Multivariate cluster analysis of trace elements and mineralogical components from some rural soils

M.M. Jordán Vidal, a J. Mateu Mahiques, b A. Boix Sanfeliu a
a Universitat Jaume I, Departament de Ciències Experimentalles, Campus Borriol, 12071 Castellón, Spain
b Universitat Jaume I, Departament de Matemàtiques, Campus de Penyeta Roja, 12071 Castellón, Spain

Abstract

During the 1994-95 two-year period, a research project has been carried out to characterise the mineralogical and chemical composition of some rural soils located close to five clay outcrops that supply the Castellon (Spanish province) ceramic industry. A total of 50 soil samples were collected in accordance with statistical criteria and their chemical elemental composition were determined by ICP-AES. The mineralogical composition of each sample was discovered using X-ray diffraction. Semi-quantitative mineralogical analysis was done following the Chung method. A hierarchical cluster analysis using Ward method has been carried out to detect statistical significative groups (clusters) and a discriminant analysis of proximities to obtain probabilities. Univariate statistical descriptions have been computed for all variables.

Statistical procedures have been implemented using the software package SPSS for Windows release 5.0.1. This is regarded as a useful software for environmental studies.

1 Introduction

In the province of Castellón, the expansion of agriculture toward the very sloping hillsides, deforestation, shepherding and fundamentally the forest fires are the causes that have produced the loss of soil, in some instances irreversible, and the road to desertization. This situation gets worse with ceramic clays extraction in the study areas, since the impact on soil can be qualified as negative and serious due to ground occupation in making hollows, building access tracks and rural roads and accumulating material stocks on the ground. Furthermore, these soils can be contaminated by solids in suspension contribution, organic material, fats and lubricants derived from the machinery used in the extractive processes.
All over this, during the biennium 1994-95 the Crystallography and Mineralogy area of the Experimental Sciences Department of the Jaume I University accomplished an investigation to characterize the mineralogical and chemical composition, from majority elements to traces, of rural soils next to five clays outcrops that supply the ceramic industry of the province of Castellón. The studied area extends to the townships of Forcall (series FM), Argelita (series LA), Zucaína (series ZC), La Jana (series LA) and Cervera (series CE).

2 Experimental

A total of 50 samples were selected, 10 for each deposit. Each classified and numbered sample and of an approximate quantity of 5 Kg has been selected for its study in the laboratory.

2.1 Mineralogical Analysis

The procedure employed in the mineralogical analysis is the breaking-up method described by Hathaway. The customary methods in the study of the clays (Carrol [2]) have been practiced, on the treated samples and without treatment making orientated aggregate preparations (normal, heated to 550 °C during 2 hours and treatment with ethylenglicol during two hours previously to the difractometric record).

For the orientated aggregates obtainment, we disperse the clay fraction of each one of the samples in distilled water (Ballbe [1]). To apply this technique the record of the X-ray graphs at ambient temperature (18-22 °C) was accomplished using the powder diffractometer Siemens D-5000 with Bragg-Bretanó geometry, graphite (monocromador) and twinkle detector. Also, a graph for the identification of the mineralogical components of the total fraction was accomplished. The spectra were taken with of Cu, K (1,54052) radiation to 40 KW and 20 mA. A normal graph from 4 to 70 of 2 using Ni filter , 0,050 steps, constant of time of 3 seconds and incidence splits combination(1 mm), of divergence (1 mm) and detector of 0,2 mm with continuous record modality was accomplished.

The identification of the present mineral phases in each one of the samples has been accomplished through the utilization of the computer program EVE of the Socabim company of the IJA and the JCPDS index cards. Through the mentioned program the elimination of corresponding diffractometer background, the diffraction peaks selection and phases search was accomplished, afterwards completed with the JCPDS diffraction index cards consultation.

A diffraction detected phases quantification has been accomplished. This is not in absolute terms quantification, but it is comparative between the different samples, reflecting the percentage of each substance respect to the total identified substances (Chung [3]).
2.2 Chemical Analysis

The instrumental analytical technique used to know the studied samples chemistry composition has been the inductive coupling plasma emission spectroscopy (ICP-AES). The inductive coupling plasma spectrometer (ICP) used corresponded to the Jobin-Yvon company, model JY-38 VHR.

The samples were put on dissolution with hot concentrated HCl, diluting afterwards with water until an appropriate volume. The used standards were containing the assault reactives of the sample to eliminate the counterfoil effect. In each one of the resulting solutions of the described protocol the following elements were determined: Al, C, Mg, K, N, Ti, Fe, Mn, Cu, Zn, Co, Ni, V, Sr, P, B, and S.

