HGCA

PROJECT REPORT No. 33

RAPID IDENTIFICATION OF BARLEY VARIETIES BY MACHINE VISION EXAMINATION OF SEED

MAY 1991

PRICE £10.00

HGCA PROJECT REPORT No. 33

Rapid identification of barley varieties by machine vision examination of seed

by

L. V. PURCHASE

Final report of a two year project which commenced in March 1989. The work was carried out by the National Institute of Agricultural Botany and was supported by a grant of £42,952 from the Home-Grown Cereals Authority (Project No. 0062/4/87).

CONTENTS

Abstract	1
Objectives	3
Introduction	4
Methods	7
Results	9
Discussion	16
Conclusions	21
Acknowledgements	24
References	25
Tables 1 - 34	26
Figures 1 - 6	60

ABSTRACT

A dedicated, prototype wheat grain image analyser, developed at the Official Seed Testing Station (OSTS) for England and Wales, was used to obtain binary images of seed samples taken from five varieties of two-rowed winter barley.

Operational difficulties associated particularly with sample presentation and image acquisition in barley preclude the use of this prototype for routine use. The curvature of both dorsal and ventral surfaces of the individual grains meant that each grain within a sample had point contact only with the sample presentation bar of the apparatus. Mechanical vibration caused by the movement of the camera caused individual grains within a sample to pivot about their longitudinal and lateral axes. This gave rise to non-uniform positioning of individual grains within a sample; some grains presented oblique lateral profiles, having pivoted about their lateral axes; movements about the longitudinal axis gave profiles in which one end of the grain was elevated relative to the other.

Despite the difficulties, it was possible to capture barley grain images which were then used to derive measurements of size and shape. Software incorporated in the prototype analyser gave sixty-nine quantitative measurements (*descriptors*) for each grain within a sample, based upon specific aspects of grain size and shape as viewed in lateral section. Sixty-nine descriptors were obtained for each grain within a sample; the software associated with the prototype grain image analyser calculated arithmetic means for each descriptor, taken from data on all the seeds within a sample for each variety. It was possible to calculate the corresponding median values for each descriptor from a particular variety sample as a separate procedure prior to analysis.

Depending upon whether a given descriptor measured some aspect of size or shape, the sixty-nine descriptors were divided between two subgroups. For each subgroup, canonical discriminant analysis was applied with the aim of characterising the five target varieties on the basis of compound measurements of size or shape.

Variety characterisation of multiple grain samples was demonstrated as being possible using measurements of either size or shape alone. Characterisation was possible using either arithmetic means or medians; generally, arithmetic means gave lower incidence of self-classification errors.

Using samples of varieties from one harvest year, it was possible to generate numerical "rules" (canonical discriminant functions) which could be used to classify samples of the same varieties taken from different harvest years; the great majority of samples were identified correctly by these "rules."

By calculating the probability of obtaining correct variety identification using one descriptor alone, it was possible to assess the relative abilities of isolated measurements of shape or size to characterise the five varieties. Some descriptors were obviously better than others in their ability to "separate" the

five varieties; this may be of significance in future extension of this technique into DUS work. Measurements of size or shape could then be ranked, or ordered, on the basis of decreasing probability of correct identification. These rankings were then used to identify the minimum numbers of measurements of size or shape from each descriptor subgroup which were required to achieve variety characterisations with rates of error comparable to those obtained when using the full complement of measurements within each subgroup.

The relevance of some of these descriptors to specific features of barley grains is discussed with a view to future development of specific descriptors for barley characterisation and/or classification in possible DUS applications.

OBJECTIVES

The overall aim of this project was to assess the value of an existing dedicated prototype wheat grain image analyser in the development of a specific image analysis technique for the differentiation of barley varieties.

Evaluation of the existing dedicated prototype indicated that, while it was by no means ideally suited for the acquisition of data in barley, it could, nevertheless, be used to demonstrate the potential of combined image analysis and statistical analysis techniques (*pattern recognition*) for variety identification in barley.

In the first fifteen months of the study, (1989-1990), the use of one particular pattern recognition technique, (canonical discriminant analysis), showed particular promise in its ability to separate two closely related varieties, *Halcyon* and *Maris Otter*, on the basis of composite measurements of size and shape.

