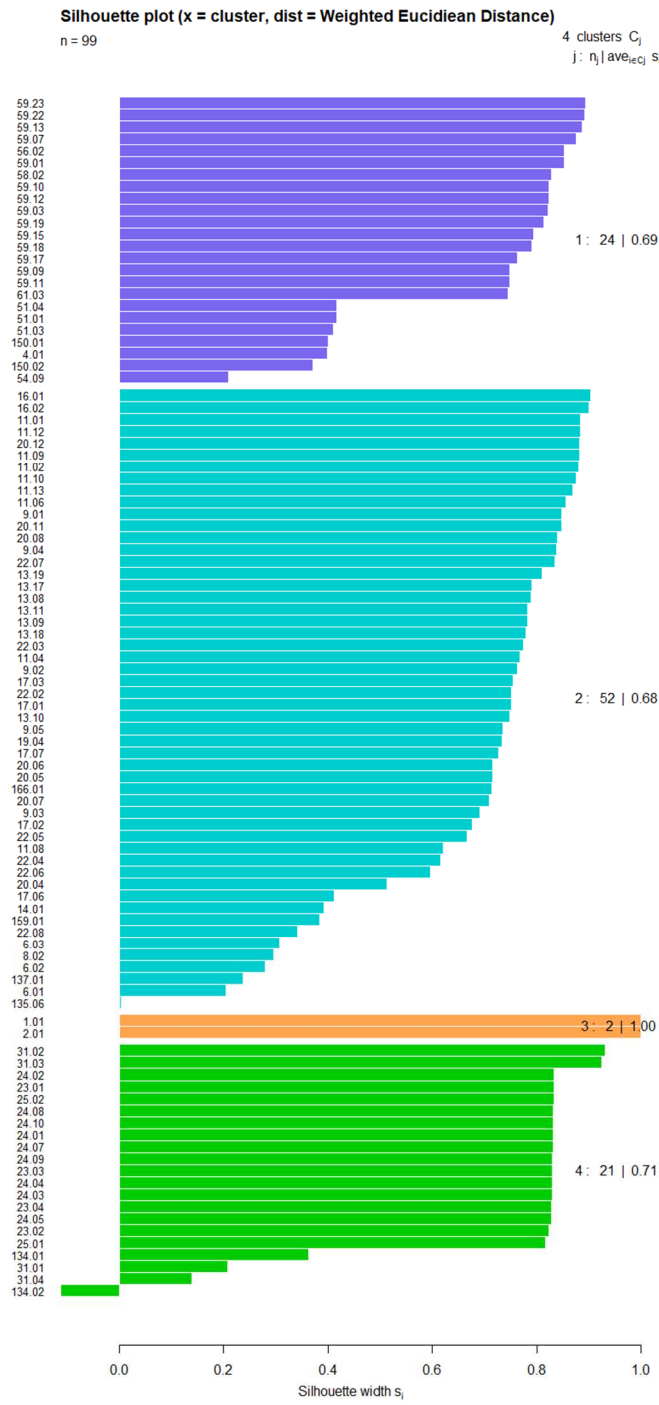
644  $s_{sil}(i)$  ranges between  $-1$  and  $1$ . A Value near  $1$  indicates that the observation  $i$  is  
645 appropriately clustered. If  $s(i)$  is close to  $0$ , the correct cluster for the observation  $i$  is  
646 doubtful.  $s(i)$  near  $-1$  indicates obvious misclustering, it would be more appropriate if  $i$  was  
647 clustered in its neighbouring cluster .

648  
649 The average of the silhouette indices defines the overall average of the silhouette widths in  
650 the different clusters that make up the score. To determine the number of clusters, the  
651 dendrogram is cut at different levels that provide different partitions and then the global  
652 average of the silhouettes widths is calculated for each possible partition. The partition that  
653 maximizes this quality index is retained.

654  
655 *Total Within cluster Sums of Squares* (WSS) measures the degree of homogeneity between  
656 belonging to the same cluster. The objective is to obtain clusters that are as homogeneous  
657 as possible and thus to make the variability between observations of the same cluster as low  
658 as possible. To determine the number of clusters to be selected, WSS is calculated  
659 according to the number of selected clusters and then the number of clusters is chosen so  
660 that adding another cluster does not improve (decrease) WSS much. In other words, after  
661 plotting the WSS vs. number of clusters curve, we observe the location of large reductions  
662 (elbow method) in WSS on the curve to determine a number of clusters.

