MULTIVARIATE MIXTURE OF NORMALS WITH UNKNOWN NUMBER OF COMPONENTS: AN APPLICATION TO CLUSTER NEOLITHIC CERAMICS FROM AEGEAN AND ASIA MINOR USING PORTABLE XRF

Ioulia Papageorgiou[(1)] and Ioannis Liritzis[(2)]

[(1)] *Department of Statistics, Athens University of Economics and Business, Patission 76, 10334 Athens, Greece (ioulia@aueb.gr)*

[(2)] *Laboratory of Archaeometry, Dept of Mediterranean Studies, University of the Aegean, 1 Demokratias Ave, Rhodes 85100, Greece (liritzis@rhodes.aegean.gr)*

ABSTRACT

Multivariate techniques and especially cluster analysis have been commonly used in archaeometry. Exploratory and model-based techniques of clustering have been applied to geochemical (continuous) data of archaeological artifacts for provenance studies. Model-based clustering techniques like classification maximum-likelihood and mixture maximum likelihood have been used to a lesser extent in this context and although they seem to be suitable for such data, they either present practical difficulties -like high dimensionality of the data- or their performance gives no evidence that they are superior to standard methods (Papageorgiou *et al.*, 2001). In this paper standard statistical methods (hierarchical clustering, principal components analysis) and the recently developed one of the multivariate mixture of normals with an unknown number of components (see Dellaportas and Papageorgiou, 2006) in the category of the model–based ones, are applied and compared. The data set comprises chemical compositions of 188 ceramic samples derived from the Aegean islands and surrounding areas.

KEYWORDS: CERAMIC COMPOSITIONS, CLUSTER ANALYSIS, MIXTURE MAXIMUM LIKELIHOOD, REVERSIBLE JUMP, OUTLIERS, PORTABLE XRF, AEGEAN.

1. INTRODUCTION

Provenance studies of the raw materials used during the prehistoric lithic industry are of key importance in research on ancient humans. During the Palaeolithic, this provides information on the extension of the territory exploited by small groups of hunter-gatherers.

In the Neolithic and Bronze Age, provenance studies contribute to the knowledge of long-distance circulation and exchanges of raw materials and goods, hence on the *chaines operatoires* of lithic and clay artifacts. Indeed, reconstructing mobility strategies is a major goal of researchers interested in prehistoric hunter-gatherers, and the use of geochemical source characterization of ceramics found at sites in a region offers a way to reconstruct the procurement range, or distance traveled to obtain resources of prehistoric groups.

Pottery, due to its remarkable storage properties, was a vital item used in every day life food activities. Not only these uses, but aesthetic qualities too were frequently used by ancient humans. Ceramics is also one of the preferred materials in provenance studies. This is because the physico-chemical properties are most often different at a major, minor but mainly trace element level because of its mode of formation from characteristic clay sources.

Early ceramic provenance studies were based on bulk physical properties, such as typology, technology, etc, as well as on petrography. Although useful for sample description, these observations generally do not provide valuable criteria for provenance studies.

The impact on characterization studies was made during the 1960s when spectroscopic methods allowed the determination of elemental compositions from small-sized samples. Since then nearly all provenance studies have been based on elemental composition. Among the destructive methods of analysis are electron microprobe (for about 10 major elements), neutron activation analysis (up to ~27 major to trace elements), ICP-MS/AES, with up to more than 50 elements determined, Optical Emission Spectroscopy, Atomic Absorption Spectroscopy, PIXE, and XRF, depending on instrumentation availability and allowance to sample in a destructive manner (Pollard & Heron, 1996). However non-destructive analysis is progressively used employing X-ray fluorescence (Liritzis *et al.*, 2002)

In this study the characterization of the analyzed ceramics was made with the application of standard statistical methods such as hierarchical clustering analysis and principal components analysis as well as model based clustering of multivariate mixture of normals with an unknown number of components (see Dellaportas and Papageorgiou 2006).

Statistical analysis and data reduction employing multivariate techniques lead to a number of variables that characterize a certain group of objects (ceramic in this context). The problem is to define groups in the data set, based on their compositional proximity. Such a comparison would result in groupings of ceramics and the raw materials they derive from. Because of the nature of the data (a number of continuous variables) and the problem of identification of such distinct groups, cluster analysis is the most appropriate multivariate method to use and has been widely used in archaeology together with principal component analysis (PCA).

The standard methods in cluster analysis are heuristic and consist of two main stages. (i) measuring the distance between data samples and (ii) application of a criterion to merge or split groups. A large number of standard methods that already exist can be adopted as a result of the various measures of distances in combination with the variety of the merging/splitting criteria. All of them are heuristic and distribution free which means they make no use of data distribution assumptions.

In the model-based techniques the groups are the results of an assumed distribution – usually the multivariate normal- that the data derive from. The two best known model-based techniques are classification maximum likelihood and mixture maximum likelihood.

In the next section we briefly describe mixture maximum likelihood, since the innovative technique employed in this work is directly linked with this methodology. In fact this approach tends to overcome the disadvantages of mixture maximum likelihood. A brief presentation and the idea behind the novel methodology is given. This is a Bayesian methodology and Bayesian approach to deal with problems in archaeology is not new (see for example, Buck *et al.* 1996).

An application is made as a case study on the chemical composition of ceramics derived from prehistoric settlements in the wide region of the Aegean and its results are presented in the section titled 'Statistical Analysis'. Finally, the obtained groupings are discussed along with current archaeological evidence and statistical evaluation.

2. STATISTICAL MODEL –BASED METHODOLOGIES FOR CLUSTERING

Model-based statistical methodologies assume that the observations forming the data are generated from a distribution. Usually the distribution is normal and because the dimension of the data is higher than one, it is multivariate normal. The assumption of normality is not essential, but quite common in such techniques. The mixture maximum likelihood approach assumes a mixture of multivariate normals regarding the data distribution. Let $\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ denote the data table with $\mathbf{x_j}$ a p-vector, representing the j^{th} observation (a ceramic sherd in this context) and p is the dimension of the data (number of variables that for each case we have measurements). Under this approach the data are coming from a population with density

$$w_1 f(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+w_2 f(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+\cdots+w_g f(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) \qquad (2.1)$$

where $f(x;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ is the density of the multivariate normal distribution, the mixture is assumed to be finite, consisting of g components with the same distributions, but different parameters, and w_k , *k=1,2,…,g* are the weights with $\sum w_k = 1$. Weights w_k represent the probabilities that a case $\mathbf{x}_j$ belongs to the *k*th component. Moreover, intuitively speaking, equation (2.1) models a data set with observations that are coming from g different populations. These populations can be described with the same form of distribution, but parameters differ among them. Weights in (2.1) are the probabilities or proportions from each population in the total distribution, if a sample has been taken. The simplest case of finite mixture is when g=2 (Everitt & Hand, 1981). For example if *x* measures the height of children of a certain age, then a mixture of two normal distributions as in (2.1) could be adopted in order to capture the difference we expect in boys and girls in a sample of children with mixed boys and girls. A descriptive measure of such a sample would probably reveal a two-mode feature, a fact that indicates a single normal would not be appropriate to describe the data. Formulating the problem under mixture densities, it could be assumed that the first component is the normal that describes height in boy's population and the same for the second in girls. Both can be seen as normals if we look at them separately, with means and variances that differ. Weights in (2.1) are now the probabilities of a member in the population to be a boy or a girl. In an archaeological context and especially in provenance problems, the assumption of a finite mixture can find application, if each group of observations having a similar composition can be seen as a subpopulation that is described from one component of the mixture. Formalizing the problem in provenance studies under this approach we provide a model-based (not distribution free) methodology for clustering observations and obtain all the merits of inference, coming from this. Moreover, although it is probably quite difficult to find a single distribution that fits the data altogether, because of the presence of different groups, it is easier to fit a distribution to each separate group. This distribution can have a common form such as the normal density for example. The likelihood function for a sample of size *n*, will be

$$\prod_{i=1}^{n}\sum_{k=1}^{g} w_k f(\mathbf{x};\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,.$$

Clustering the data will result after estimating the unknown parameters in the population, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, w_k (*k=1,…,g*). $\boldsymbol{\mu}_k$ is a vector of *k* parameters and $\boldsymbol{\Sigma}_k$ is a symmetric matrix, thus there are p×(p+1)/2 parameters for each of the g matrices. Following the estimation of the parameters and weights in the mixture, clustering the data in groups (components) will be performed on the basis of w_k. More precisely, an observation $\mathbf{x}_j$ is classified to the component with the largest weight w_k.

