A model of discriminant analysis on the basis of descriptor variables for the ampelography of *Vitis* sp.

A. Camussi [1]), A. Calo [2]), A. Costacurta [2]), C. Lorenzoni [3]) and E. Ottaviano [1])

[1]) Dipartimento di Genetica e di Biologia dei Microorganismi, Università di Milano, Italy

[2]) Istituto Sperimentale per la Viticoltura, Conegliano, Italy

[3]) Istituto di Agronomia, Botanica e Genetica Vegetale, Università di Piacenza, Italy

Summary: Use of descriptor variables in ampelography is recommended to simplify recording of data and to enable useful comparisons. Parametric assumptions are, however, poorly satisfied especially with regard to statistical interference. In the paper some statistical procedures to improve the discriminant ability of descriptor variables are considered. The use of variances and covariances of variety by year interactions is suggested for the error matrix within a multiple discriminant analysis procedure. The adequacy of this model is verified in a 3-year experiment with Italian wine varieties. The discriminant power, as evaluated on the basis of the estimated distances among varieties, is satisfactory.

Key words: ampelography, shoot, leaf, berry, biometry, analysis, descriptor variables, multiple discriminant analysis, genotype by environment interaction, normality assumptions.

The basic aim of ampelography is the complete description of vine varieties in order to provide the most precise definition of their features. The identification of highly discriminating factors is important in order to derive an efficient and reproducible classification of varieties. Use of descriptor variables, as suggested by the O.I.V. protocols, is recommended to simplify data recording and to favor useful comparison. These variables, expressed as rating scales, are difficult to analyse statistically especially with regards to statistical inference.

In the present paper, some statistical procedures are considered to improve the discriminant ability of descriptor variables, particularly as regards data collected in different years of experimentation. As an example, discriminant analysis is applied to data from a set of 31 vine varieties evaluated over 3 years at the Istituto Sperimentale per la Viticoltura (Italy).

Multivariate approach: the canonical analysis

Discriminant analysis is a well established statistical procedure (see, for example, Srivastava and Carter 1983, 231-252), nevertheless a short summary of the main characteristics is given below.

Multiple discriminant analysis (often referred to as canonical analysis) is a very powerful tool used to reduce the complexity of a multivariate system of observations by means of linear functions (discriminant functions) of original variables: they are estimated so that the divergency among groups (here varieties) will be maximized on the basis of the variability existing within groups. The coefficients (a) defining the canonical variates (y) are found by the maximization of

$$\Sigma_i^k N_i (a' \bar{x}_i - a' \bar{x})^2 \; (a' S \; a)^{-1}$$

where $\bar{x}_i$ is the mean vector of the i-th group (variety) and $\bar{x}$ is the overall mean vector. N_i is the number of observations of the i-th group and S is the variance-covariance matrix of errors. The solution is found by solving

$$(B - \lambda S) a = 0$$

where λ is an eigenvalue of $S^{-1} B$, and B is the between groups variance-covariance matrix:

$$B = \Sigma_i^k N_i (x_i - \bar{x}) (x_i - \bar{x})' / (k - 1)$$

If we achieve a dimensional reduction taking into account only the canonical variates related to the largest eigenvalues of $S^{-1} B$, the following results can be obtained:
(1) an elimination of most of the redundancy of the original multivariate system, where the traits are correlated to a different extent;
(2)a statistical tool to assign a new object (plant), whose origin is unknown, to one of the groups incluced in the analysis;
(3) a projection of the mean values (centroids) of the groups in the orthogonal space defined by canonical variates, with possible taxonomic derivations on the basis of a reduced but significant number of axes. As the last point is concerned, use of standardized discriminant functions is generally suggested:

$$z_i = a'_i (x - \bar{x})$$

An application of discriminant analysis to the classification of vine varieties

As matter of exemplification, the procedure was applied to describe and discriminate 31 vine varieties. Data were recorded at the Istituto Sperimentale per la Viticoltura in Susegana (Eastern Venetia) in 3 different years considered as repetitions with 3 stocks per variety within each year (Calo et al. 1989).

The traits used in the analysis are reported in Table 1. Numerous traits were recorded but only those showing in the 3 years a variability among groups (varieties) greater than the variability between plants within groups are considered here. Following this simple criteria, a minimal level, as regards the discriminant power, is assured. Most of the traits are expression of underlying quantitative continuous variables. Others describe qualitative characteristics.

Normality assumptions and the matrix of errors

Our data are expressed according to different rating scales. Since the canonical analysis is an application of, or better is based on, the multivariate analysis of variance (MANOVA), it is assumed that the variables used in the analysis are continuous and jointly follow a multivariate normal distribution. When the variables are discrete, as in the present case, the assumption of normality is generally not obvious.

