

## ASSESSING CLUSTER VALIDITY IN THE STUDY OF ENEOLITHIC COPPER ARTIFACTS FROM ROMANIA

M. KADAR

"1 DECEMBRIE 1918" UNIVERSITY OF ALBA IULIA, ROMANIA

I. ILEANA

"1 DECEMBRIE 1918" UNIVERSITY OF ALBA IULIA, ROMANIA

L. MARINA

"1 DECEMBRIE 1918" UNIVERSITY OF ALBA IULIA, ROMANIA

SEE THE CD FOR THE EXTENDED VERSION

ABSTRACT

High purity copper artifacts dated to the Eneolithic period from Romania have been grouped by cluster analysis in order to establish the geological nature of raw materials used by the early metalworking in this area. The chemical composition in major and trace elements has been determined by ICP-AES. Trace elements used as markers for native copper have been analyzed and grouped in order to distinguish between made techniques of the heavy implements.

Here we discuss some aspects regarding the number of "real" clusters and methods of validation. A subset of samples of known origin that was "a priori" supposed to group together have been introduced in order to monitor the clustering. Additionally, principal component analysis has been used to supplement the cluster analysis.

## INTRODUCTION

An important and still debated issue is the use of native copper in the early stages of the human prehistory and its influence to the development of the metallurgy world wide. Scholars have been studied ways of distinguishing between objects made of native copper from those made of smelted copper, some methodology being elaborated by combining chemical composition analysis with metallographic investigation (Maddin et al. 1980, Rapp 1982, Hancock 1991, Wayman and Duke 1999).

In order to understand the inception of metallurgy as one of the most important events of the prehistory, it is critical to recognize and to describe native copper in the unmelted state and after melting, to distinguish and to set up markers for several ore types and to assess the distribution of impurities remained concentrated in the copper objects.

In this archaeo-metallurgical project, 100 copper and bronze artifacts belonging to several museums from Transylvania have been investigated by ICP-AES and ICP-MS in the framework of project financed by the European Community - Access to Research Infrastructures Action of the Improving Human Potential Programme, Contract HPRI-CT-1999-00008 awarded to Prof. B.J. Wood from the EU Geochemical Facility, University of Bristol, UK. For this study have been chosen 31 artifacts of high purity concentration in copper and another 9 samples of native copper in unmelted and melted status.

## STATISTICAL APPROACH

Two main goals have been set up in this research:

1. What is the geographical component in the classification?
2. What are the major type of copper artifacts and their chief characterizations and best representations?

As software, has been used Clustan Graphics package elaborated by D. Wishart and SPSS 8.0.

The method was hierarchical agglomerative clustering analyses. Experimental matrix is consisted of 40 cases containing copper objects with 98-100% Cu and 5 variables. The variables chosen are those impurities which are considered to be markers for different ore types: Co, Fe, Sb, Ag and As. Standardization of data has been achieved by Z-score, dissimilarities between all pairs of individuals have been measured by squared Euclidian distance defined as:

$$d_{2h}^2 = \frac{\sum_j (x_{ij} - x_{hj})^2}{v}$$

Where for each variable  $j$ ,  $x_{ij}$  and  $x_{hj}$  are the values in cases  $i$  and  $h$  and the summation is over all  $v$  variables. This method has been selected as the main interest was to find clusters that are relatively homogenous with respect to all variables.

Objects have been clustered by Ward's method or the method of increase in sum of squares. Increase in sum of squares assumes that the cases can be represented by points in Euclidian space and requires a proximity matrix of Squared Euclidian Distances.

$$E_p = \frac{\sum_i \epsilon_p \sum_j (x_{ij} - \mu_{pj})^2}{v}$$

$$I_{p \cup q} = E_{p \cup q} - E_p - E_q$$

The Euclidian Sum of Squares  $E_p$  for a cluster  $p$  is the sum of the Squared Euclidian Distances between all members of cluster  $p$  and is represented by the formula above, where for each variable  $j$ ,  $x_{ij}$  is the value in case  $i$ ,  $\mu_{pj}$  is the mean in cluster  $p$  and  $v$  is the number of variables. The total Euclidian Sum of Squares over all clusters for a given classification is  $E = \sum_p E_p$ . Increase in Sum of Squares combines two clusters  $p$  and  $q$  which result in the least increase  $I_{p \cup q}$  in  $E$  that is for which  $I_{p \cup q} = E_{p \cup q} - E_p - E_q$  is minimum. Increase in Sum of