There are two disadvantages of a different nature in this technique. A practical one, that induces problems in the progress of the method, and a rather fundamental and methodological one. The practical problem arises from the fact of the large number of parameters that need to be estimated in contrast with the small number of observations available from excavation. Estimation suffers and a way to overcome this is to restrict ourselves by imposing constraints on the parameters among components to minimize the total number. The most usual constraint is to assume that matrices $\boldsymbol{\Sigma}_k$ are equal across *k=1,…,g*. Expectation-Maximization Mixture (EMMIX) is an algorithm discussed in McLachlan *et al.* (1999) that implements mixture maximum likelihood.

The second problem is more fundamental. It is necessary to predetermine the number of the components in the mixture in order for the technique to work. This leads to the necessity of reliable statistical tests to define the number of components in the finite mixture as a separate problem from the estimation. The methodology of maximum likelihood has to be executed for a variety of values for g (the number of components) and at a later separate stage, tests like the approximate weight of evidence AWE (Banfield and Raftery, 1993) and the Bayesian information criterion BIC (Fraley and Raftery, 1999) suggest the best *g* value. Unfortunately AWE, BIC and other similar tests are all approximate tests. As a result, they might even not provide the same suggested *g* value (Fraley and Raftery 1998).

In an attempt to deal with the problems of estimation and choice of the number of components in the finite mixture simultaneously, another approach, based in Bayesian inference, was developed and presented in Dellaportas and Papageorgiou (2006). The assumption for the basis of the problem is the same: a finite mixture with an unknown number of normal components. Making use of the powerful Bayesian technique of

Reversible Jump Markov Chain Monte Carlo (RJMCMC) (Richardson and Green, 1997) that allows the testing of models with different number of unknown parameters, it is possible to estimate the parameters in a mixture of *k* components and compare this with another mixture of *l* components with *l≠k*.

Some applications of univariate normal mixtures that use RJMCMC are presented in Nobile and Green (2000), Robert *et al.* (2000), Fernandez and Green (2002), Green and Richardson (2001), Bottolo *et al.* (2003). An extension to multivariate mixtures is the novelty in the approach by Dellaportas and Papageorgiou (2006). The multivariate context is appropriate for the application in clustering and moreover there are no constraints in the form of variance-covariance matrices of the components.

Under this methodology a mixture of normals as in (2.1) is assumed for the population density. With the Bayesian approach the data vector, say $\mathbf{x}_i$, given the set of parameters $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, w)$ follows a multivariate normal, i.e $[\mathbf{x}_i|\theta] \sim \mathrm{N_p}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\mathrm{N_p}$ denotes the p-dimensional normal distribution. In the parameter vector θ, $\boldsymbol{\mu}$ represents the vector of the means of the set of components, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k)$, where $\boldsymbol{\mu}_j$ is the mean of the j component, and in the same way $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_k)$, $w = (w_1, w_2, ..., w_k)$. The Bayesian formulation assumes prior distributions for the parameter *θ.* Given a known number of *k*, conjugate priors are assumed for the above situation. Priors for $\boldsymbol{\mu}_j$ given $\boldsymbol{\Sigma}_j$ are assumed to be normal with the mean depending on a partition of the data range (following Richardson and Green, 1997) and variance $\boldsymbol{\Sigma}_j / c_j$, where c_j are precision parameters. For $\boldsymbol{\Sigma}_j$ an inverse Wishart distribution is assumed, with scale parameter Ξ a p-diagonal matrix given a Gamma prior to each of the p elements. Parameters of gamma depend again on hyperparameters and range of the data. Finally, the prior for ***w***, the vector of weights, is chosen to be a Dirichlet p-dimensional distribution with parameters all equal to delta, where delta is a hyperparameter. For the analytical technical details on how to choose the priors, we refer to Dellaportas and Papageorgiou (2006).

The powerful Bayesian technique of RJMCMC (Green, 1995) is then applied for “jumping” between mixtures of different numbers of components. The tool for this is to have available an algorithm for the merging/splitting of components. Among the existing components a randomly selected one is a candidate to split into two in the ‘split’ move, so that the resulting mixture will be increased by one in the total number of components. In the ‘merge’ move, two components (again randomly selected among all) are merged in one and the total number of components in the mixture is reduced by one. The algorithm must be a one-by-one mathematical function so that it can be inverted and any move can be reversible. This means that starting from one component and after applying the ‘split’ move, we should be able to compose the two resulting smaller components with the help of the opposite ‘merge’ move and resulting in one that coincides with the original –before split- component. Because of this reversibility there is no need to give expressions for both moves. We present the mathematical details for the ‘split’ move in (2.2). Inverting the mathematical expressions in (2.2) can give us the complete details for ‘merge’ move.

Say that $\boldsymbol{\mu}_*$ is the mean, $\boldsymbol{\Sigma}_*$ the covariance matrix and w_* the weight of the candidate component for split. Then the split algorithm that provides the weights, the means, and the covariance matrices for the two new components, is

$$
\begin{aligned}
w_1 &= u_1 w_* \\
w_2 &= (1-u_1) w_* \\
\boldsymbol{\mu}_1 &= \boldsymbol{\mu}_* - \left(\sum_{i=1}^{\pi} u_2^i \sqrt{\boldsymbol{\lambda}_*^i \mathbf{V}_*^i} \right) \sqrt{\frac{w_2}{w_1}} \\
\boldsymbol{\mu}_2 &= \boldsymbol{\mu}_* + \left(\sum_{i=1}^{\pi} u_2^i \sqrt{\boldsymbol{\lambda}_*^i \mathbf{V}_*^i} \right) \sqrt{\frac{w_1}{w_2}} \qquad (2.2) \\
\boldsymbol{\Lambda}_1 &= diag(u_3)\, diag(\mathbf{I}-u_2)\, diag(\mathbf{I}+u_2) \boldsymbol{\Lambda}_* \frac{w_*}{w_1} \\
\boldsymbol{\Lambda}_2 &= diag(\mathbf{I}-u_3)\, diag(\mathbf{I}-u_2)\, diag(\mathbf{I}+u_2) \boldsymbol{\Lambda}_* \frac{w_*}{w_2} \\
\mathbf{V}_1 &= \mathbf{P}\, \mathbf{V}_* \\
\mathbf{V}_2 &= \mathbf{P}'\, \mathbf{V}_*
\end{aligned}
$$

where w_1, w_2 are the new weights that add to one, $\boldsymbol{\mu}_1$ $\boldsymbol{\mu}_2$ are the new means and $\mathbf{V}_*$, $\boldsymbol{\lambda}_*$ the matrices of eigenvectors and eigenvalues after spectral decomposition of $\boldsymbol{\Sigma}_*$. Following the same notation $\mathbf{V}_1$, $\boldsymbol{\lambda}_1$ and $\mathbf{V}_2$, $\boldsymbol{\lambda}_2$ correspond to $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively. u_1, u_2, u_3, **P** are random in order to complete the algorithm (same total number of variables between one component and two components). They are generated from certain distributions assumed. Matrix **I** is the p-dimensional identity matrix.

The method works iteratively and each iteration include a stage of testing if a move (either split or merge) would be accepted or staying in the same number of components and in any case estimate the parameters of the mixture based on the data. A more detailed technical description is presented in Dellaportas and Papageorgiou (2006).

3. CLUSTERING DATA FROM THE AEGEAN AND ANATOLIA

The present new clustering approach is applied to a large data set. In fact initially it was considered to serve as a test rather than a provenance question, albeit the latter can not be excluded under the light of new finds. The data set under study consists of 188 samples, deriving from eight archaeological excavation sites of Mesolithic, Neolithic and Bronze ages: Ftelia at Mykonos, Yali and Pergussa near Nissiros (Dodecanese), Kalithies cave in the island of Rhodes, Sarakinos Cave in Beotia, Central Greece, two settlements in Cyprus and Ulucak in Asia Minor near Smyrna (data available in our web sites: www.rhodes.aegean.gr/tms and www.stat-athens.aueb.gr/~ioulia). The project of chemical analysis and provenance of prehistoric pottery from the Aegean, Cyprus and Anatolia started in 2000 funded by the Ministry of the Aegean. Most samples have an age overlap during the late Neolithic and Bronze Age period and a question of interest was to provide statistical evidence for exchange of goods and ideas via a *chaines operatoires* model. This question was pursued in the light of recent excavation evidence and further at the recommendation of field archaeologists (Prof. A.Sampson, 2005, University of the Aegean

Rhodes and Prof. Altan Cilingiroglu, University Ege, Smyrna, 2005, personal communication).