663  
664

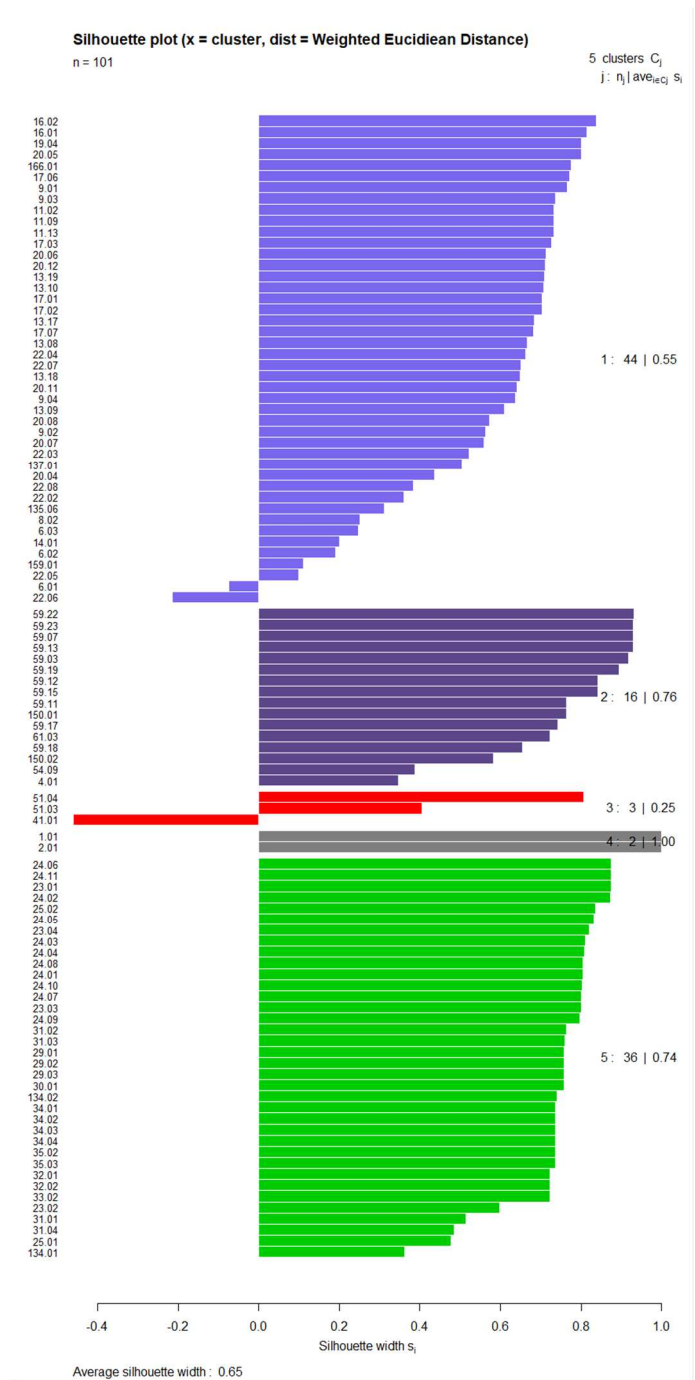
665 Appendix D: Silhouette plot from the white-beige productions MapClust  
 666 clustering  
 667



668  
 669  
 670

Fig. D.1. White-beige - Silhouette index by MapClust cluster.

671 Appendix E: Silhouette plot from the multidimensional MapClust  
 672 clustering



673

674

675

676

Fig. E.1. Ochre-red & white-beige - Silhouette index by MapClust cluster.