For the remaining nine months of the study, (1990-1991) three aims were identified:

(*i*) demonstration that observed differences in average outline shape of five barley varieties were sufficient to allow variety characterisation within a shared discriminant space defined by canonical discriminant analysis;

(*ii*) demonstration that observed differences between the varieties were sufficiently stable from year to year to allow subsequent classification of "unknown" samples based on previously measured samples;

(*iii*) the preparation of a paper (or papers) which described the work and results, demonstrating the potential of image analysis for barley variety screening. It was hoped that, by generating a wider awareness of the technique, it may be possible to attract further funding needed to support the realisation of a usable technology.

INTRODUCTION

Preliminary investigation of the application of image analysis techniques to the problem of identification in barley varieties indicated that the methodology had considerable potential. (Purchase, 1990.)

Despite operational difficulties (associated with sample presentation and image capture), it was shown that a prototype wheat grain image analyser could be used to obtain measurements of size and shape in barley grains. While not an ideal system for the acquisition of these types of data from barley, parameters designed specifically for characterising shape in wheat grains were shown to be adequate for the demonstration of consistent and measurable differences in the average outline shape of grains from five barley varieties. The prototype facilitated the quantification of continuously variable morphological characters which formed the basis for the development of multivariate variety characterisations by the application of statistical analysis techniques (*pattern recognition*) to the collected data.

Since present UPOV regulations do not admit the use of multivariate variety characterisations in DUS work, the "usefulness" of this application of combined image analysis/pattern recognition techniques very much depends upon the ability to apply these characterisations in the subsequent classification of "unknown" variety samples. In previously reported work, successful use of numerical "rules" obtained from variety characterisation by two pattern recognition methods, (cluster analysis and canonical discriminant analysis), in the classification of a limited number of "unknown" samples has been demonstrated.

Characterising the varieties by canonical discriminant analysis may be thought of as a process in which the descriptor scores or measurements for each sample of the target varieties are transformed to give composite scores which locate the particular sample within an n-dimensional space. Ideally, this process is such that all samples of the same variety form unique, well-defined entities within the discriminant space, with each group of variety samples, or entity, isolated from the other groups of different variety samples within the same discriminant space. This is achieved by maximising the distances between entities and reducing the spread of variety samples about the average location of each entity within the discriminant space.

As a technique for characterising and subsequently classifying varieties such as *Halcyon* and *Maris Otter* (two varieties so closely related that even on the basis of biochemical techniques, such as polyacrylamide gel electrophoresis, they are indistinguishable; see table 1) canonical discriminant analysis indicated particular promise.

In canonical discriminant analysis, the aim is to select "compounds" (linear functions) of the original descriptor scores which maximise **B** (the between-varieties sum-of-squares and cross-products matrix) relative to **W** (the corresponding

within-varieties sum-of-squares and cross-products matrix). These "compounds", (termed canonical eigenvectors or canonical variates) are transformed axes which are chosen so that the first one of them lies in the direction of the greatest variability of the variety means. The second axis is inclined in the direction of the next greatest variability and is orthogonal to the first, subject to zero correlation between the two. (The axes are required to be mutually uncorrelated to increase the separation between varieties.) Hence, each axis may be considered as a composite of the original descriptor scores, with differing "contributions" from each of them.

The number of such transformed axes that may be extracted is determined as *min(g-1,n)*: this is owing to a constraint of the analysis method, which requires that the extracted vectors and corresponding eigenvalues be non-zero. (In this application, the maximum number of canonical eigenvectors and associated eigenvalues that may be extracted is four, since *g* (the number of varieties = 5) is generally less than *n*, the number of descriptors.) However, not all of the possible *g-1* axes may be required to describe the total variation within a particular set of descriptors, for reasons explained below.

Differences between the varieties within discriminant space prior to extraction of each successive eigenvector can be examined, by testing the significance, or otherwise, of differences between vectors containing the mean values for each descriptor for each group of variety means. Successive eigenvectors may be extracted until either all *g-1* non-zero eigenvectors have been obtained or a null hypothesis is accepted. Acceptance of any of the null hypotheses indicates that the differences between the variety means in the remaining space may be attributed to chance alone. Beyond this point, there is no gain in further extraction as it serves only to add the effects of random variation to the discriminant process which may "blur" the distinction between the varieties.

