CONTENTS

Chapter 4 Multivariate morphological classification

4.1 Introduction

The National Soil Inventory (NSI) contains a profile description for each sampling site. The descriptions include records of the pedological properties of the topsoil. These are not of great interest individually, but together they constitute the data on which soil maps are made by morphological classification. They also form the basis upon which new profiles are allocated to an existing classification.

The soil surveyors, when they described the sites for the NSI, assigned the profiles as they observed them to the classes of the national soil classification (Avery, 1980b). The relations between the assignments, as representatives of the national classes, and the geochemistry are described throughout this report.

In this Chapter we describe how we have analysed the morphological data to produce a classification that is in a sense optimal (Chapter 2). We also assess its performance in relation to the geochemistry.

4.2 The data

The morphological data comprise 21 variates. Some are measured, some are recorded as multistate characters, and some are binary (present or absent). The variates and their types are listed in Table 4.1. Methods for classifying individuals from such mixed data are available. However, the packages in which they are programmed will handle, in general, no more than a few hundred units, because they are based on a similarity matrix of size N ? (N+1)/2, where N is the number of units. For our data - with N=5671 - the size of the matrix would be too large for all but a supercomputer to handle. Therefore, we proceeded in a sequence of steps to circumvent this apparent impasse.

4.3 Principal Coordinates Analysis (PCO)

4.3.1 Method

The first step in the analysis was to convert the mixed set of data into a fully metric system. We did this using Gower's (1966) method of principal coordinates in which a similarity matrix is transformed into a Euclidean framework. In this method

similarities between units (sites) are computed using Gower's (1971) general similarity coefficient:

$$S_{ij} = \frac{\sum_{k=1}^{P} x_{ijk} w_{ijk}}{\sum_{k=1}^{P} w_{ijk}}, \qquad (4.1)$$

where x_{ijk} is a value for the comparison of the kth variate between units i and j, and w_{ijk} is the weight assigned to it. For continuous variables:

$$x_{ijk} = 1 - \frac{|z_{ik} - z_{jk}|}{r_k} \qquad (4.2)$$

where r_k is the range of the variable. For multistate characters, $x_{ijk} = 1$ if $z_{ik} = z_{jk}$ and 0 otherwise. Variables compared by matching in this way are designated in Table 4.1 by the word 'matching', and they include several binary variables. For some binary variables, presence is significant - but absence is not. For these, $x_{ijk} = 1$ if both are present, $x_{ijk} = 0$ if one unit has the character and the other does not, and no comparison is made if both lack the character. These are denoted in Table 4.1 by the word 'Jaccard'.

Following Gower (1966), we converted the similarities, S_{ij}, into dissimilarities or 'distances':

$$d_{ij} = \sqrt{2(1 - S_{ij})}, \qquad (4.3)$$

to give a dissimilarity matrix. After adjusting the elements of this matrix and scaling them, as described in Chapter 2, we obtained a matrix, the eigenvectors of which contain the co-ordinates of the units in a Euclidean space. The first eigenvector defines the dimension in which the variance is largest, the second, orthogonal to the first, accounts for the second largest contribution to the total variance, and so on.

Since the similarity matrix for the whole set of sites in the NSI would be 5671 × (5671 + 1)/2 we made use of another development due to Gower (1968) in which points can be added to a principal coordinate analysis.

We selected, as a representative set of sites, those at 30-km intervals on the original grid; i.e. every sixth point in each row and column. This gave us 155 sites. We computed the similarities between all pairs using the equation above, and transformed

the resulting similarity matrix into the principal coordinates. These defined the Euclidean space of the coordinate system, and the remaining points were added in batches of 100 to this space. The procedure was implemented by the directive (`Addpoint`) in Genstat. In this way we obtained the principal coordinates for all 5671 sites.