## 677 References

- 678 Aggarwal, C., Reddy, C., 2014. Data Clustering: Algorithms and Applications. Boca Raton:  
679 Chapman and Hall/CRC.
- 680 Allan, J., 2020. The importation of French Ceramics to South-West England, c AD 900-1600,  
681 In. Henigfeld Y., Husi P., Ravoire F. (eds.), 2020, L'objet au Moyen Âge et à l'époque  
682 moderne, Fabriquer, échanger, consommer et recycler, actes du XIe congrès  
683 international de la Société d'Archéologie médiévale, moderne et contemporaine  
684 (Bayeux, 28-30 mai 2015), CRAHAM, Presses Universitaires de Caen (PUC), Caen,  
685 p.456. : 181-196.
- 686 Baxter, M.J., 1994. *Exploratory multivariate analysis in archaeology*, Edinburgh University  
687 Press, Edinburgh.
- 688 Baxter, M.-J., 2006, A review of supervised and unsupervised pattern recognition in  
689 archaeometry, *Archaeometry* 48, 4, Oxford: 671- 6.4.
- 690 Bellanger, L., Tomassone, R., 2014. *Exploration de données et méthodes statistiques : Data*  
691 *analysis & Data mining avec R*. Collection Références Sciences, Editions Ellipses,  
692 Paris. 480 pages.
- 693 Bellanger, L., Husi, P., 2012. Statistical tool for dating and interpreting archaeological  
694 contexts using pottery, *Journal of Archaeological Science* 39, Issue 4, 777–790  
695 <https://doi.org/10.1016/j.jas.2011.06.031>
- 696 Bellanger, L., Tomassone, R., Husi, P., 2008. A statistical approach for dating archaeological  
697 contexts, *Journal of Data Science (JDS)* 6, n° 2, revue en ligne,  
698 <http://www.sinica.edu.tw/~jds/Content-v-6-2.html>
- 699 Bellanger, L., Husi, P., Tomassone, R., 2006. Statistical aspects of pottery quantification for  
700 dating some archaeological contexts, *Archaeometry* 48, Oxford, 169-183.
- 701 Bellanger, L., Husi, P., Laghzali, Y., 2015. Spatial statistic analysis of dating using pottery: an  
702 aid to the characterization of cultural areas in West Central France, in : Traviglia A. ed.,  
703 *Across Space and Time, Proceedings of the 41th International Conference on*  
704 *Computer Applications and Quantitative Methods in Archaeology (CAA-2013)*, Perth  
705 (Australie), Amsterdam University Press : 276-282.
- 706 Bellanger, L., Coulon, A., Husi, P., 2021. PerioClust: A Simple Hierarchical Agglomerative  
707 Clustering Approach Including Constraints. In: Chadjipadelis T., Lausen B., Markos A.,  
708 Lee T.R., Montanari A., Nugent R. (eds) *Data Analysis and Rationality in a Complex*  
709 *World*. IFCS 2019. Studies in Classification, Data Analysis, and Knowledge  
710 Organization. Springer, Cham. [https://doi.org/10.1007/978-3-030-60104-1\\_1](https://doi.org/10.1007/978-3-030-60104-1_1)
- 711 Benzecri, J.-P., 1973. L'analyse des Données. Volume II. L'analyse des Correspondances.  
712 Paris: Dunod.
- 713 Chabert, S., 2019. La céramique arverne Durant l'Antiquité tardive et le haut Moyen Âge (fin  
714 Ille – milieu VIIIe s.), approche chrono-typologique, économique et culturelle, 71eme  
715 Supplément à la revue Archéologique du Centre de la France, FERAC, Tours, p. 374.
- 716 Coulon, A., Bellanger, L., Husi, P., 2021. SPARTAAS: Statistical Pattern Recognition and  
717 daTing using Archaeological Artefacts assemblageS. R package version 1.0.0.  
718 <https://CRAN.R-project.org/package=SPARTAAS>.
- 719 Delage, R., 1998. Première approche de la diffusion des céramiques sigillées du Centre de  
720 la Gaule en occident romain, *SFECAG*, Actes du Congrès d'Istres : 271-314.
- 721 Dervin, S., Bocquet –Liénard, A., 2020. Les aires d'approvisionnement en céramique en  
722 Basse-Normandie de la fin du XIe siècle à la fin du Moyen Âge, In. Henigfeld Y., Husi  
723 P., Ravoire F. (eds.), 2020, L'objet au Moyen Âge et à l'époque moderne, Fabriquer,

724 échanger, consommer et recycler, actes du XIe congrès international de la Société  
725 d'Archéologie médiévale, moderne et contemporaine (Bayeux, 28-30 mai 2015),  
726 CRAHAM, Presses Universitaires de Caen (PUC), Caen, p.456. : 153-166.