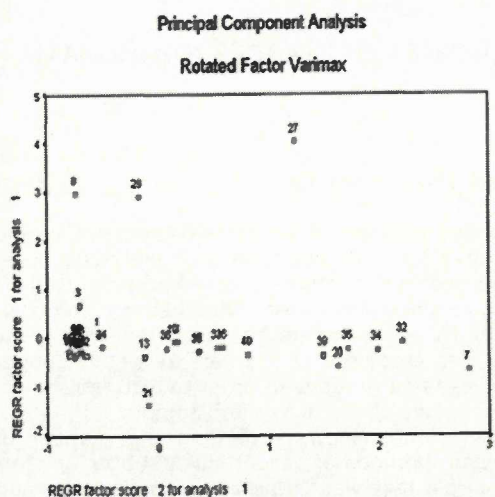


Figure 1 Principal component analyses by SPSS 8.0

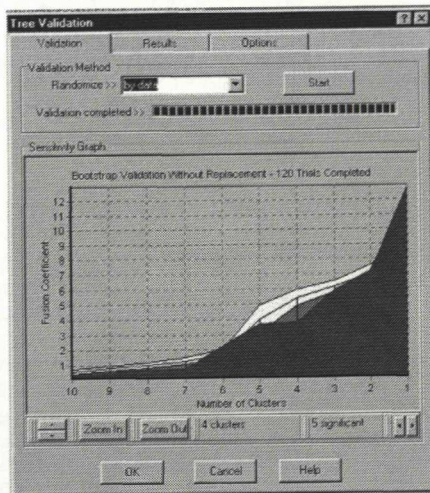


Figure 3 Tree Validation option in Clustan Graphics

Squares assumes that the cases can be represented by points in Euclidian space and requires a proximity matrix of Squared Euclidian Distances. *Clustan Graphics* can convert similarities to dissimilarities and this conversion is invoked automatically if Increase in Sum of Squares is selected with a similarity matrix (Wishart 2003:29).

METHODS OF VALIDATION APPLIED

The real number of clusters is difficult to assess from the typical output of the dendrogram. The success of archaeological cluster analyses is connected with the ability to reproduce known archaeological groups. In the case of the provenancing studies it is confirmed that archaeologically-defined groups are chemically distinct in some cases (Baxter 1994:164).

A feed-back procedure is represented by the introduction of individuals known to group together, as a subset of the total number of cases studied. Native copper material has been introduced to monitor the clustering. Cases 32-40 are the samples of native copper, which have grouped together .

Graphical procedure by principal component analyses has been used to supplement cluster analysis. Clear structure on the component plot serve to confirm the reality of structure and offers additional information, if the cluster analysis is partitioning the data rather than identifying distinct clusters (Fig.1). Absence of structure on the plots does not however, imply that there are no clusters, since these may well exist in a higher number of dimensions than are capable of the representations on the plots.

Heuristic stopping rules for determining the number of clusters have been implemented by D. Wishart in *Clustan Graphics*. The real number of clusters in this case has been obtained by applying the Best Cut option and the significance tests to the series of fusion levels in the current tree. By choosing signi-

ficance test from the list a table shows the proposed cluster partitions in the current tree, which are significant. The corresponding partition of the tree is shaded (Fig.3) and the partition is saved as current cluster model.

Another procedure which can be used in *Clustan Graphics* is the Tree Validation to test the best number of clusters in the hierarchical classification. It compares the tree obtained for the data set with the family of trees generated by random permutation of the same data or the associated proximity matrix. A distribution is obtained for the set of trees from the randomly permuted data and a confidence interval is constructed about the mean. The tree for the given data is then compared with the confidence interval and significant departures from random are identified. Validate tree seeks to reject the underlying hypothesis that the data are randomly distributed. It searches for tree sections that correspond to the greatest departure from randomness, and in trials when random data evaluated it reassuringly reported no significant clusters (Wishart 2003:46).