4.3.2 Results of the principal coordinate analysis

Table 4.2 gives the eigenvalues and the percentage variance accounted for by the leading coordinates. The scatter of the 155 sampling sites in the plane of the two leading principal coordinates is shown in Figure 4.1. It appears as a single cloud of points with no evident gaps or clusters. Figure 4.2 shows the scatter of all 5671 points in the same plane. The general configuration is the same, and the points occupy the same part of the space, and there are no outliers. This shows that the sample of 155 sites spans the space adequately; it was a properly representative sample of the full data.

4.4 Classification

4.4.1 Method

Given the lack of evident clusters in the scatter of points in the space of the principal coordinates, classification might appear entirely arbitrary. Nevertheless, it is still possible to seek a classification at a single categoric level that is optimal, in the sense of minimizing some function of within-group variation or between group variation or a combination of the two. There are numerous variations on the basic method known as *k*-means classification (Chapter 2), and we have used the implementation already embodied in Genstat. The starting point in this method is a data matrix of N units by P quantitative variates. With N = 5671 units this process would be enormously time-consuming, and we therefore followed the same logic as when computing the principal coordinates. In addition, the method requires quantitative variables. Therefore, it would not have been possible with the raw pedological data. We classified the sample of 155 sites on their first 10 principal coordinates.

The 155 units were partitioned into g = 10 classes and the chosen criterion for classification was computed. We assigned the 155 units to the classes in approximately equal proportions, simply in the order in which they appeared in the

file. We used the determinant of the within-class sums-of-squares-and-products (SSP) matrix, |**W**|, as the criterion of optimality. Units were then transferred from group to group and exchanged, and at each transfer or exchange |**W**| was recomputed. If it was smaller than the former value the change was kept; otherwise the earlier partition was restored. The procedure continued until no further improvement seemed possible, and the classification into the 10 classes was regarded as optimal. The most similar pair of classes was then combined, and the procedure repeated with g now one less than the original partition. Again the most similar pair of classes was combined, and the whole cycle of iterative transfer computation of the criterion for optimality and combination can be continued until there were only two groups.

To the optimal classification based on 155 sites we allocated the remaining 5516 sites to the classes created by multiple discriminant analyses.

4.4.2 How many classes?

In principle, one can have as many optimal classifications as there are classes. Some make more sense than others, and one way of choosing a particular classification is to plot the criterion value, or some function of it, against the number of classes, and choose that value of g for which the function falls farthest below any trend. When |**W**| is the criterion, the graph to plot is of $g^2|\mathbf{W}|$ against g, and Figure 4.3 shows this graph for the classification of the 155 sites. Two values of the function stand out, namely those at $g = 4$ and $g = 6$.

4.4.3 Canonical variates

To see how the optimal classification at $g = 6$ performed, we transformed the space of the principal coordinates into that of canonical variates for the classification. Table 4.3 summarizes the canonical variate analysis: the first two canonical variates account for 97.7 % of the variance. Figure 4.4 shows the scatter of the 155 sites in the transformed space, using point symbols. The circles are drawn with their centres at the mean canonical points in the space and with radius 2.45, which is $?^2$ with two degrees of freedom at the 95 % confidence level. Thus, each circle encloses 95 % of the distribution around its centroid.

Figure 4.4 shows three distinct classes, top left and right and bottom right. The three groups of points, and their associated circles in the lower left quadrant, overlap one

another in the space. By plotting canonical variate 2 against variate 3, instead of variate 1, we see that they are separate, though still close to one another. We could have considered the three classes as a single class, which would accord with the trough in the graph of $g^2|\mathbf{W}|$ against g (Figure 4.3) at four classes. However, the separation of the classes in this plane suggests that a classification based on four classes would account for less of the variation than one based on six classes.