Each of the canonical eigenvectors extracted for each model has an associated eigenvalue, which may be thought of as representing the variation in the data for the corresponding eigenvector. Summation across the *g-1* eigenvalues for each model gives the total variability of the system; hence, the percentage of the total variability accounted for by individual eigenvalues may be calculated. Thinking of variability as being synonymous with "information" in this context, examination of these percentages shows how much information is lost by, for example, the elimination of one axis from a graphical representation of the data. Generally, the first two or three canonical variates account for most of the variability in the system. This is useful from the point of view of representing the structure of the data within the transformed discriminant space.

Following on from the success of initial investigation of the application of this technology to barley, the aim of this study is to demonstrate how the derivation of a set of "rules" by canonical discriminant analysis (*variety characterisation*) enables identification (*classification*) of "unknown" samples of barley

grains to be made. Further, the observed differences between the average outline shape of five barley varieties are sufficiently stable from year to year to allow the use of characterisations developed from one years' samples to be used for classification of "unknown" samples of the same varieties taken from different years.

METHODS

Materials

The five barley varieties used in this study, with their recommended uses and PAGE (polyacrylamide gel electrophoresis) groupings are given in table 1.

Bulk samples of each variety were taken from ten different seed lots amongst the reference stocks of harvested varieties held by the Seed Production Department (SPD) at the National Institute of Agricultural Botany for each of the harvest years 1988, 1989 and 1990.

Subsamples, containing fifty seeds each, were drawn from the bulk samples; these were used to form training and test series of data using seed lot subsamples in the combinations given in table 2. Each of the fifty seeds included in a subsample was subject to two selection criteria in that:

(i) any seed with obvious morphological abnormality or defect was excluded from the subsample

(ii) any seed longer than 11mm would be too wide for the camera field of view: this was a constraint imposed by the optical geometry of the grain image analyser. Consequently, it was necessary to trim the awns of some seeds before presentation as part of a sample.

Methods

Image analysis

A description of the dedicated, prototype wheat grain image analyser and its general operation has been reported previously. (Keefe and Draper, 1986; Purchase, 1990.)

As noted elsewhere, definition of certain measurements (descriptors) provided by the image analyser remains confidential, owing to commercial interests associated with the development of the prototype; the software associated with the grain image analyser will, however, produce predictable measurements of any convex, polygonal shape. Table 3 is a list of those descriptors which are in the public domain; for convenience, descriptors are referred to hereafter as v1 to v69 inclusive.

For 2 samples of each of the 10 lots available for each variety collected from 1988 harvested stocks, (sample series B_{88}) measurements of size and shape were made for each grain in four different orientations; differences between each resulting grain profile A,B,C and D are given in table 4.

For the remaining sample series given in table 2, measurements were made using one orientation of the seeds only, this being profile A.

Numerical methods

For all data collected, both arithmetic means (calculated by the analyser software) and median values (calculated *post hoc* in a separate exercise prior to pattern recognition) for each descriptor were available.

Each of the sixty-nine descriptors was assigned to one of two subgroups: if the descriptor was a specific measure of size (*e.g.*, area, height, length), it was allocated to the subset {SIZE}; conversely, if the measurement was size-invariant, (*e.g.*, aspect ratio, shape factor, Q) it was assigned to the subgroup {SHAPE}. Hence, two subgroups of descriptors were established; {SIZE}, comprising 34 measurements (v1 to v31 and v67 to v69, inclusive) and {SHAPE}, containing 35 measurements (v32 to v66, inclusive).

Given that the subgroups {SIZE} and {SHAPE} contain 34 and 35 variables respectively, it is clearly impossible to give a graphical presentation of the original data which includes all the variation simultaneously. Effectively, the variation in the data can be examined by using canonical (transformed) axes to reduce the system from 34-or 35- dimensions down to 3- dimensions, which can be represented graphically in a 2- dimensional medium.

Using the data from each subgroup in turn, canonical variates were calculated using the MGLH (multivariate general linear hypothesis) module of the statistical analysis package SYSTAT (Wilkinson, 1988) on a Hewlett Packard Vectra ES (IBM-AT compatible) PC under Microsoft DOS 3.2. Input data for each model were the arithmetic means or medians for each descriptor, calculated over the fifty seeds in each sample from the ten lots for each variety.