Randomize can be done by data or by proximities with or without replacements.

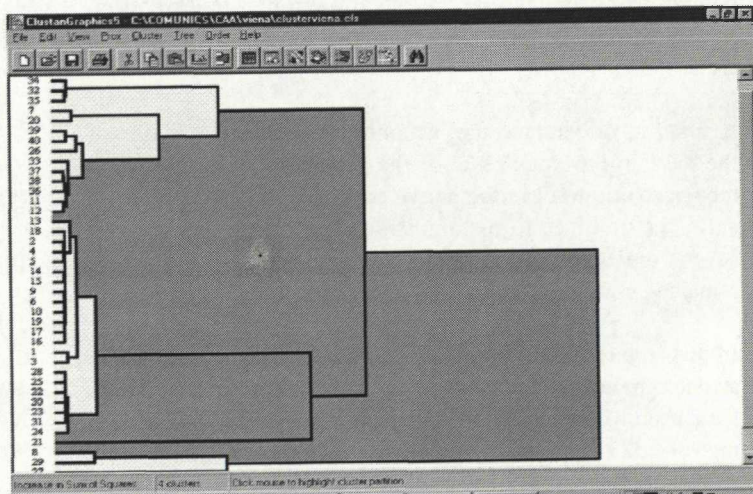


Figure 2 Dendrogram obtained by Tree validation

CONCLUSIONS

In this case study, satisfactory results have been obtained by combining principal component analyses with the clustering and validation methods implemented in *Clustan Graphics*.

As proposed by several scholars there is no generally applicable solution to determining the appropriate number of clusters, each case should be analyzed and checked by other additional methods. Reliance on the form of the dendrogram

is unsafe therefore methods combining informal and subjective criteria based on subject expertise should be taken into consideration when dealing with statistical analyses of archaeological data, especially of chemical composition data.

#### Total Variance Explained

Component	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.005	40.092	40.092	2.005	40.092	40.092	1.651	33.014	33.014
2	1.350	27.003	67.096	1.350	27.003	67.096	1.251	25.015	58.029
3	.748	14.962	82.058	.748	14.962	82.058	1.201	24.029	82.058
4	.666	13.327	95.385						
5	.231	4.615	100.000						

Extraction Method: Principal Component Analysis

#### Component Matrix

	Component		
	1	2	3
CO	.534	-.487	.548
FE	.489	.596	.441
SB	.913	-.106	-.129
AG	-7.546E-02	.848	8.029E-02
AS	.800	.161	-.480

Extraction Method: Principal Component Analysis.

a 3 components extracted

#### Rotated Component Matrix

	Component		
	1	2	3
CO	.135	-1.795E-02	.897
FE	.214	.833	.221
SB	.836	7.175E-02	.398
AG	--5.887E-02	.735	-.433
AS	.940	.101	-4.421E-02

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 5 iterations

#### REFERENCES

BAXTER, M.J., 1994. Exploratory multivariate analysis in Archaeology. Edinburgh.

HANCOCK, R.G.V., PAVLISH, L.A., FARQUHAR, R.M., SALLOUM, R., FOX, W.A. and WILSON G.C., 1991. Distinguishing European trade copper and north-eastern North American native copper. *Archaeometry* 33:69-86.

MADDIN, R., MUHLY, J.D. and WHEELER, T.S., 1980. Distinguishing artifacts made of native copper. *Journal of archaeological Science* 7:211-225.

RAPP, Jr.G., 1982. Native copper and the beginning of smelting: chemical studies. In *Early Metallurgy in Cyprus, 4000-500 B.C.*, Nicosia:33-40.

WAYMAN, L.M., and DUKE, M.J.M., 1999. The effects of melting on native copper - The Beginnings of Metallurgy. *Proceedings of the International Conference, Bochum 1995:55-67.*

WISHART, D., 2003. *Clustan Graphics Primer: A Guide to Cluster Analysis Second Edition.*