We have undertaken the discriminant analysis on the basis of the following considerations:
(1) The generalization of the central limit theorem to the multivariate case says that if $X_1 \ldots X_p$ have variances $\sigma_1^2 \ldots \sigma_p^2$ and correlations ρ_{ij} ($i = 1 \ldots p, j = 1 + 1 \ldots p$), then the means $\bar{x}_1 \ldots \bar{x}_p$ of a sample of size N have a joint distribution that as N increases approaches to a multivariate normal distribution with variances $^1/_N \sigma_1^2 .. ^1/_N \sigma_p^2$ and with correlations ρ_{ij} the same as those of the X's. The robustness of most multivariate statistical tests is based on this theorem, provided that variances are independent from means (see Mariott 1974, 15). For this reason our sampling design includes replications in different years and within year.
(2) Even if the distribution of single variable is not normal, the distribution of a linear function of numerous variables is approximately normal and the normality increases with the number of variables entering the linear function (Seal 1964, 139). A 'caveat' is the fact that the coefficients applied to 1 or 2 of the non-normal variables allow to dominate the results (this point will be raised later).
(3) Incidental deviations from normality of single component variables will not cause distortion of the point estimates, which are of main interest in discriminant applications (for a discussion on the consequences of deviations from normality, see Scheffé 1959, chap. 10).

Table 1: Descriptor variables from O.I.V. protocol

Variable	Code	Description	Scale
003	I2	Young shoot: intensity of anthocyanin coloration of tip	F
004/005	I3	Young shoot: density of hairs of tip	F
007/008	I4	Shoot: color of internodes	A
009/010	I5	Shoot: color of nodes	A
012/014	I7	Shoot: density of hairs on internodes	F
068	I10	Mature leaf: number of lobes	C
068/B	I11	Mature leaf: angle between weins (L and L1)	
068/C	I12	Mature leaf: ratio L1/L	
075	I13	Mature leaf: blistering of the upper side	F
079	I16	Mature leaf: general shape of petiole sinus	E
081	I17	Mature leaf: particularieties of petiole sinus	A
084/085	I18	Mature leaf: density of hairs between the veins	F
090/091	I21	Mature leaf: density of hairs on petiole	F
093	I22	Mature leaf: lenght of petiole as compared to middle vein	F
202	I28	Bunch: size	F
206	I29	Bunch: length of peduncole	F
220	I31	Berry: size	F
225	I33	Berry: color of skin	D
236	I37	Berrry: particular flavor	B
238	I38	Berry: length of pedicel	F
301	I41	Time of bud burst	F

(scale values: A = 1 2 3, B = 1 2 3 4, C = 1 2 3 4 5,
D = 1 2 3 4 5 6 7, E = 1 2 3 4 5 6 7 8 , F = 1 3 5 7 9)

Choice of the matrix of error variances and covariances

The choice of an adequate model in MANOVA is also of importance for the arguments in the previous section. As the replication within years refers to single plants (stocks), we have considered as variance-covariance S matrix, the matrix from the effects of interaction varieties by years (see table below). In this way we achieve two important points. First, the unit of observation becomes the mean value over three stocks within each year assuring a better fit to normality than individual plants. Second, the weighting of the variability among varieties is performed on the basis of the joint performance over 3 years, which increases the discriminating power of more stable traits. This aspect is in agreement with the simple coefficient of discrimination of LUBISCHEW (1962) as regards single traits.

MANOVA model

Items	D.F.	Matrices
Between years	2	-
Between varieties	30	B
Int. year x varieties	60	S
Residual from the model	186	-

It is possible to asses the adequacy of this model with the MANOVA assumptions by the examination of the frequency distribution of the errors (variety x years effects) as shown in Fig. 1 for some traits included in the analysis. The fit to normal distribution is generally good and particularly with the traits that are more relevant to the discriminant process: this aspect is in agreement with the expectations of point (2), above.

The use of interaction effects in the contest of discriminant analysis is original and we think that it can meet the requirements of a multivariate analysis based on rating scores.