The adequacy of variety characterisation was assessed by the number of errors in self-classification amongst the samples within a training set when the original data were transformed from feature to discriminant space in canonical analysis. Samples taken from 1989 and 1990 harvests (samples series X_{89},Y_{89},X_{90} and Y_{90}; see table 2) were intended primarily as test series to demonstrate the "competence" of classifiers based on the characterisation of the varieties in the training series. In the course of the investigation, data from different years were used to produce mixed training series, as indicated in table 2, to determine whether or not variety characterisations was possible using variety samples from different harvest years.

Samples used as test series for any of the classifiers were "unknown" in the sense that they were not included in the sample series used to establish characterisation and subsequent classifiers. Test series were used to assess the performance of each classifier in terms of the number of incorrectly identified (*i.e.*, misclassified) samples as a percentage of the total number of samples in the test series.

RESULTS

Variety characterisation

Comparisons of characterisations obtained using all descriptors from each subgroup {SIZE} and {SHAPE} amongst profiles A,B,C and D.

Whether using arithmetic means or medians as input variables for canonical discriminant analysis, it was possible to characterise all five varieties in samples series B_{88} on the basis of measurements of size or shape in all four orientations of the seeds, A,B,C and D.

Generally, in terms of the %error figures for each of the four profiles A,B,C and D, the differences between characterisations based on either {SIZE} or {SHAPE} were slight, as were the differences between characterisations based on medians and arithmetic means. (Table 5.)

The values for three multivariate statistics, Λ,ν, and θ are given in table 6 as a quantitative description of the differences between the four profiles in terms of the spatial arrangement of the varieties in feature (*i.e.*, defined by the untransformed descriptor scores) space. The multivariate F-approximation to Wilks' Λ describes the variation of the descriptor scores within-varieties as a proportion of the total variation present in the feature space; as Λ decreases, so the value of the multivariate F-approximation increases. The multivariate F-approximation to the Pillai trace, ν, represents a similar measure of the variation between variety groups. θ represents the ratio of the between-varieties variation relative to the variation within samples of the same variety; hence, its value increases as the variation within-varieties decreases relative to that of the variation between-varieties (Wilkinson, 1988; Kendall, 1975).

In both {SIZE} and {SHAPE} descriptor subgroups, both profiles A and B have greater values for Λ,ν and θ, indicative of superior separation of the variety groups. (Table 6.) This was true whether dealing with data based on arithmetic means or medians. Hence, from table 6, it is apparent that groups of variety samples are best defined as separate entities on the basis of either size or shape using data from either profile A or B.

Canonical discriminant analysis is generally regarded as "robust", meaning that its operation is not adversely affected by failure of any input variable to meet the assumptions underlying the technique. While use of input variables which do not meet the assumption of normality and homogeneity of variance within the variety samples need not necessarily compromise characterisation of the varieties, the full predictive power of the technique, in terms of assignment of variety samples on a probabilistic basis, may be affected. Thus far, all descriptors within each of the two subgroups {SIZE} and {SHAPE} have been used for characterisation on the basis of both arithmetic means and medians.

Tables 7 and 8 list those descriptors from both subgroups {SIZE}

and {SHAPE} within each of the four profiles which fulfilled the dual criteria of (*i*) univariate normality of the descriptor data within each variety and (*ii*), for a given descriptor, homogeneity of variances across all five varieties when both arithmetic means (table 7) and medians (table 8) were used. Univariate normality was tested within each variety by a Kolmogorov-Smirnov one-sample test with Lilliefors probabilities; homogeneity of variance was tested across all five varieties for a given descriptor using Bartlett's test. (Sokal and Rohlf, 1981.) For both tests, rejection of the null hypothesis at $p < 0.001$ was taken to indicate non-compliance with the assumption being tested. While most of the descriptors from each subgroup "passed" both tests in profiles A and B, the numbers of descriptors remaining in the subgroups for profiles C and D were reduced considerably; *e.g.*, for profiles A and B, using arithmetic means the number of {SIZE} descriptors failing to meet criteria (*i*) and (*ii*) was 7 and 4 respectively; for {SHAPE} descriptors, it was 9 and 3 respectively. Corresponding figures for profiles C and D were 14 and 25 respectively for {SIZE} and 24 and 23 for {SHAPE}.

Repetition of the characterisation of the target varieties within discriminant space using only those descriptors within each subgroup {SIZE} and {SHAPE} which met criteria (*i*) and (*ii*) above gave the results shown in table 9. Comparing these error percentages with those given in table 5, for arithmetic means, data from profiles A and B still gives flawless characterisation on the basis of either size or shape, while data from profiles C and D show some deterioration of performance. Using medians, the restriction on "available" descriptors has an obvious effect in all four profiles, A,B,C and D using {SHAPE} descriptors and profiles A,C and D using {SIZE} descriptors.