Though distant sites (e.g. Boeotia and Cyprus) may exclude any possible contact, in spite of the overlapping period, nevertheless cultural contact between Neolithic and Bronze age sites in Asia Minor and the Aegean has been documented. Evidence from settlements and specific areas involved include Youra, Euboea, Skyros, Boeotia, Nea Makri Attica, Tigani at Samos, Vathi in Kalymnos, Ulucak central Anatolia, St. George's cave Kalithies Rhodes, Agio Gala Chios (Furness, 1956; French, 1965; Hood, 1981; Kamil, 1982; Ozdogan & Pendik, 1983; Yakar, 1985; Seher, 1990; Eslick, 1992; Mountjoy, 1998; Sampson, 2006).

Moreover, Aegean population theories since the onset of the Holocene, assigning human movements from Orient and the Balkans are also proposed (Ammerman & Cavalli-Sforza, 1984; Runnels, 1995; Van Andel & Runnels, 1988; Cherry, 1990, 1985; Broodbank, 2000). Various practical matters concerning seafaring have affected the navigation, exploration and colonization of the Aegean. The traveling of long distances through navigation is not a surprise. Melian obsidian has been found in the cave of Cyclop's in Youra, northwest Aegean, some 300 km far from Melos, and at Maroulas Mesolithic site in Kythnos (Sampson et al., 2002), indicating knowledge of navigation in the Aegean as early as the Holocene (Keegan & Diamond, 1987, Davis, 1992). In addition, the stone industry at Youra shows such contacts / similarities with other Anatolian caves, while the most recent field work (summer 2005) has revealed three Mesolithic sites in Icaria island, off the Anatolian coast, where stone artifacts have many similarities with those of Maroulas in Kythnos (Kaczanowska & Kozlowski, 2006; Sampson & Koslowski, 1999).

On the other hand, several similarities are observed during the 9th to 7th mil. BC between western Asia Minor and the Aegean which are synchronized away from SE Anatolia. Thus, at Dodecanese, the pottery shapes and decoration of 6th-5th mil BC Kalithies and Koumelo caves derive from various Anatolian and Aegean LN prototypes (Sampson, 1984, 1987; Melas, 1988). Although the pottery distribution is not uniform, no clear borders can be set or recognized for separating cultural or even stylistical zones. This situation stresses the role of interaction and exchange from this early period. For example the presence of obsidian at Kalithies underlines these contacts, especially when 80% comes from Melos and the rest from Yiali and a source in central Anatolia (Sampson, 1984, 72). It is remarkable when Aegean shifted to a full pottery Neolithic economy around 6500-6300 BC, almost at the same time in mainland Greece, the islands and Crete, following similar developments in Anatolia and the Near East; Cyprus although fully neolithicized, developed a distinct aceramic culture, after a hiatus postdating Shillourokambos (Stanley-Price, 1979; Lebrun, et al., 1987). Architectural similarities (circular huts) between Cyprus and Kythnos during Mesolithic, indicates association of both sites with an archaistic mentality, an idiosyncratic conservativism, although their economies are different. Even if we exclude population movements between these distant islands by the sea, it would be realistic to argue a gradual spread of ideas through indirect contacts, taking into account the lower sea level by about 40-50 meters during 7th-9th mil. B.C. (Pirazzoli, 2000; Katsarou-Tzeveleki, 2001; Efstratiou & Mantzourani, 1997). This is postulated by Katsarou-Tzeveleki & Sampson (2006, 112) as *"...Moreover, a likely introduction of the pre-pottery terminology in our analysis of the Aegean corroborates the typical view of the East as the birthplace and the Aegean as the periphery. And as the recent discovery of the Cypriote* *preceramic phase highlighted the naval background of the SW Asian civilizations, linking them with the Aegean seafarers by common provenance, the domestication techniques and the circular architecture of the Aegean are degraded to just another diffusion symptom".* However, navigation techniques progressively developed had been highly established by late Neolithic (5th mil. BC) and thereafter large islands (e.g. Cyprus, Crete) break isolation and are capable of maintaining continuous contacts with the surrounding areas (Aegean, Near East), "…*which discourages conservatism and ensures renovation and normal cultural sequence*" (Katsarou-Tzeveleki, 2001). During the early Bronze Age several socio-economic and technological (pyrometallurgical) changes have taken place in the Aegean and southern eastern Mediterranean, highlighting the role of exchanges, contacts and maritime networks. Contacts and interaction allowed the circulation of ideas, symbols and objects between Rhodes and Anatolia, Cyclades and eastern Greece (French, 1968; Marketou, 1990).

The cultural overlapping is a common image in the Aegean throughout its prehistory, underlying the significance of interaction and not necessarily revealing cultural domination from some adjacent cultures. Autochthonous and semi-autochthonous development seems to be the case of cultural interaction affected by local and interregional development.

With these in mind we attempted to cluster characteristic pottery finds from the aforementioned sites in the Aegean, western Asia Minor and Cyprus, even if the latter seems a remote possibility.

Samples derive from well stratified archaeological sections dated by C-14 and represent characteristic typology provided by the excavator per case.

Table 1 gives the sites involved, the sampled code number used in the present analysis, the dating and respective references.

Site	**Code (Labels)**	**Date, BC**	**Reference**
Ulucak, Asia Minor, near Smyrna, Turkey	RHO-38 to RHO-108	ca.6000-2000 (Late Neolithic, LN), (Middle Neolithic, MN) , (Late Chalcolithic, LC)	Prof. A. Ciliniroglou (MAA 2005, in press)
Ftelia Neolithic settlement, Mykonos	MFC1-28	ca.4500-5100 Late Neolithic (LN)	Sampson (2002)
Kalithies cave Rhodes	KR1-10	Late Neolithic, ca.5300-4500 (LN)	Sampson (1987)
Yali Neolithic settlement, near Nissyros and Pergoussa	YNB…, PERG	ca.5000-3000 (LN, Early Bronze Age, EBA)	Sampson (1988)
Sarakinos, cave, Boeotia	SARA….	ca.5800-2500 (LN, EBA)	Prof.A.Sampson, pers. comm.
Cyprus, Koufovounos and Sotiras	CK1-6 and CS1-11	Later part of Late Neolithic, ca.4000-2500 (LN, EBA)	Mantzourani and Liritzis (2005)

Table 1: Samples classified by origin of excavation, dating, archaeological period and references

3.1 Sample Preparation

In all ceramic sherds the outer surface was discarded to avoid weathering implying leaching/ infiltration of ions, thus altering elemental composition. Solid pieces of ceramic pieces and soils were powdered (<90 μm), dried, and measured by a portable ED-X-Ray Florescence analyzer (ED-XRF).

3.2 The Analyser ED-XRF

The EDXRF field portable analyzer Spectrace 9000 TN was used with a mercuric iodide (HgI_2) detector, which has a spectral resolution of about 260 eV FWHM at 5.9 keV, and three excitation sources of radioisotopes within the probe unit – Americium Am-241 (26.4 KeV K-line and 59.6 KeV L-lineV) measuring Ag, Cd, Sn, Ba, Sb; Cadmium Cd-109 (22.1 K-line, 87.9 K- & L-line KeV) measuring Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Sr, Zr, Mo, Hg, Pb, Rb, Th, U; and Iron Fe-55 (5.9 KeV K-line) measuring K, Ca, Ti, Cr.
The system was calibrated on several standard clays and bricks, and the application software 'Fine particle of soil application' was used.

The performance of the portable XRF instrumentation in the laboratory has already been reported (Potts et al., 1995, 2001). Here, a wide range of silicate rock reference materials were analysed as powder pellets to evaluate accuracy, precision and detection limits. Reference samples included Bonn clay Univ. of Bonn, Montana soil NIST, CFA ash NIST, Brick clay NIST. The study showed the capability of the instrument to determine major and minor elements (K, Fe, Ca, Mn, Ti, Cr) and selected trace elements (Sr, Zr, Rb, Ba, Ni, Ag). Other trace elements were not measured because their lower counting sensitivities mean that the concentrations were near to or below detection limits.

4. STATISTICAL ANALYSIS AND DISCUSSION

For the total number of samples (188) the analyzer described above provided us with measurements of nine elements: Ba, Fe, Rb, K, Ti, Mn, Sr, Zr, Ca. The set of nine elements was the common subset of elements for which measurements were available for all the samples.