2.3 Statistical Methodology

Cluster analysis

Cluster analysis is concerned with the discovering group structure amongst the cases of a n by p matrix. Two general references are Gordon⁴ and Hartigan⁵. Almost all methods are based on a measure of the similarity or dissimilarity between cases. A dissimilarity coefficient \( d \) is symmetric (\( d(A,B)=d(B,A) \)), non-negative and \( d(A,A) \) is zero. A similarity coefficient has the scale reversed. Dissimilarities may be a metric:

\[
d(A, C) \leq d(A, B) + d(B, C)
\]

or an ultrametric:

\[
d(A, B) \leq \max(d(A, C),d(B, C))
\]

but need not be either.

There are several families of similarity and dissimilarity measures to produce clusters. Similar objects should appear in the same cluster and dissimilar objects in different clusters. A well known interobject measure of similarity is the correlation coefficient between a pair of objects measured on several variables. Correlational measures represent similarity by the correspondence of patterns across the characteristics (X variables) and are not influenced by differences in scales between objects. But correlational measures are rarely used because emphasis in most applications of cluster analysis is on the magnitudes of the objects, not the patterns of values. Distance measures of similarity, which represent similarity as the proximity of observations to one another across the variables in the cluster variate, are the similarity measure most often used. These measures, like Euclidean, Gamma, Pearson..., are significantly affected by differences in scale so before computing them, it is needed to standardize the data across the measured attributes. The most commonly used is the Euclidean distance.
On the other hand, various linkage methods can be used to compute the distance of one object or cluster from another and to determine whether the two should be merged in a given step. The *Centroid Linkage* method uses the average value of all objects in a cluster as the reference point for distances to other objects or clusters. The *Ward’s method* (Ward [8]) resembles Centroid Linkage but adjusts for covariances.

Cluster algorithms can be classified into two general categories: *hierarchical* and *nonhierarchical*. We’ll just discuss the former. Hierarchical procedures involve the construction of a hierarchy of a treelike structure. There are basically two types of hierarchical clustering procedures: *agglomerative* and *divisive*. In the agglomerative methods, each object or observation starts out as its own cluster. In subsequent steps, the two closest clusters are combined into a new aggregate cluster. Eventually, all individuals are grouped into one large cluster. This process is usually shown in a tree graph or dendrogram. In divisive methods, we begin with one large cluster containing all observations. In succeeding steps, the observations that are most dissimilar are split off and made into smaller clusters.

**Multiple discriminant analysis**

Discriminant analysis is the appropriate statistical technique when the dependent variable is categorical and the independent variables are metric. In many cases, the dependent variable consists of two groups or classifications. In other instances, more than two groups are involved, such as a three-group classification. Discriminant analysis is capable of handling either two groups or multiple groups. When two classifications are involved, the technique is referred to as *two-group discriminant analysis* (Hair et al [5]). When three or more classifications are identified, the technique is referred to as *multiple discriminant analysis* (MDA).

Discriminant analysis involves deriving a *variate*, the linear combination of the two (or more) independent variables that will discriminate best between a priori defined groups. Discrimination is achieved by setting the variate's weights for each variable to maximize the between-group variance relative to the within-group variance. The linear combination for a discriminant analysis, also known as the *discriminant function*, is derived from an equation that takes the following form:

\[
Z = W_1X_1 + W_2X_2 + \ldots + W_nX_n
\]

where

- \(Z\) = Discriminant score
- \(W_i\) = Discriminant weight for variable \(i\)
- \(X_i\) = Independent variable \(i\).
Suppose that $g$ is the number of groups, and for each case we know the group (assumed correctly). We can then use the group information to help reveal the structure. Let $W$ denote the within-group covariance matrix, that is the covariance matrix of the variables centred on the group mean, and $B$ denote the between-groups covariance matrix, that is of the predictions by the group means. Let $M$ be the $g \times p$ matrix of group means, and $G$ be the $n \times g$ matrix of group indicator variables (so $g_{ij}=1$ if and only if case $i$ is assigned to group $j$). Then the predictions are $GM$. Let $\bar{x}$ be the means of the variables over the whole sample. Then the sample covariance matrices are

$$W = \frac{(X - GM)^T (X - GM)}{n - g}, \quad B = \frac{(GM - 1\bar{x})^T (GM - 1\bar{x})}{g - 1}$$

If there are more than two groups in the dependent variable, discriminant analysis will calculate more than one discriminant function. As a matter of fact, it will calculate $g-1$ functions. Each discriminant function will calculate a discriminant score. In the case of a three-group dependent variable, each object will have a score for discriminant functions one and two, allowing the objects to be plotted in two dimensions, with each dimension representing a discriminant function.

### 2.4 Computer Comands

The SPSS software package provides many useful statistical techniques to look, explore, analyze and present environmental data. Some of them are used in this paper and the SPSS comands to run these methods are provided here.