Table 10 shows the new values obtained for Λ,ν and θ for the restricted descriptor subgroups, indicating the relative "distinctness" of the five target varieties within feature space, as explained above. The obvious effects of reducing the number of descriptors within each subgroup are demonstrated by the relative changes in values for the multivariate statistics. Comparing table 2 with table 10, these changes are particularly marked in the data for profiles C and D, where restriction (on the basis of non-normality and heterogeneity of variances) reduced the availability of descriptors for inclusion within each subgroup {SIZE} and {SHAPE} for both medians and arithmetic means. These differences can be seen in figures 1 and 2, in which the first three canonical axes for each profile using {SIZE} (figure 1) and {SHAPE} (figure 2) descriptors meeting the dual criteria have been plotted. The figures show the disjointed distribution of the samples in discriminant space, representing groups of variety samples. The distribution of the varieties is, however, more distinct in the plotted data from profiles A and B.

Investigation of the minimum number of descriptors required to obtain variety characterisation.

Univariate F-tests performed on data from profile A gave a means of assessing the discriminatory "power" of each descriptor. Considered in isolation within each subset, {SIZE} and {SHAPE},

the descriptors may be ranked by decreasing order of their univariate F-ratios, taken to indicate the "power" of each descriptor to differentiate between the varieties. Using the method described in Lubischew (1962) and Sneath and Sokal (1973), estimates of the probability of correct variety identification using single descriptors may be determined; these are given in tables 11 and 12. As univariate F-ratios (indicating the ratio of the data spread between-varieties relative to that within-varieties) decreases, so the value of *p*(correct identification using a given isolated descriptor) decreases.

These rankings were then used to "remodel" the canonical transformations to determine which serial combination of descriptors had the greatest effect in terms of characterising the target varieties within the discriminant space defined by the transformed canonical axes. The "effect" of any combination of descriptors from the two subgroups {SIZE} and {SHAPE} on the configuration of the variety samples within discriminant space may be monitored, indirectly, by the changes in values for the three multivariate statistics Λ, ν and θ mentioned above. The aim of this "trial and error" process is to minimise the variation within groups of samples of the same variety within feature space (seen as a reduction in the value of Λ) whilst maximising the variation between those groups (visible as an increase the value of ν). θ can be used to indicate the simultaneous effect on both within-variety and between-varieties variation.

The descriptors from each subgroup {SIZE} and {SHAPE} used in each model for canonical discriminant analysis are given in tables 13 and 14 for arithmetic means and tables 15 and 16 for medians. Groups of descriptors from each subgroup were added to canonical discriminant analysis in order of descending values of the univariate F-ratio and *p*(correct identification).

Tables 17 and 18 give the values of the three multivariate statistics (Λ, ν and θ) associated with each combination of descriptors. The changing values of these statistics describe the changes in location of groups of variety samples within feature space. The values show increasing group separation and decreasing spread within groups of samples from the same variety as more descriptors from each subgroup {SIZE} and {SHAPE} are submitted as input variables for canonical discriminant analysis.

The position of each group of variety samples relative to others in the same discriminant space defined by four canonical axes may be established. Using the average location, or centroid of each group of variety samples as a reference point, Mahalanobis distances, (representing the Euclidean distances between centroid pairs) can be calculated. (Sneath and Sokal, 1973.) This allows assessment of the effects of various combinations of input variables on the final configuration of variety groups following canonical discriminant analysis. These relative distances are summarised graphically in figures 3, 4, 5 and 6. Here, each graph plotted shows the changes in relative distances between one named variety and the other four target varieties sharing the discriminant space. Figures 3 and 4 refer to arithmetic mean and median data for combinations of descriptors from the subgroups

{SIZE}; figures 5 and 6 contain similar plots for combinations of descriptors from the subgroup {SHAPE}.