Thus the processed data set is a nine-dimensional (188×9) data matrix. An initial hierarchical clustering allows us to separate some very clear and compact groups that separate well from the remaining. Several hierarchical techniques, like complete linkage, average linkage, single linkage and Ward's method have been applied and agreement among all was possible about this issue. Figure 1 shows the dendrogram of average linkage of the standardized data set of the 188 samples. The result of this clustering is plotted in the first 2 Principal Components and presented in Figure 2. The obvious groups (and singletons/outliers) are groups noted by "1's", "2's", "5's" "6's", "7", "8", "9" and"10" (the last four are singletons, e.g. groups of size one).

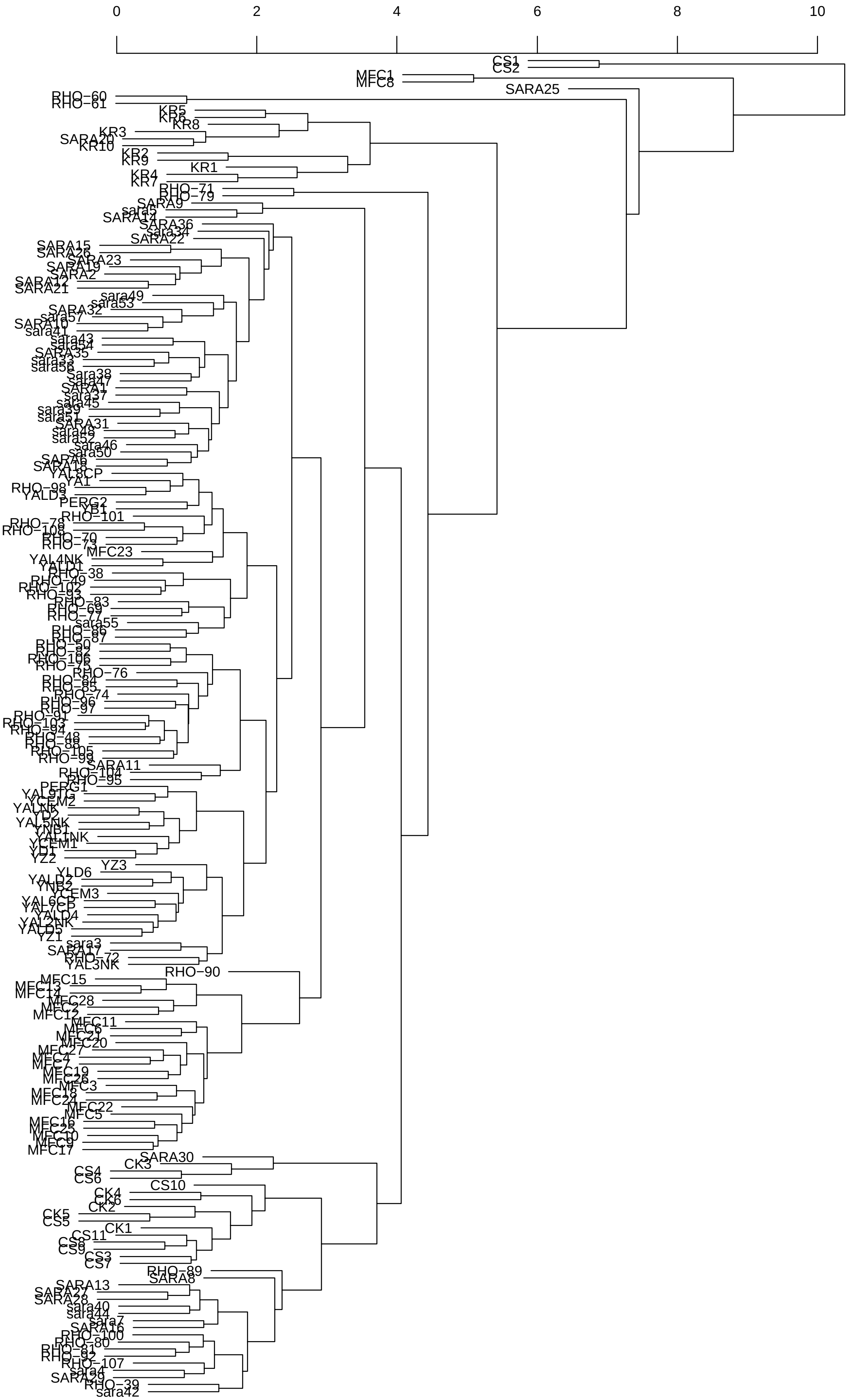

Figure 1A dendrogram based on cluster analysis (hierarchical, average linkage) of the data set in total.

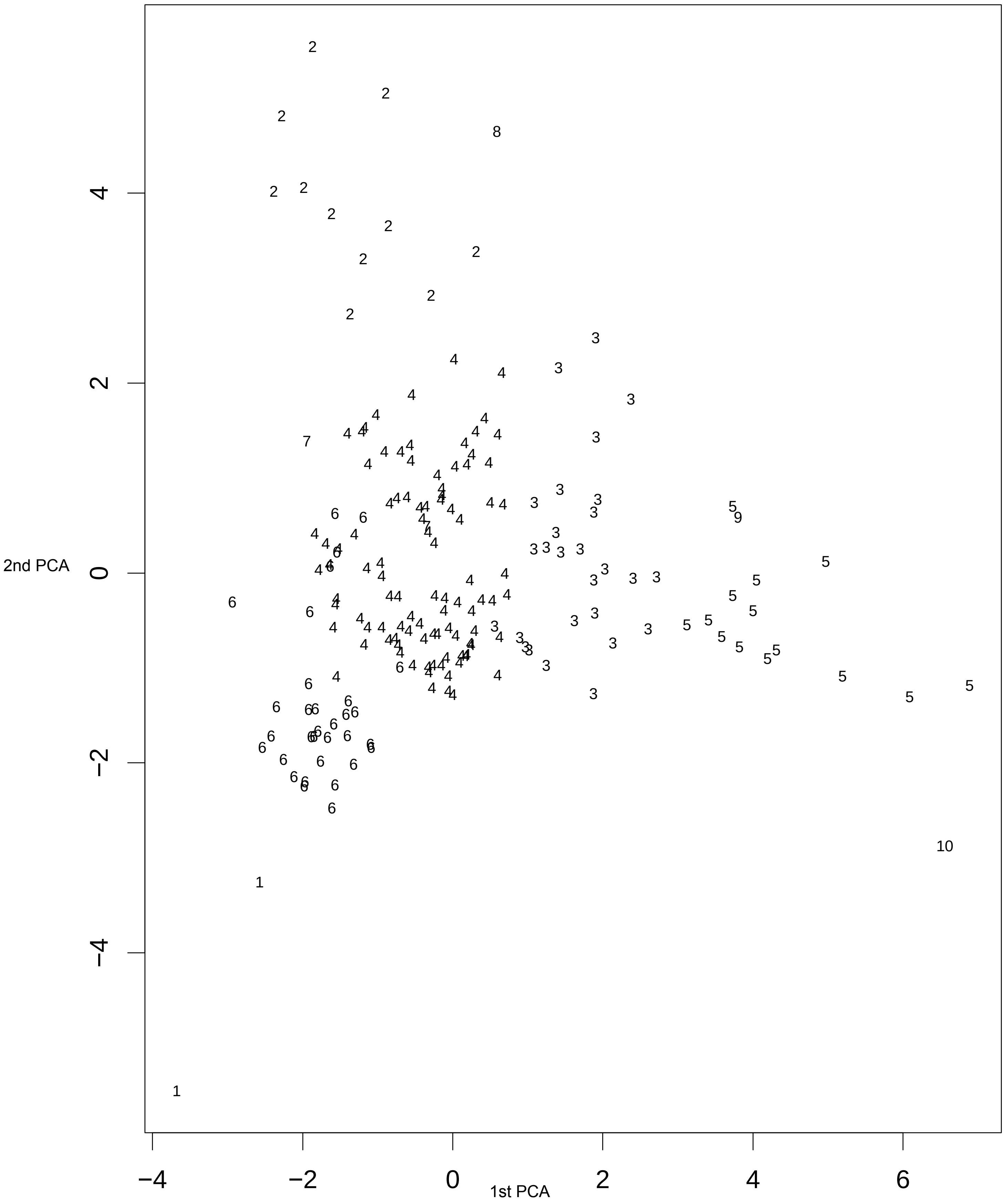

Figure 2 *A two-dimensional component plot based on principal component analysis of the resulting groups from the dendrogram in Fig. 1. Each number, ‘1’, ‘2’, . . . , ‘10’, corresponds to a separate group.*

In a comparison between the figures and studying each branch of the tree, the distinct groups are:

Group "6"	Samples from Ftelia and Mykonos (MFC-1 to MFC-28)
Group "1"	A group of 2 members: MFC1 and MFC8 that separate from gr. "6"
Group "2"	10 samples from Kalithies (Rodos) KR1-10
Group "5"	Samples from Cyprus (both origins) and 2 samples from Rodos: RHO-60, RHO-61 and one sample from Sarakino, SARA-30.
Group "9", "10"	Two singleton groups with members CS1 and CS2 respectively, the two exceptions of group "5".