#### a) Cluster analysis

```spss
PROXIMITIES
variable names /MATRIX OUT ('C:\WINDOWS\TEMP\spssclus.tmp') /
VIEW=CASE
/MEASURE=SEUCLID /PRINT NONE /STANDARDIZE=NONE.
CLUSTER /MATRIX IN ('C:\WINDOWS\TEMP\spssclus.tmp') /METHOD
WARD
/PRINT SCHEDULE /PLOT DENDROGRAM.
ERASE FILE='C:\WINDOWS\TEMP\spssclus.tmp'.
```

#### b) Discriminant analysis

```spss
DISCRIMINANT /GROUPS=grouping variable /VARIABLES=variable names
/ANALYSIS ALL /STATISTICS=MEAN STDDEV UNIVF COEFF
/CLASSIFY=NONMISSING POOLED.
```
3 Results and Discussion

The X Ray Diffractograms analysis of the clay fraction shows the predominantly illitic character of all the analyzed samples. Chlorite and the Kaolinite are also present. Quartz is also detected as integrating the clay fraction.

The semi quantitative analysis of the total fraction has stated a great similarity in the LA and ZC series mineralogical composition. Both possess a strong illitic character (40-54 %), a content in quartz around 30 % and a total percentage of Chlorita + Caolinita around 10 %. The medium content in Hematites is found about 5 %. The potasinic feldspars prevail as compared to the plagioclases though in no case surpass 7 %. The samples belonging to the CE series possess a high content in clay’s minerals. The Quartz does not exceed 21% in any case. The Illita and Chlorita+Caolinita percentages are located around 45 % and 30 % respectively. In most of the samples it is not detected the Hematites, Dolomite or Plagioclases presence. The Calcite content is similar to the LA and ZC series, no surpassing 2 %. These samples can be classified as illitic-caolinitic clays with under content in sand fraction. The series JT is characterized by a percentual similarity concerning Illite and Caolinite+Clorite located around 30 %. The content in Quartz is variable, oscillating between 27 % in JT1 sample and 44 % in JT8 and JT9 samples. This series is the one which possesses a greater content in carbonates as much in the form of Calcite as Dolomite reaching a 14 % in the JT7 sample for the Calcite and a 3 % for the Dolomite. On the contrary in JT8, JT9 and JT10 samples only have indicia. The FM series is quite similar to LA and ZC series, whose results have been discussed previously. However, it emphasizes the presence of potasinc feldspars as important part of the sand fraction.

About the mineralogy of the sandy fraction, at the sight of the results, we observe that this fraction is constituted mainly by Quartz followed quantitatively by the Illite, that in some samples and in the fraction fine sand, arrives to surpass even the Quartz percentage in the same fraction.

Between the analyzed trace elements emphasize by its relative importance and "abundance" (> 100 ppm) the following elements: P, B, S and Mn. Consequently, the trace element analysis results has been mainly focused in the discussion of the presence of these 4 elements in each one of the considerate series. The phosphorus analysis (P) states that its content varies sensibly of
some samples to other, no existing a clear concentration of this element in given series. Something similar occurs when it is analyzed the Barium (B), whose content varies from 145 ppm analyzed in the CE5 sample and 506 ppm of LA4. It is not observed a concentration of this element in a given series either. The S is an essential element in the nourishment of the plants, and results indispensable for the cysteine and certain vitamins synthesis. When it is analyzed the S, is observed the concentration of this element in the CE series, arriving to surpass 5000 ppm in the CE2 sample and being its concentration above 3000 ppm in CE5, CE6 and CE7. In the rest of analyzed series the concentrations are found far below 300 ppm. Thus, the samples of the LA series only exceed 100 ppm in the LA2 sample, while the samples of the ZC and FM series are found below 200 ppm and the corresponding samples to the JT series do not reach 200 ppm. The manganese analysis (Mn) permitted to observe a meaningful increase of this element in samples of the CE (643 ppm in CE7) and FM (573 ppm in FM3) series, as well as an important decrease of this element when samples of the ZC series are analyzed. The LA and JT series possess variable concentrations according to the analyzed sample.

A statistical cluster analysis of data provided by elemental chemical analysis (ICP-AES) as well as by quantitative mineralogical analysis (Chung [4]) has been accomplished. Since the analyzed data suggest the existence of more than a form of soil, from the geological age point of view, it has been turned to the cluster or grouping analysis and to the discriminant analysis for its classification. Previously, univariate descriptions of the variables have been calculated.

In first place, it was accomplished a cluster analysis of the samples, through the Ward's hierarchical agglomerative method based on the euclidean distance. Two types of dendrograms have been accomplished: the first corresponds to the grouping of each one of the samples according to its chemistry composition (Fig.1) and the second corresponds to the grouping of these according to its mineralogical composition (Fig.2).