At the final stage in the classification we allocated the remaining 5516 sites to the six classes by calculating the Mahalanobis distance, D_M, for each, between it and every class centroid, and assigning it to the class for which D_M was least. Figures 4.5 and 4.7 show the scatter of all sites allocated to the optimized classification in the canonical variate space. The circles are drawn with their centres at the mean canonical points in the space. The circles in these figures are those for the 99% confidence level with radius 3.03; each circle encloses 95% of the distribution around its centroid. The groups have a very similar distribution to those for the analysis with 155 sites. The canonical variate analysis was then repeated, giving new centroids, and the Mahalanobis distance between each site was recalculated and any site in the 'wrong' class for which its D_M was least. Figures 4.6 and 4.8 show the final classification after the iteration (reallocation) described above. There is little difference between these figures and Figures 4.5 and 4.7. Figures 4.6 and 4.8 show a somewhat more compact distribution of the sites, however.

4.4.4 Analysis of variance

An analysis of variance was performed for selected elements and properties based on the six classes of the classification described above. Table 4.4 gives the results of this analysis.

Of the elements, $\log_{10}$Cr has the largest percentage variance explained, 20.6%, by these soil groups. The largest percentage variance accounted for is 27.4% for pH, and next is $\log_{10}$C with 21%.

For all of the variables listed, only $\log_{10}$Cr has a larger percentage variance explained by this classification than that using geology. This might be expected for the elements, but not for pH, clay and organic C, for which one would expect the soil morphology to provide a good description.

4.5 Summary

The results of the ordination show that soil data are poorly clustered in property space, which is what we have come to expect from quantitative analyses. The variation in terms of soil characters forms a continuum, and this may be particularly true when the greatest number of properties used to classify the soils comes from the topsoil only - as in this case of the NSI. This analysis indicates that there are no well defined soil groups, and any subdivision of the NSI sites is likely to be arbitrary. This does not preclude some form of subdivision, but it does suggest that some kind of optimizing procedure of the kind described above is the sensible approach. The results of the analysis of variance confirm that the subdivision does not account for a large proportion of the variance of individual properties. The use to which such a classification is put needs to be treated with caution in terms of interpreting the variation and of soil management. The morphological properties for selected individuals for each class were examined, but they showed no clear characteristics overall. Certain classes, such as group 4 is not gleyed, had reddish hues, and a large organic C content. Group 4 also has lower pH values. Group 3 has the largest pH values and the soil generally has a fine texture. Group 1 has sandy textured individuals and few mottles.

Table 4.1: Morphological data used for classification.

Variate	Type of variable	Matching
pH	Quantitative	Euclidean
Clay	Quantitative	Euclidean
Organic C	Quantitative	Euclidean
Ped grade	Multistate	Simple matching
Ped size	Quantitative	Euclidean
Silt	Quantitative	Euclidean
Very fine sand	Quantitative	Euclidean
Moderately fine sand	Quantitative	Euclidean
Medium sand	Quantitative	Euclidean
Coarse sand	Quantitative	Euclidean
Matrix value	Quantitative	Euclidean
Matrix chroma	Quantitative	Euclidean
Depth to gley	Binary	Jaccard
Depth to slightly gleyed horizon	Binary	Jaccard
Depth to skeletal gley	Binary	Jaccard
Depth to rock	Binary	Jaccard
Depth to slowly permeable layer	Binary	Jaccard
Carbonates	Multistate	Simple matching
Mottle abundance	Multistate	Simple matching
Stone abundance	Multistate	Simple matching
Matrix hue	Quantitative	Euclidean

Table 4.2: Eigenvalues and percentage variance for the leading principal co-ordinates.

	Eigenvalues	% variance
1	6.433	19.73
2	3.618	11.10
3	3.197	9.81
4	7.42	2.42
5	6.89	2.45

Table 4.3: Eigenvalues and percentage variance from the canonical variate analysis.

	Eigenvalues	% variance
1	52.98	61.95
2	31.15	36.43
3	1.083	1.267
4	0.3035	0.3549
5	0.00206	0.0024

Table 4.4: Analysis of variance for data classified according to soil morphological features.