The second stage of the analysis is to remove the above described clear groups of Ftelia, Kalithies, Cyprus and singletons/outliers like SARA-25 (group 8 in figure 2) and continue further analysis with the remaining samples. After this "peeling-off" procedure of the data we end up with 125 samples, mainly consisting of Ulucak, Yali and Sarakinos. For this remaining data set, both hierarchical clustering and model-based (with RJMCMC) techniques have been applied and results have been compared. The algorithm converged rather quickly and suggested a three-component mixture as the most powerful model. Figures 3 and 4 present the predictive density based on all iterations of RJMCMC. The predictive density of future data is a posterior inference statistic. The predictive distribution is the distributions of future observations that is new ceramics in our context. It includes all the information of how a future observation would distribute according to the model of our existing data and given the parameters. If y is a future observation and x past (existing data) then the predictive density for $y \mid x$ is

$$y \mid x \sim f(y \mid x) = \int f(y \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \qquad (4.1)$$

where, $f(y|\boldsymbol{\theta})$ is the model of our data given parameter vector and $\pi(\boldsymbol{\theta})$ is the posterior density for the parameter $\boldsymbol{\theta}$.

Samples from the predictive density can be generated by sampling one, or more, data points for each sampled point of parameter θ selected from the RJMCMC iterations. The density is plotted in all possible 2-dimensional projections of the first 5 principal components. The samples (points in the figures) are also plotted in the same images. Although the images are the projections of the density, they show that the predictive density captures the data quite well. The multidimensionality of the data, and the difficulty of testing the fit of the model (2.1) to our data, was the reason to choose the predictive density for inference. Plotting this density in two dimensions solves this problem and gives also graphical as well as numerical information about the means and variances of the model we fit. In the case of clustering this is important as we have available how information about the means differ in various groups, or why some data points form a group (probably different dispersion or orientation of a neighbor group).

Figure 5 shows the projection of the predictive density in the first two PCs with the labels (codes) of the data simultaneously shown. Figure 6 gives the dendrogram of an average linkage cluster analysis applied to the same standardized subset of samples. There are also three main groups. Some possible outliers that hierarchical clustering suggests (e.g. those of RHO-89, SARA-8, YAL4NK, YALD1) were not expected to form a separate group in the model based methodology. To verify the result made of this methodology the posterior probabilities for these samples may be checked. Although in all cases the

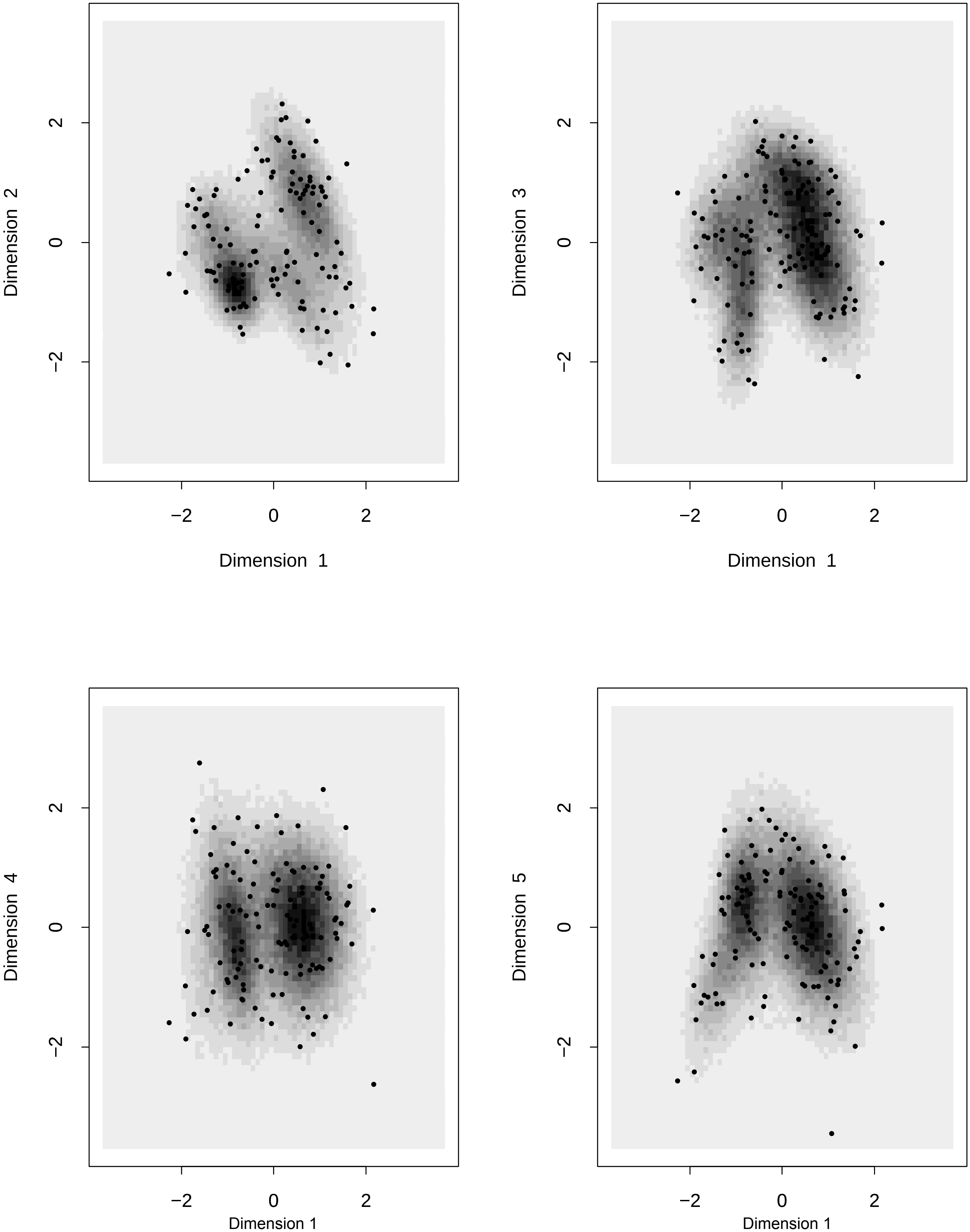

Figure 3 *The two-dimensional component plots of the predictive density derived from RJMCMC methodology applied in the first five PCAs of the reduced data set.*