The samples classification according to its chemistry drove to establish three well differentiated groups. As checking, a discriminant analysis was accomplished. Group I consisted of ZC series and LA series. The group II gathered samples of the FM and JT series, being the FM series samples closer to the group I than the ones corresponding to the JT series. It is convinient to indicate that the groups I and II could be included in only one group since, as it is shown in the dendrogram, the differences are minimal. In fact, after a T-test, there are only meaningful differences between both groups for the Mn, Zn, Co and Th variable. Finally, to indicate that the group III is constituted by the samples of the CE series. These samples have some typical compositional characteristics by its high content in S and alkalinoterreous elements. The averages and standard deviations for each group and only for trace elements come given in table 1. Differences between groups I and II with respect to III,
are confirmed with a new T-test for independent samples given a statistical significance value of p=0.01.

The classification or grouping of the samples attending to its mineralogical composition permitted to establish three groups, see the dendrogram of the Fig.2, confirmed by the applied discriminant analysis. This grouping coincided with the accomplished one attending to the chemistry composition of the samples solely in the establishment of the group I (ZC series + LA series ). The group II gathered solely the samples of the JT series with the exception of the FM1 sample that was included in this group. The group III gathered the samples of the CE and FM series.

To confirm the existence of meaningful differences between the three formed groups, an analysis of the variance (ANOVA) for each one of the variables under study was made. In all the variables, except in F (Feldspars) in the one which is reached a statistical significance level of p=0.56, there are statistical significant differences between the three groups formed by the cluster analysis with very close statistical significances to the p=0.0. The univariates descriptions for each group and variable come given in table 2.

### Table 1. Univariate descriptions for groups of chemical trace elements.

<table>
<thead>
<tr>
<th>Element</th>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std. dev</td>
<td>mean</td>
</tr>
<tr>
<td>Mn</td>
<td>116.87</td>
<td>72.16</td>
<td>310.38</td>
</tr>
<tr>
<td>Cu</td>
<td>4.66</td>
<td>4.26</td>
<td>28.89</td>
</tr>
<tr>
<td>Zn</td>
<td>24.69</td>
<td>11.3</td>
<td>50.7</td>
</tr>
<tr>
<td>Co</td>
<td>1.16</td>
<td>1.86</td>
<td>7.96</td>
</tr>
<tr>
<td>Ni</td>
<td>22.08</td>
<td>7.1</td>
<td>31.9</td>
</tr>
<tr>
<td>V</td>
<td>76.47</td>
<td>26.9</td>
<td>69.75</td>
</tr>
<tr>
<td>Sr</td>
<td>75.42</td>
<td>21.75</td>
<td>125.64</td>
</tr>
<tr>
<td>P</td>
<td>234.71</td>
<td>82.4</td>
<td>301.41</td>
</tr>
<tr>
<td>Ba</td>
<td>355.57</td>
<td>98.11</td>
<td>315.43</td>
</tr>
<tr>
<td>S</td>
<td>84.41</td>
<td>69.62</td>
<td>138.52</td>
</tr>
</tbody>
</table>
Table 2. Univariate descriptions for groups of mineralogic elements.

<table>
<thead>
<tr>
<th>Element</th>
<th>Group I mean</th>
<th>std.dev</th>
<th>Group II mean</th>
<th>std.dev</th>
<th>Group III mean</th>
<th>std.dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>32.05</td>
<td>4.17</td>
<td>18.89</td>
<td>3.21</td>
<td>32.8</td>
<td>8.52</td>
</tr>
<tr>
<td>Cc</td>
<td>1.45</td>
<td>1.39</td>
<td>0.77</td>
<td>0.64</td>
<td>4.5</td>
<td>4.32</td>
</tr>
<tr>
<td>D</td>
<td>0.65</td>
<td>0.74</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.99</td>
</tr>
<tr>
<td>M + I</td>
<td>47.1</td>
<td>4.97</td>
<td>43.83</td>
<td>2.95</td>
<td>26.7</td>
<td>6.81</td>
</tr>
<tr>
<td>C + K</td>
<td>9.1</td>
<td>4.22</td>
<td>29.22</td>
<td>3.45</td>
<td>28.1</td>
<td>6.24</td>
</tr>
<tr>
<td>Hm</td>
<td>4.95</td>
<td>3.06</td>
<td>2.44</td>
<td>2.45</td>
<td>2.4</td>
<td>1.77</td>
</tr>
<tr>
<td>F</td>
<td>4.7</td>
<td>2.73</td>
<td>4.55</td>
<td>2.09</td>
<td>3.7</td>
<td>2.45</td>
</tr>
</tbody>
</table>

References

Figure 1: Dendrogram using Ward method for chemical elements.

Figure 2: Dendrogram using Ward method for mineralogic elements.