727 Dillman, P., L'Héritier, M., Dissier, A., Leroy, S., Vega, E., 2020. Produire, échanger, utiliser  
728 les métaux ferreux au Moyen Âge : nouveaux éclairages interdisciplinaires In.  
729 Henigfeld Y., Husi P., Ravoire F. (eds.), 2020, L'objet au Moyen Âge et à l'époque  
730 moderne, Fabriquer, échanger, consommer et recycler, actes du XIe congrès  
731 international de la Société d'Archéologie médiévale, moderne et contemporaine  
732 (Bayeux, 28-30 mai 2015), CRAHAM, Presses Universitaires de Caen (PUC), Caen,  
733 p.456. : 197-210.

734 Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis. London UK: Arnold.

735 Greenacre, M.J., 2007. Correspondence Analysis in Practice. Chapman & Hall/CRC,  
736 BocaRaton, second edition.

737 Hartigan, J.A., 1981. Consistency of single linkage for high-density clusters. Journal of the  
738 American Statistical Association 76(374), 388–394.

739 Henigfeld, Y. dir., 2021, to be published. La céramique dans les pays de la Loire et en  
740 Bretagne de la fin du Xe au début du XVIIe siècle, Presses Universitaires de Rennes  
741 (PUC), collection Archéologie, Rennes.

742 Hodges, R., 1981. *The Hamwih pottery: the local and imported wares from 30 years'*  
743 *excavations at Middle Saxon Southampton and their European context*, cba research  
744 Report 37, London.

745 Holmqvist, E., Larsson, A.-M., Kriiska, A., Palonen, V., Pesonen, P., Mizohata, K., Kouki, P.,  
746 Räisänen, J., 2018. Tracing grog and pots to reveal Neolithic Corded Ware Culture  
747 contacts in the Baltic Sea region (SEM-EDS, PIXE), *Journal of Archaeological Science*  
748 91, 77-91, <https://doi.org/10.1016/j.jas.2017.12.009>.

749 Hurst, J. G., 1969. Red-painted and glazed pottery in Western Europe from the eighth to the  
750 twelfth C., *Medieval Archaeology* XIII, 92-146.

751 Husi, P. (dir.), 2003. *La céramique médiévale et moderne du Centre-Ouest de la France (11<sup>e</sup>*  
752 *– 17<sup>e</sup> siècle), chrono-typologie de la céramique et approvisionnement de la vallée de la*  
753 *Loire moyenne*, 20<sup>e</sup> supplément à la Revue Archéologique du Centre de la France,  
754 FERAC, Tours, 110 p. et un cédérom (Ouvrage imprimé épuisé) (Persée : Ouvrage en  
755 ligne : [www.persee.fr/issue/sracf\\_1159-7151\\_2003\\_ant\\_20\\_1](http://www.persee.fr/issue/sracf_1159-7151_2003_ant_20_1) ; Preuves en ligne :  
756 <http://citeres.univ-tours.fr/cera2003>).

757 Husi, P., 2010. Red-painted and glazed ware of the early medieval period in western  
758 France : new data for previous interpretations, an assessment for northwestern  
759 Europe, in : De Groote K., Tys D. Marnix P. (ed.), *Excavation Medieval Material*  
760 *Culture, Studies on archaeology and history presented to Frans Verhaeghe*, Relicta  
761 Monografieën, 4, Archeologie, Monumenten – en Landschapsonderzoek in  
762 Vlaanderen, Heritage Research in Flanders. : 81-91.

763 Husi, P. (dir.), 2013. *La céramique du haut Moyen Age dans le Centre-Ouest de la France :*  
764 *de la chrono-typologie aux aires culturelles*, 49<sup>e</sup> Supplément à la Revue Archéologique  
765 du Centre de la France, FERAC-ARCHEA, Tours, 283 p. (volume papier, et en ligne  
766 Persée : [https://www.persee.fr/issue/sracf\\_1159-7151\\_2013\\_arc\\_49\\_1](https://www.persee.fr/issue/sracf_1159-7151_2013_arc_49_1) ; Preuves en  
767 ligne : <http://citeres.univ-tours.fr/cera2013>).