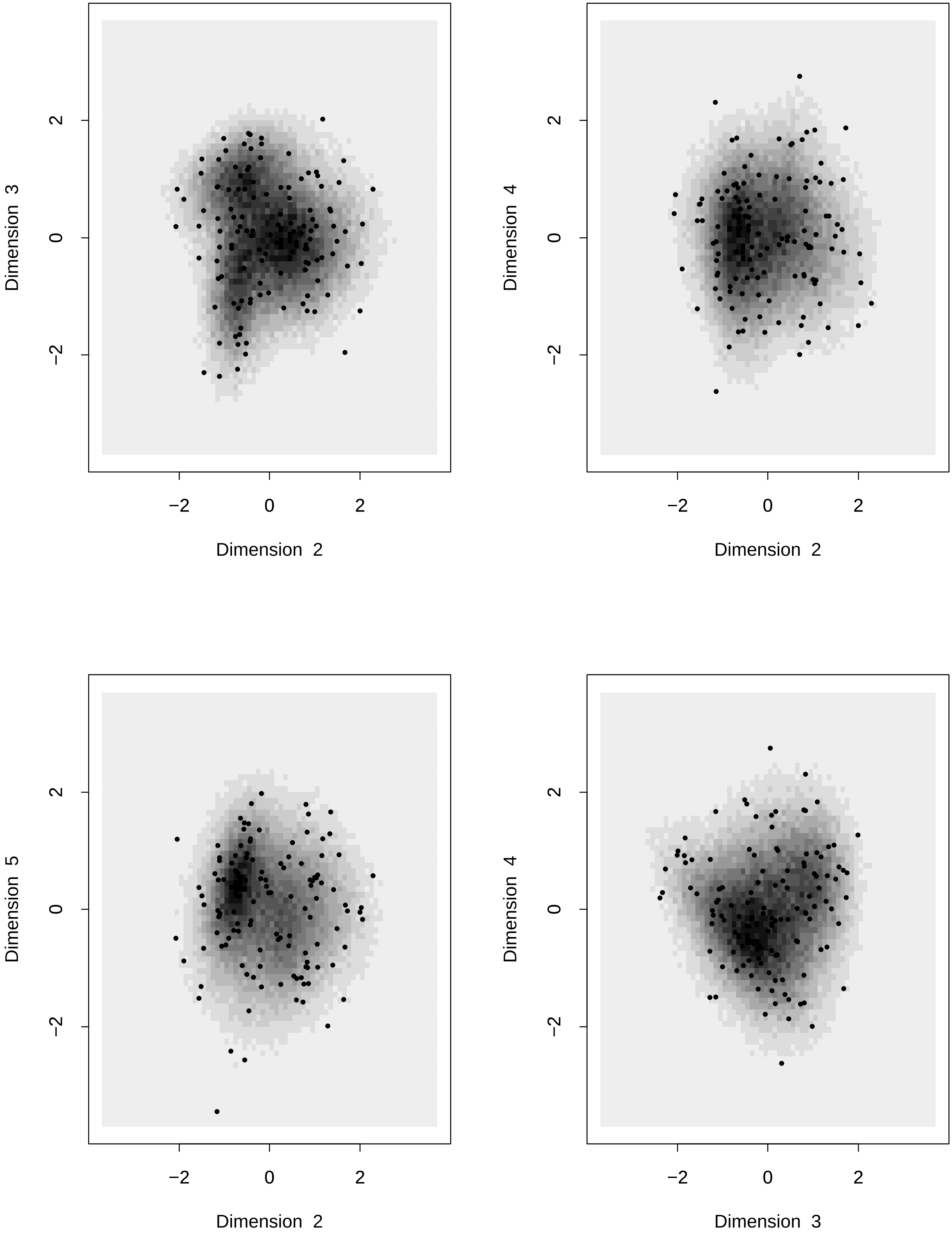

Figure 4 *The two-dimensional component plots of the same result of predictive density as in Fig. 3 for the remaining combinations of components.*

probability that a sample belongs to a group will be dominant for one component and classification will be obvious, in cases of outlier samples the probabilities will be almost equally shared among two or three different components. A more detailed comparison and conclusions from the analysis are given in the next discussion section.

With a detailed examination of the remaining samples, and Figure 4 and 5 in addition, with the results of the analysis in the first stage, the following observations are made:

1) The pottery from the eight sites indicate a clear intra- and inter-site interaction. Robust clusters of the major sites are clearly seen. For example, the *Ftelia group, the Yali group, the Kalithies* and *Sarakinos caves,* the two *Cyprus settlements.*

2) There appears an intrasite distribution and occasionally extremely distant "outliers". This is the case with Sotiras CS2, and Koufovouno Cyprus, CK2. In fact these two are a combed bowl and a monochrome flask respectively, derived from different floors and phases too, in relation to the rest. In fact the ceramics from two settlements have sub-groups implying more than one clay source, and at the same time, clusters comprising samples from either site, indicating communication through pottery exchange. This is anticipated result because of their proximity and same cultural phase.
Similarly, the two Ftelia MFC1, 8 of the latest date (c.4500-4700 BC) indicate a quite different origin of clay in relation to the rest.

3) The Sarakinos cave group exhibits a greater spread around an apparent central nucleus, with an obvious "outlier" SARA25, and others falling within neighbour clusters, - e.g. SARA20 close to Kalithies Rhodes, SARA30 along the elongated distribution of Cyprus groups, and several RHO (Ulucak) (77, 86, 101, 83, 87, 75, 96) form separate distinct sub-groups within the SARA main cluster. Though some Ulucak sherds (RHO39 EBA, RHO107 LN, RHO80 LC, RHO81 LN, RHO92 LN) of LN, EB and LC periods, overlap with some SARA (4 EBA, 29 EBA, 3 MN, 17 MN) from Early Bronze and Middle Neolithic, while SARA42 of LN I a-b period belongs to the same subgroup of RHO: 72 LC, 95 LN, 69 LC. This interesting pattern implies possible interactions (exchange of ceramics and/or sharing same clay source), enhanced by the fact they are of the same period i.e. Late Chalcolithic / Early Bronze Age (4000-2500 BC), Late Neolithic and Middle Neolithic. This finding needs further verification.

4) Two soil samples from the local floor of Ulucak settlement (RHO60, 61), form a group as expected. In all techniques they both are quite distant from the main Ulucak cluster(s) which is consisted of ceramics. This is expected because although the origin is the same, the source of its production is not clay. Clay is on the other hand the source of production for pottery.

5) Yali and Pergussa ceramics form distinct subgroups. Several RHO (Ulucak) ones (49, 93, 38, 102, as well as, those of 72, 95, 69, 78, 108) fall within Yali subgroups but form distinct clusters, and RHO-102 is close to Pergussa one- both of LN period- but on another tree-branch. Also RHO98 resembles Yali YALD3, both of LN period, too. Such interaction is possible during the Late Chalcolithic (for Ulucak) and Late Neolithic (Greek Neolithic at Yali). The two sites are close to the Asia Minor coastline, Ulucak being c.15 km from Smyrna.

6) In Ulucak, a quite interesting observation is the apparent use of a particular clay source throughout the long period of successive cultural phases (Early Bronze, Late