Variable	Variance accounted for	Variable	Variance accounted for
Log_{10}Cd	2.47	K	6.92
Log_{10}ext. Cd	0.12	Log_{10} ext.K	5.30
Log_{10}Cr	20.6	Log_{10} P	2.93
Log_{10}Cu	0.59	Log_{10} ext. P	1.78
Log_{10}ext. Cu	5.22	Log_{10} Mg	13.5
Log_{10}Ni	15.8	Log_{10} ext.Mg	6.20
Log_{10}ext. Ni	6.13		
Log_{10}Pb	10.4	Log_{10}organic C	21.0
Log_{10}ext. Pb	10.1	pH	27.4
Log_{10}Zn	3.58	Clay	6.38
Log_{10}ext. Zn	2.67		

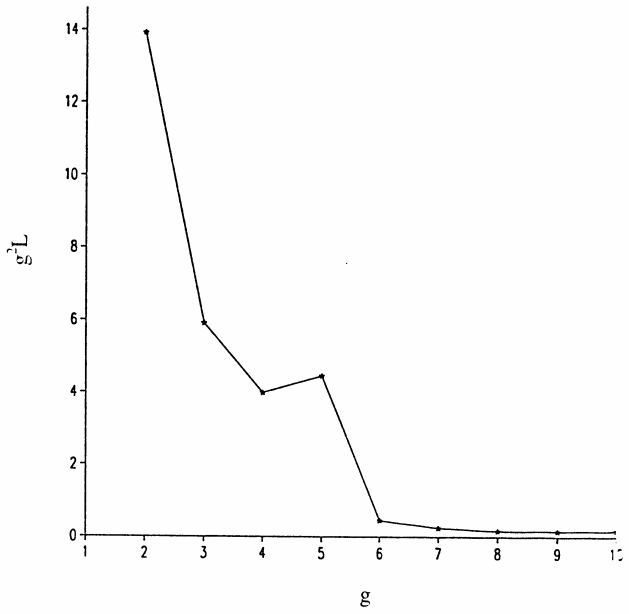

Figure 4.3: Plot of g^2L, where g is the group number and L is Wilk's criterion, against g.