768 Husi, P. (dir.), 2021 to be published. *La céramique médiévale et moderne du bassin de la*  
769 *Loire moyenne (Centre-ouest de la France), transformation des aires culturelles dans*  
770 *le temps long (6<sup>e</sup> – 19<sup>e</sup> s.)*, Supplément à la Revue Archéologique du Centre de la  
771 France, FERAC-ARCHEA, Tours, publication en ligne.

- 772 Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data - An Introduction to Cluster  
773 Analysis. A Wiley-Science Publication John Wiley & Sons.
- 774 Legoux, R., 2010. Diffusion en Île-de-France et en Picardie de la céramique ornée entre la  
775 fin du Ve et le VIIe siècle. In, Hincker V., Husi P. (dir.) 2006 - *La céramique du haut*  
776 *Moyen-Age dans le Nord-Ouest de l'Europe, Actes du colloque de Caen, mars 2004*,  
777 Éditions NEA, Condé-sur-Noireau, 429, 326-336.
- 778 Macnaughton-Smith, P., Williams, W.T., Dale, M.B., Mockett, L.G., 1964. Dissimilarity  
779 analysis: A new technique of hierarchical sub-division, *Nature*, **202**, 1034-1035.
- 780 Madsen, T., 1988. Multivariate statistics and archaeology, in Multivariate archaeology (ed. T.  
781 Madsen), Aarhus University Press, Aarhus, 7–27.
- 782 Orton, C. R., and Tyers, P. A., 1992. Counting broken objects: the statistics of ceramic  
783 assemblages, in New developments in archaeological science; a joint symposium of  
784 the Royal Society and the British Academy, February 1991 (ed. A. M. Pollard), 163–84,  
785 Proceedings of the British Academy, 77, published for The British Academy by Oxford  
786 University Press.
- 787 Pagès, J., 2014. Multiple Factor Analysis by Example Using R (1st ed.). CRC Press, 272  
788 Pages. <https://doi.org/10.1201/b17700>
- 789 Parzen, E., 1962. On estimation of a probability density function and mode. *Ann. Math. Stat.*  
790 **33**, 1065–1076. Rand, W. M., 1971, "Objective criteria for the evaluation of clustering  
791 methods". *Journal of the American Statistical Association*. American Statistical  
792 Association. **66** (336): 846–850. *arXiv:1704.01036*. *doi:10.2307/2284239*.  
793 *JSTOR 2284239*.
- 794 R Core Team, 2016. *R: A Language and Environment for Statistical Computing*. Vienna,  
795 Austria. Available at: <https://www.R-project.org/>
- 796 Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of  
797 cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65.  
798 [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- 799 Tyers, P. A., Orton, C. R., 1991. Statistical analysis of ceramic assemblages, in *Computer*  
800 *applications and quantitative methods in archaeology* (eds. K. Lockyear and S. Rahtz),  
801 117–20, British Archaeological Reports, International Series 565, Oxford.
- 802 Vince, A., 1993. Form, functions and manufacturing techniques of late 9<sup>th</sup> and 10<sup>th</sup> century  
803 wheelthrown pottery in England and their Origins, In. PITON eds. *La céramique du Ve*  
804 *au Xe siècle dans l'Europe du Nord-Ouest, actes du colloque d'Outreau (avril 1992)* ,  
805 n° hors-série de Nord-Ouest Archéologie, 151-164.
- 806 Vince, A., 2011. Characterising French whiteware imports in Britain and Northwest Europe.  
807 In Bocquet-Liénard A. et Fajal B., A propos de l'usage de la production et de la  
808 circulation des terres cuites dans l'Europe du Nord-Ouest autour des XIVE-XVle  
809 siècles, Actes des tables rondes du CRAHM, publication du CRAHM, Caen.
- 810 Woillez, M., Poulard, J.C., Rivoirard, J., Petitgas, P., Bez, N., 2007. Indices for capturing  
811 spatial patterns and their evolution in time, with application to European hake  
812 (*Merluccius merluccius*) in the Bay of Biscay. *ICES J. Mar. Sci.* **64**, 537–550.
- 813 Woillez, M., Rivoirard, J., Petitgas, P., 2009. Notes on survey-based spatial indicators for  
814 monitoring fish populations. *Aquat. Living Resour.* **22**, 155–164.
- 815
- 816