Table 2: Eigenvalues and percentages of variance explained

Root No.	Eigenvalue	Pct.	Cum. Pct.
1	210.27516	61.62073	61.62073
2	44.70333	13.10022	74.72096
3	27.76431	8.13628	82.85723
4	16.22094	4.75351	87.61075
5	12.59509	3.69097	91.30171
6	7.89587	2.31387	93.61558
7	4.74879	1.39162	95.00720
8	3.19846	.93730	95.94451
9	3.17576	.93065	96.87516
10	2.50119	.73297	97.60813
11	1.93070	.56579	98.17391
12	1.45800	.42726	98.60118
13	1.13728	.33328	98.93445

Evaluation of the discriminant power

The eigenvalues (λ_k) whose value is greater than 1 and the relative percent of variance explained are reported in Table 2. The first 7 eigenvalues account for 95 % of the total variation among varieties and we achieve a dimensional reduction by ignoring the smallest 14 eigenvalues. Canonical loadings (correlations between original traits and the canonical variates) allow a biological interpretation of the results of linear transformation. They are reported in Table 3.

The null hypothesis of no differences between varieties was rejected by a Hotelling-Lowley test following MANOVA, which, when approximated by an F statistics, received the value 20.09.

Mean values of varieties according to the first 7 canonical variates were used to derive taxonomic aspects. Since canonical variates are orthogonal, the simple square distances between

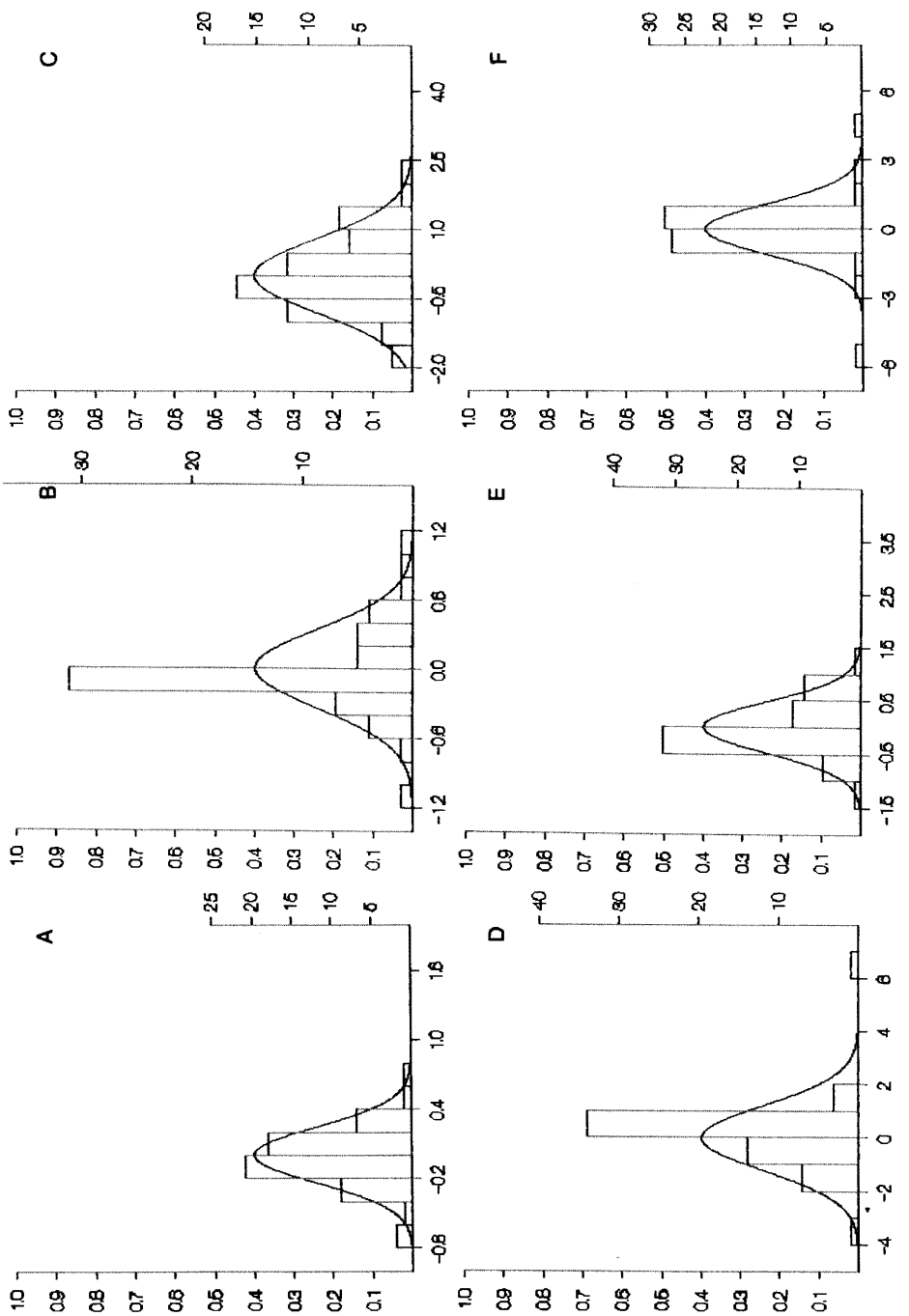

Fig. 1: Frequency distribution of errors (variety x years interaction effects) as regards some traits included in the analysis: A) shoot, color of internodes; B) shoot, color of nodes; C) shoot, density of hairs on internodes; D) mature leaf, density of hairs between the veins; E) berry size; F) color of berry skin. In the ordinates: count of cases (on the right); proportion of case per standard unit (on the left).