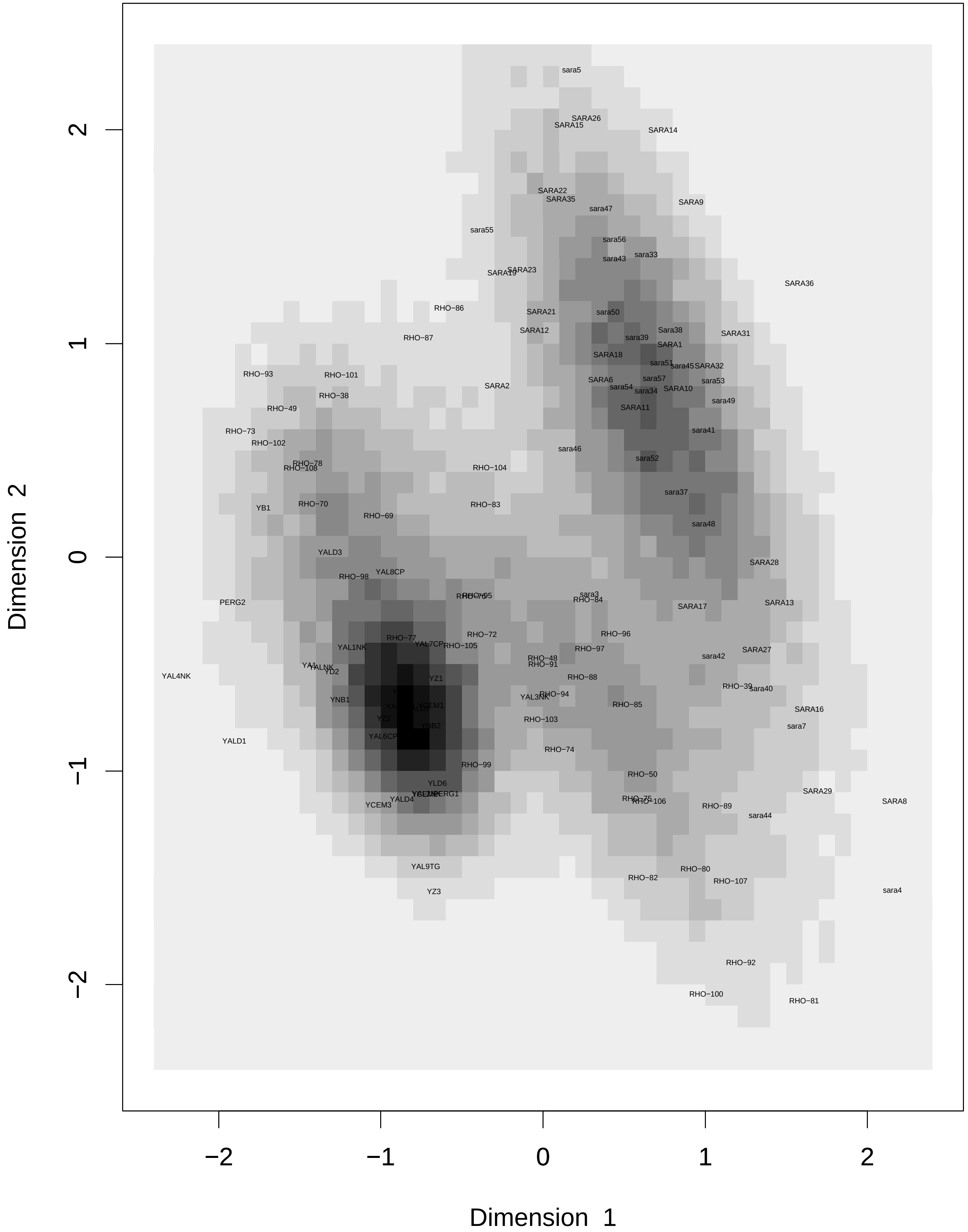

Figure 5 *A two-dimensional plot of the projection of the predictive density in the first two PCAs from the results of RJMCMC methodology, with the labels (codes) of the data shown. This is in order to confirm that the suggested grouping from RJMCMC methodology agrees with the existing grouping in data.*

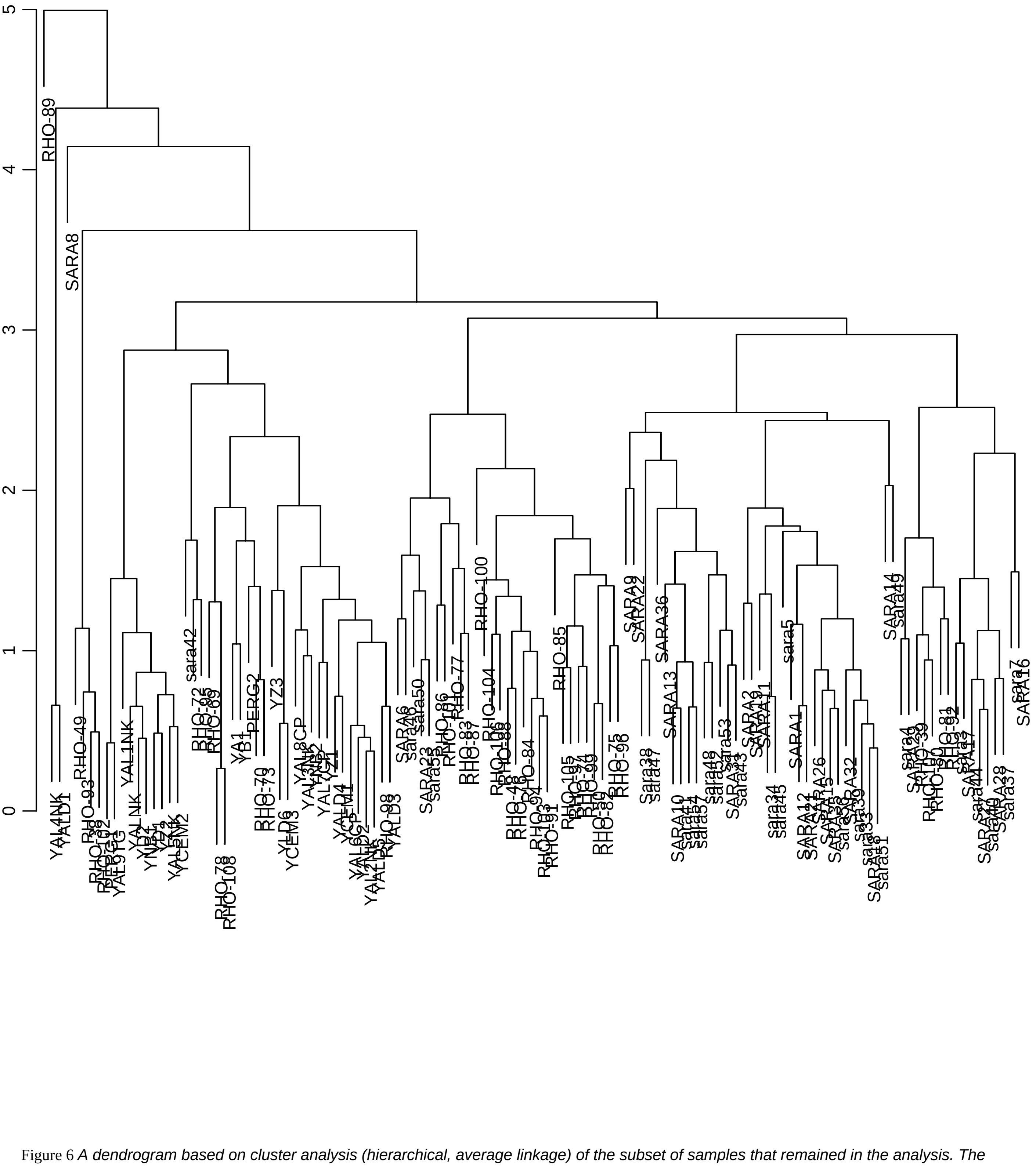

Figure 6 *A dendrogram based on cluster analysis (hierarchical, average linkage) of the subset of samples that remained in the analysis. The main aim here was to compare RJMCMC with hierarchical clustering. The results suggest the same number of main groups (except outliers/singletons).*

Chalcolithic, Late Neolithic, late Early Neolithic), evidenced from subgroups containing ceramic sherds from these periods.

The extremely interesting Ulucak- Yali-Pergoussa and Sarakinos-Ulucak interaction needs further verification.

5. CONCLUSION

The attempted characterization on a diverse set (temporal, contemporary and geographical) of pottery samples, mainly to test the success of the novel grouping model-based method, has proved highly satisfactory. An additional advantage in contrast with the non-model based clustering techniques, is that full estimation of the parameters exist after model-based classification and it is possible to classify a new incoming sample to one of the existing groups (discriminant analysis).

The obtained results indicate useful information regarding long distance trade exchange, usage of the same clay source by successive cultural phases, and interaction of settlements via sea routes. Some 'outliers' imply very different clay sources.

The standard methods in cluster analysis used here are distribution free which means they make no use of data distribution assumption. However, in the model-based techniques applied the resulting groups follow the multivariate normal. The recently introduced iterative methodology that is applied in this paper is a model-based technique with the same philosophy as mixture maximum likelihood, under a different formulation (Bayesian) and improved in the sense that it is devoid of disadvantages that mixture maximum likelihood has. Standard and model-based techniques were used in our application made as a case study on the chemical element composition of ceramics derived from prehistoric settlements in the wide region of the Aegean.

It is the first time to our knowledge that this endeavor to group prehistoric ceramic fabric derived from seemingly distant cultures in and across the Aegean has been undertaken. Questions posed by archaeologists often refer to the use of common clay sources, exchange trade routes, diachronical accessibility of same clay source.

Acknowledgements: We thank Prof. A.Sampson, Prof. E.Mantzourani and Prof. A. Cilingiroglu for providing samples and Ministry of Culture for granting permission, the General Secretariat for Research & Technology for partial funding of the project and the Ministry of the Aegean for funding initial stages of the project.

REFERENCES

Ammerman, A.J and Cavalli-Sforza, L.L (1984) *The neolithic tradition and the genetics of populations in Europe*. Princeton.

Broodbank,C (2000) *An island archaeology of the Early Cyclades*. Cambridge University Press, Cambridge.

Banfield, J. D. and Raftery, A.E (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803-21.

Bottolo L., Consonni G., Dellaportas P. and Lijoi A. (2003) Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6, 25-47.

Buck, C.E., Cavanagh, W.G. and Litton C.D. (1996). *Bayesian Approach to Interpreting Archaeological Data*. John Wiley and Sons, Inc. New York.

Cherry, J (1990) The first colonization of the Mediterranean islands: a review of recent research. *Journal of Mediterranean Archaeology*, vol.3, 145-221.

Cherry, J.F (1985) Islands out of the stream: isolation and interaction in early east Mediterranean insular prehistory. In *Prehistoric production and exchange: the Aegean and eastern Mediterranean,* A.B.Knapp and T.Stetch (eds.), 12-29, Los Angeles.

Davies, J.L (1992) Review of Aegean prehistory I: the islands of the Aegean. *American Journal of Archaeology,* 96, 699-756.