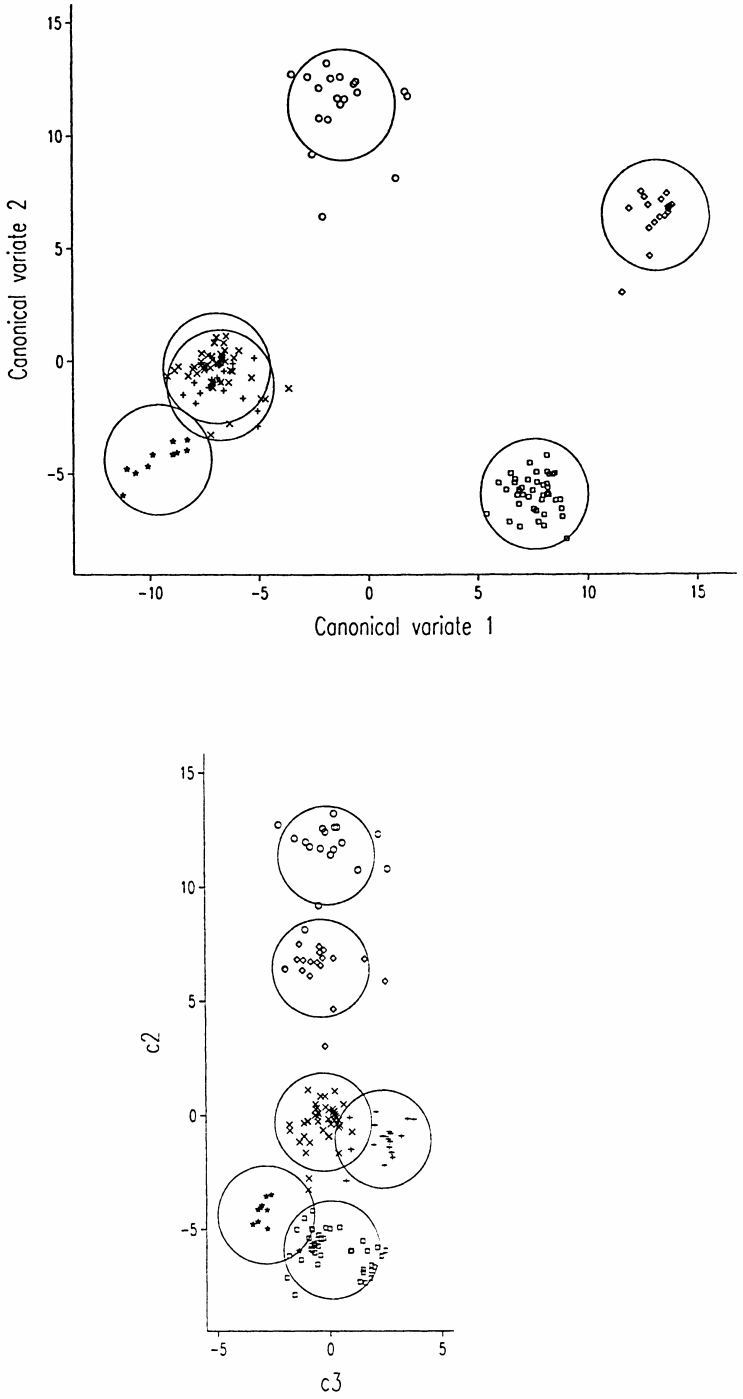

Figure 4.4: Scatter of the 155 sampling sites in the transformed space.

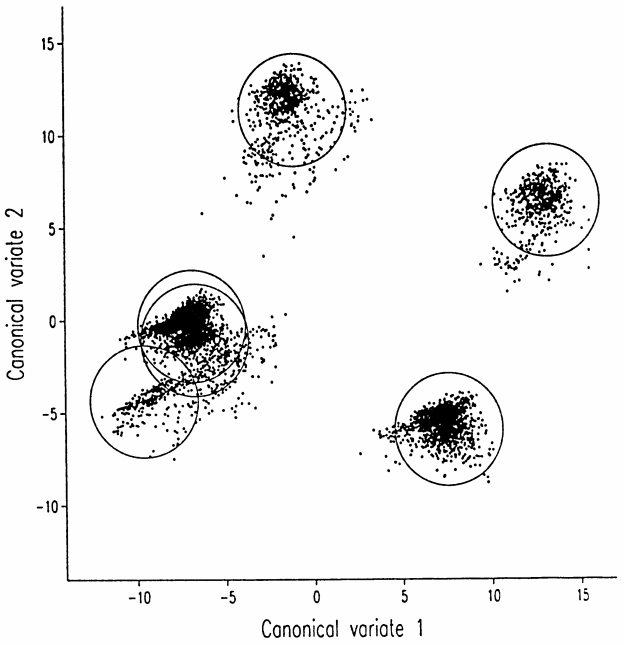

Figure 4.5: All sites allocated to the optimal classification based on the initial 155 sites and plotted according to their class in the plane of canonical variates 1 and 2.