varieties correspond to Mahalanobis' generalized distances:

$$D^2_{ii'} = (\dot{z}_i - \dot{z}_{i'})' (\dot{z}_i - \dot{z}_{i'}) = (\dot{x}_i - \dot{x}_{i'})' S^{-1} (\dot{x}_i - \dot{x}_{i'})$$

The significance of a distance can be obtained referring to the critical values at different levels of α. These critical values can be obtained by an approximation of the T^2 statistic to F:

$$T^2 = [(\nu - p) / (\nu - p + 1)] F_{\alpha;\, p,\, \nu-p+1}$$

and noting that $T^2_{ii'} = [(N_i N_{i'}) / (N_i + N_{i'})] D^2_{ii'}$. ν are the degrees of freedom of S and p the number of variates used in computing the Distance (here p = 7).

The critical values of distances received the following values:

$$\alpha = 0.05 \;\; D^2 = 3.22, \;\; \alpha = 0.01 \;\; D^2 = 4.05$$

Table 3: Canonical loadings

	Canonical variates						
	1	2	3	4	5	6	7
I2	-.031	-.005	-.075	-.148	-.212	.028	.240
I3	-.034	-.202	-.046	.163	-.107	.035	.173
I4	-.037	-.070	*-.347*	-.210	-.007	-.265	*.466*
I5	-.020	-.030	*-.330*	-.225	-.010	-.257	.244
I7	-.017	-.238	-.013	.168	-.064	-.189	-.115
I10	-.038	-.158	.034	.137	-.255	*.394*	.135
I11	-.005	.001	.040	.082	-.101	-.220	-.036
I12	.003	.017	-.026	.044	-.016	-.052	.069
I13	-.052	-.003	-.062	.147	-.164	-.103	.188
I16	-.018	-.108	.170	.211	*-.528*	*-.388*	-.163
I17	.008	-.006	-.015	.078	-.055	-.121	-.049
I18	-.132	*-.495*	-.068	.188	.060	-.091	.086
I21	-.020	-.094	.052	.010	.012	-.264	-.062
I22	-.059	-.032	-.119	.083	.059	-.187	-.121
I28	-.006	-.265	.063	*-.342*	-.192	*.310*	*-.465*
I29	.011	-.085	-.065	-.099	-.084	.027	-.247
I31	-.020	-.027	-.006	-.246	-.170	.234	-.272
I33	*-.774*	.273	.187	-.078	.134	-.170	-.290
I37	-.002	.110	.027	.181	-.060	.060	.244
I38	-.013	-.034	.024	-.119	-.198	.059	-.209
I41	.033	-.095	*.614*	*-.433*	-.032	-.243	*.319*

A graphical representation of the discriminant results is the projection of varieties in the plot of the first 2 canonical variates. Even though two-dimensional diagrams may be somewhat misleading, these plots can reveal the divergencies among groups (clusters) of varieties. Varieties whose relative distances are not significant at the $\alpha = 0.01$ level were included in the same cluster, following a UPGMA (unweighted pair group with arithmetic mean) procedure (SNEATH and SOKAL 1973, 230-234). The plot shows some varieties not included in clusters and clusters with a reduced number of members (Fig. 2). This pattern can be regarded as a good result for a discriminant process.