In all four figures, a general pattern in the plotted data is apparent; initially, as the model number (and hence, the number of descriptors combined in canonical discriminant analysis; see tables 13 to 16) increases, there is an initial rise in the relative distances between pairs of centroids within discriminant space. Thereafter, as more descriptors are added as input variables, further divergence between centroid pairs is not so marked; it appears that successive additions of more input variables has less effect upon the resulting configuration of the variety samples in discriminant space following analysis. This suggests that the descriptors which convey the most "useful" information about each variety may be contained within perhaps the first three or four models. Certain descriptors convey the most important information from the point of view of locating groups of variety samples within discriminant space; subsequent addition of data from other descriptors serves only to "fine-tune" group location, producing no visible re-location of variety groups relative to each other within the co-ordinate system of the canonical axes.

Figure 6, showing plots of distances between centroid pairs based on arithmetic mean data from the subgroup {SHAPE}, presents a slightly different pattern. The final combination of input variables appears to have had a marked effect on the divergence of group centroids, breaking the trend established by previous combinations and possibly implying the existence of synergism within this final combination of descriptors.

The data plotted in these figures demonstrates the close proximity of the three varieties *Halcyon*, *Pipkin* and *Maris Otter*. There is very little divergence between the centroids of these three varieties; this is particularly true of *Halcyon* and *Maris Otter*. In terms of the relative distances between centroid pairs, the closer together two varieties are, the more chance there is that there will be overlap in the data points forming a variety group around each centroid. Overlap may lead to "blurring" of the boundary between two such varieties and the canonical discriminant analysis may thus fail to distinguish all samples of two such varieties correctly.

Tables 19 and 20 show how the "performance" of canonical discriminant analysis improves, in terms of correct allocation of each sample to the appropriate group of varieties. The allocation "rule" involves comparison of the Mahalanobis' distances between the co-ordinates of each of the five variety centroids within discriminant space and the co-ordinates of a point representing the location of the variety sample. The sample point is allocated to the closest group of variety points in discriminant space. As more sample points are allocated to each group, locations of the group centroids are adjusted. The figures given in tables 19 and 20 show the total number of samples incorrectly allocated as a percentage of the total number of samples. Whether based on arithmetic means or medians, as more descriptors from each subgroup are added as input variables for canonical discriminant

analysis, %error declines.

For descriptors based on {SIZE}, as the number of input variables based on arithmetic means increases from 1 to 9 (table 13) the number of samples incorrectly assigned drops from 63 to 4 out of a total of 150 (table 21): thereafter, all errors are attributed to incorrect assignations of variety *Halcyon* (table 21). Using medians, comparable results are seen: %error drops from 75 to 5 out of 150 (table 22) as the number of input variables increases from 1 to 10, subsequent errors involving two varieties only, *Halcyon* and *Maris Otter* (table 22).

Using combinations of descriptors from the {SHAPE} subgroup, as the number of input variables increases from 1 to 14 (for arithmetic means; table 14) and 1 to 9 (for medians; table 16), the number of samples incorrectly assigned falls from 61 to 6 for arithmetic means and 61 to 4 for medians (tables 23 and 24 respectively). Again, subsequent errors involve *Halcyon* and *Maris Otter* only.

Sample classification

Use of variety characterisations based on all descriptors within each subgroup {SIZE} and {SHAPE}.

Quantitative assessment of the "usefulness" of both size and shape characterisations of the five varieties as the bases for classification of further independent "pure" but "unknown" samples was possible. (Here, "pure" is used to indicate that samples comprise grains of one variety alone; "unknown" is in the sense that such samples were not used in the development of each classifier.)

Canonical discriminant models developed from the 1988 sample series C_{88} for each descriptor subgroup were used to obtain the canonical scores of further independent samples taken from 1989 and 1990 harvested stocks. On the basis of these scores and a distance criterion (Mahalanobis' D^2), a given grain sample was assigned to the variety to which it was closest in the discriminant space defined by the canonical axes for each descriptor subgroup. (Sneath and Sokal, 1973; Dunn and Everitt, 1982.) Significantly, variety characterisations based on either shape or size data from the 1988 samples alone define classifiers which show generally adequate classification performance on samples taken from subsequent season's harvested stocks, 1989 and 1990.

In terms of total misclassification error, (the total number of variety samples incorrectly classified as a percentage of the total number of samples within a test series) the 1988 classifier developed from samples series C_{88}, using all 34 arithmetic mean-based {SIZE} descriptors 1.3% and 5.3% incorrect assignations in the 1989 and 1990 sample series X_{89} and X_{90} respectively. For all 35 {SHAPE} descriptors, the corresponding misclassification errors in X_{89} and X_{90} were 4.0% and 5.3%. (Table 25.)