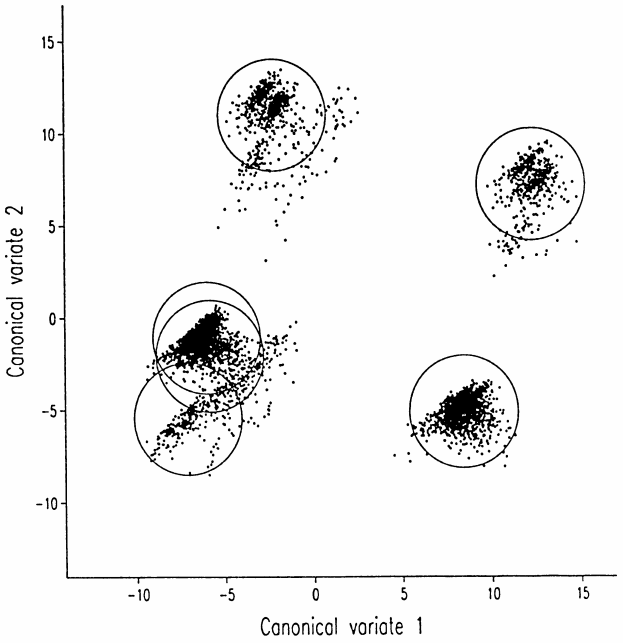

Figure 4.6: All sites allocated to the optimal classification based on the initial 155 sites after iteration, and plotted according to their class in the plane of canonical variates 1 and 2.

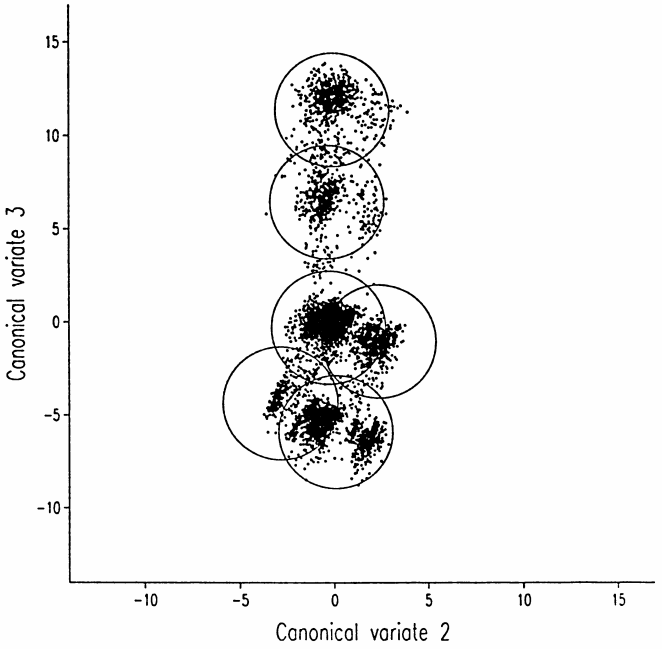

Figure 4.7: All sites allocated to the optimal classification based on the initial 155 sites and plotted according to their class in the plane of canonical variates 2 and 3.

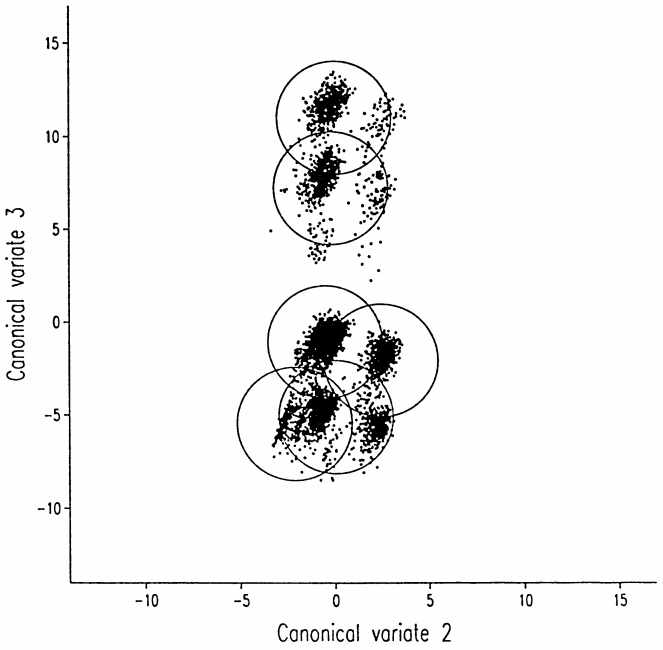

Figure 4.8: All sites allocated to the optimal classification based on the initial 155 sites after iteration, and plotted according to their class in the plane of canonical variates 2 and 3.