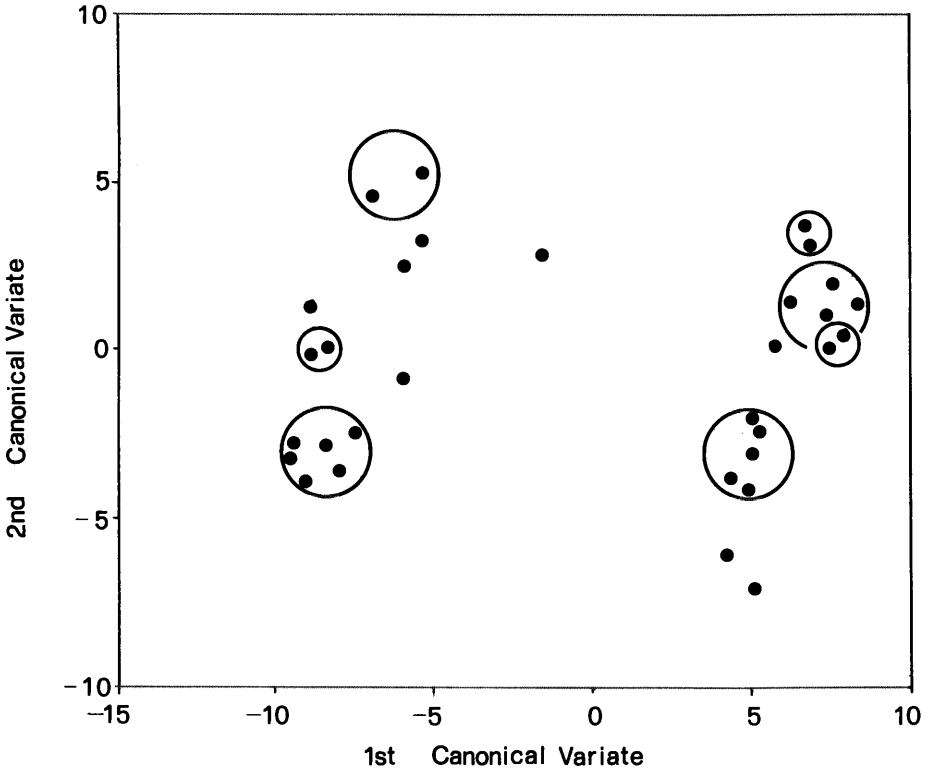

Fig. 2: Projection of variety centroids in the plot of the first 2 canonical variates. Circles include variates with relative distances not different from zero at α = 00.1 level of significance.

Regarding the discriminating power of single original descriptor variables, the matrix of loading reveals that the first canonical variate is mainly dominated by the color of the berry skin, while in general the large weight is assumed by traits linked to vegetative characteristics of the plant. Though these results are preliminary and presented merely as an example of a statistical procedure, the standardized discriminant coefficients of the first 3 canonical variates are reported in Table 4. The inspection of figures can be of great interest when choosing the variables with a higher discriminant power in the context of the examined covariation structure.

The use of interaction effects in the context of discriminant analysis is original and we think that it can meet the requirements of a multivariate analysis based on rating scores. More particularly, assuring a good fit to normality, it allows the use of usual statistical procedures, as stepwise selection and related statistical tests to choose the most discriminant descriptors.

Acknowledgements

C. A. is grateful to Prof. T. CALINSKI for useful discussions on the topic.

Table 4: Standardized discriminant function coefficients

	Function No.		
Variable	1	2	3
I2	-.19235	.36414	-.13183
I3	-.15294	-.13837	-.05140
I4	-.22991	-.13317	-.56496
I5	.00150	-.05174	-.45131
I7	-.05234	-.34306	.04656
I10	-.31959	-.19086	.02672
I11	.40305	.31301	.03755
I12	-.23082	-.30224	-.11383
I13	-.34818	.15428	-.10186
I16	.00050	-.09670	.18503
I17	.18351	.26616	-.15830
I18	-.47610	-1.29967	.04751
I21	.40265	.35096	.07206
I22	-.25843	-.39268	-.46230
I28	-.00298	-.28686	.08117
I29	.10030	-.19566	-.25988
I31	-.18423	-.29449	-.38828
I33	-1.12760	.23569	.11122
I37	-.04240	.26009	.16466
I38	-.06007	.23255	-.07179
I41	.00432	-.38915	.80115

Literature cited

CALO, A.; COSTACURTA, A.; GIUST, M.; OTTAVIANO, E.; CAMUSSI, A.; LORENZONI, C.; 1989: Preliminar contribution to the identification of ampelographic and ampelometric features for the characterization of vine varieties (*Vitis* sp.). Riv. Viticolt. Enol. **42**, 71-76.

LUBISCHEW, A. A.; 1962: On the use of discriminant functions in taxonomy. Biometrics **18**, 455-477.

MARIOTT, F. H. C.; 1974: The Interpretation of Multiple Observations. Academic Press, London, New York.

SCHEFFÉ, H.; 1959: The Analysis of Variance. J. Wiley and Sons, New York.

SEAL, H.; 1964: Multivariate Statistical Analysis for Biologists. Meuthen and Co. Ldt., London.

SNEATH, P. H. A.; SOKAL, R. R.; 1973: Numerical Taxonomy. The Principles and Practice of Numerical Classification. W. H. Freeman and Co., San Francisco.

SRIVASTAVA, M. S.; CARTER, E. M.; 1983: An Introduction to Applied Multivariate Statistics. North-Holland, New York.