Dellaportas, P. & Papageorgiou, Ioulia (2006) Multivariate mixtures of Normals with unknown number of components. *Statistics and Computing*, **16**, 1, 57-68.
(Available online in http://stat-athens.aueb.gr/~ptd/finmix.pdf)

Eslick, C. *Elmalı-Karataş I, The Neolithic and Chalcolithic Period*, Bryn Mayr, 1992.

Eftsratiou, N and Mantzourani, E (1997) The beginning of the Neolithic period in Greece and Cyprus: Common research and interpretation problems. In *Cyprus and the Aegean in Antiquity. From the Prehistoric period to 7th c. B.C.* Proceedings of International Congress, Nicosia 7-20.

Everitt, B.S and Hand D.J. (1981). *Finite Mixture Distributions*. Chapman and Hall, Cambridge University Press.

Fernandez C. and Green P.J. (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Socisety, Series B*, 64, 805-826.

Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis, *Computer Journal,* **41**, 578-88.

D.H.French, *Anatolia and the Aegean in the third Millenium B.C.*, (Basılmamış Doktora Tezi),Cambridge, 1968.

D.H.French, “Early Pottery Sites from Western Anatolia”*Bulletin of the Institute of Archaeology* 5 (1965), 15 vd.

Furness, A (1956) Some early pottery from Samos, Kalimnos and Chios. *PPS* 22, 173.
Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28, 355-376.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Hood, S. (1981) *Excavations at Chios 1938-55. Prehistoric Emborio and Agio Gala,* Thames and Hudson, London.

KaczanowskaM & Kozlowski,J (2006) Palaeolithic traditions, Mesolithic adaptations and Neolithic innovations as seen through lithic industries. In *Sampson,A The Prehistory of the Aegean, Palaeolithic- Mesolithic- Neolithic*, Publ. House Atrapos, Athens, 67-87 .

T.Kamil, *Yortan Cemetery in the Early Bronze Age of Western Anatolia* BAR International Series 145, Oxford,1982

Katsarou-Tzeveleki, S (2001) Aegean and Cyprus in the early Holocene: Brothers or distant relatives? *Mediterranean Archaeology & Archaeometry*, Vol.1, no. 1, 43-55.

Keegan, W and Diamond, J (1987) Colonization of islands by humans: A biogeographical perspective. *Advances in Archaeological Method and Theory*, vol.10, 49-92.

Lebrun, A, Cluzan. S, Davis. S, Hansen.J, Renault-Miskovski, J (1987) Le neolithique preceramique de Cypre. *L' Anthropologie*, vol.91, 283-316.

Lindsay B. G. (1995) *Mixture models: Theory, Geometry and Applications*. Hayward: Institute of Mathematical Statistics.

Liritzis.I (2005) ULUCAK (Smyrna, Turkey): chemical analysis with clustering of ceramics and soils and obsidian hydration dating. *Mediterranean Archaeology & Archaeometry*, Vol.5, Special Issue (in press).

Liritzis.I, Drakonaki.S, Vafiadou.A, Sampson.A and Boutsika.T (2002) Destructive and non-destructive analysis of ceramics, artifacts and sediments of Neolithic ftelia (Mykonos) by portable EDXRF spectrometer: first results. In Sampson.A (ed.) (2002) *The Neolithic settlement at Ftelia, Mykonos.* Dept. of Mediterranean Studies, Univ. of the Aegean, Rhodes, Greece, Chapter 11, 251-271.

Mantzourani H. and Liritzis, I. (2005) Notes on the chemical analysis of pottery samples from Kantou-Kouphovounos and Sotira-Tepes (Cyprus): a comparative approach. *Cyprus Archaeological Reports*, Nicosia (in press)

Marketou, T (1990) Asomatos and Seraglio: EBA production and interconnections. Hydra, *Working Papers in Middle Bronze Age Studies*, 7, 40-48.

MAA (2005) Mediterranean Archaeology & Archaeometry, Vol.5, Special Issue. ‘Ulucak Hoyuk Project’. Rhodes, Greece (www.rhodes.aegean.gr/maa_journal).

McLachlan G. J., Peel, D., Basford, K. E. and Adams, P. (1999). The EMMIX algorithm for the fitting of mixtures of normal and t-components, *Journal of Statistical software*, **4**, 2.

McLachlan G. J. and Basford K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Melas, E.M (1988) The Dodecanese and West Anatolia in Prehistory: Interrelationships Ethnicity and Political Geography, *Anatolian Studies* 38, 109-120.

Mountjoy, P.A (1998) The east Aegean- west Anatolian interfacein the Late Bronze age: Myceneans and the kingdom of Ahhiyawa. *Anatolian Studies* 48, 33-68.

Nobile, A. and Green, P.J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, **87**, 15-35.

Özdoğan M., Pendik,A (1983) A Neolithic Site of Fikirtepe Culture in the Marmara Region. In *Beitrage zur Altertumskunde Kleinasiens, Feschrift für Kurt Bittel* (ed. R.M.Boehmer-H.Hauptmann), Mainz, 401-411.

Papageorgiou, Ioulia, Baxter M.J. & Cau M. A. (2001) Model-based clustering techniques in archaeological ceramic provenance studies. *Arhaeometry*, **43**, 4, 571-588.

Pirazzoli, P.A (2000) *Sea level changes. The last 20,000 years*. J.Wiley & Sons, England.

Pollard.M and Heron.C (1996) *Archaeological Chemistry*. The Royal Society of Chemistry, London.

Potts, J.P, Webb.P.C, Williams-Thorpe,O 1995: Analysis of silicate rocks using field-portable X-ray fluorescence instrumentation incorporating a mercury (II) iodide detector: a preliminary assessment of analytical performance. *Analyst*, **120**, 1273-1278.

Potts, J.P, Ellis, A.T, Kregsamer, P, Marshall, J, Streli, Ch, West, M and Wobrauschek, P 2001: Atomic spectrometry update. X-ray fluorescence spectrometry. *Journal of Analytical and Atomic Spectroscopy.* **16**, 1217-1237.

Richardson, S. and Green, P. J. (1997). On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Socisety, Series B*, 59, 731-792.

Robert C.P., Rydén T. and Titterington D.M. (2000) Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 62, 57-76.

Robert C. P. (1996) Mixtures of distributions: inference and estimation, *Markov chain Monte Carlo in Practice*, (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 441-464, Chapman and Hall, UK.

Runnels, C (1995) Review of Aegean prehistory IV: The Stone age in Greece from the Palaeolithic to the advent of the Neolithic. *American Journal of Archaeology,* vol.99, 699-728.

Seher, J. "Coşkuntepe-Anatolische Neolithikum am Nordostufer der Agais" Istmitt 40 (1990), 9-15.

Stanley-Price, N (1979) *Early prehistoric settlement in Cyprus: A review and gazetteer of sites, c.6500-3000 BC.* BAR International Series **65**, Oxford.

Sampson.A (ed.) (2002) *The Neolithic settlement at Ftelia, Mykonos.* Dept. of Mediterranean Studies, Univ. of the Aegean, Rhodes, Greece, pp.332

Sampson, A (1984) The Neolithic of the Dodecanese and Aegean Neolithic. Annals of the *British School at Athens* 79, 239-249.

Sampson, A and Koslowski,J (1999) The cave of cyclope in the northern Aegean: a specialized fishing shelter of the Mesolithic and Neolithic periods. *Neo-lithic*, 3, 5-7.

Samson.A (1987) *Neolithic period in Dodecanesse*, Ministry of Culture, TAPA, Athens.

Sampson,A (1988) *The Neolithic habitation at Yali, Nisyros*. Evoiki Archeofilos Etaeria, (in Greek with English summary), Athens.

Sampson.A, Koslowski, J, Kaszanowska, M and Giannouli, B (2002) The Mesolithic settlement at Maroulas, Kythnos. *Mediterranean Archaeology & Archaeometry*, vol.2, No.1, 45-67.

Sampson,A (2006) *The Prehistory of the Aegean, Palaeolithic- Mesolithic- Neolithic*, Publ. House Atrapos, Athens (in Greek with extended English version).

Van Andel, Tj.H and Runnels, C.N (1988) An essay on the 'emergence of civilization' in the Aegean world. *Antiquity* vol.62, 234-247.

Yakar, J (1985) *The Later Prehistory of Anatolia. The Late Chalcolithic and Early Bronze Age Age,* BAR International Series 268, Oxford.