For median-based classifiers, misclassification error in X_{89} and

X_{90} was 4.0 and 8.0% respectively using all {SIZE} descriptors; using all available {SHAPE} descriptors, misclassification errors were 12.0 and 9.3% for 1989 and 1990 test sample series respectively. (Table 26.)

Characterisation of the varieties on the basis of size was largely unaffected by adding five samples of each variety from the 1989 sample series to those in the 1988 sample series to form training set C_{88+89}; the same was true when five samples from the 1990 series were added to the 1988 series. In both cases, variety characterisation remained largely error-free, even though the data upon which it was based came from two different harvest years. (Tables 27, 28.)

Assessed by the "performance" of numerical classifiers, the use of combined data from two years as sample series for the development of classifiers for subsequent use on "unknown" samples had a variable effect.

Classifiers for {SIZE} and {SHAPE} descriptor subgroups based on arithmetic means and sample series C_{88+89} did not improve classification of the remaining 1989 samples in series Y_{89}. This contrasted with the apparent improvement in performance of classifiers based on a mixture of data from 1988 and 1990 (sample series C_{88+90}) on the remaining samples in sample series Y_{90}; here, for both {SIZE} and {SHAPE} descriptor subsets, the classifiers developed from two years' data reduced %misclassification error from 5.3 to 2.0%. However, using all available data for 1988 and 1989 to form one large sample series ($C_{88}+X_{89}$) did not improve the ability of resulting classifiers for either {SIZE} or {SHAPE} to identify 1990 samples in series X_{90}. (Table 25.)

Conversely, where classifiers were based on medians, there was an apparent "benefit" in terms of improved classifier performance: in sample series Y_{89} and Y_{90}, %misclassification error was approximately half that seen in sample series X_{89} and X_{90} for both subgroups of descriptors. Combining two years' samples (1988 and 1989) gave rise to classifiers which reduced the %misclassification error in the 1990 sample series X_{90}. (Table 26.)

Use of variety characterisations based on different combinations of descriptors from each subgroup {SIZE} and {SHAPE} in subsequent classification of unknown variety samples.

Tables 29 and 30 show how each of the classifiers developed for different serial combinations of arithmetic mean and median data from both subgroups {SIZE} and {SHAPE} "performs" when used to classify 15 unknown samples of each of the target varieties taken, this time, from 1989 and 1990 harvested stocks.

For both years' samples, the %error figures given in these tables show generally similar patterns to those given in tables 19 and 20. Whether using arithmetic mean data or medians, as the number of 1988 based descriptors used as input variables for canonical discriminant analysis increases, so the %error drops in the 1989

and 1990 samples. The tabulated data shows that the %error for both 1989 and 1990 samples was generally higher than that for 1988 samples; %error for 1990 samples was generally higher than that for 1989 samples. Both arithmetic mean and median based 1988 samples showed sudden marked reduction in %error as the number of descriptors from each subgroup used as input variables increased beyond 3 or 4 (see tables 19 and 20); similar marked decrease was not evident in %error from either 1989 or 1990 samples. (Tables 29 and 30.)

Using arithmetic mean data from descriptors in either subgroup {SIZE} or {SHAPE}, the main sources of classification error amongst both 1989 and 1990 samples were the three varieties *Halcyon*, *Maris Otter* and *Pipkin*. (Tables 31 and 32.)

Using medians, amongst 1990 samples, most error was associated with *Halcyon* and *Maris Otter*. For 1989 samples, again, the three varieties *Halcyon*, *Pipkin* and *Maris Otter* had the highest error associated with them, whether dealing with combinations of {SIZE} or {SHAPE} based descriptors. (Tables 33 and 34.)

DISCUSSION

The results presented above have demonstrated that a combination of image analysis and pattern recognition techniques holds considerable potential for the classification of barley varieties.

Even within the limited range of varieties chosen, consistent and measurable differences in specific aspects of grain morphology have been shown, regardless of the fact that these measurements had been designed specifically for the characterisation of shape in wheat grains.

Variety characterisations could be made using either arithmetic means or medians. Generally, the results indicated that higher %error was obtained when using medians, but in terms of both self-classification error in characterisation and misclassification error in identification of unknown variety samples, these differences were very slight. Where differences did exist, they corresponded to one or two misclassified samples, rather than differences amounting to orders of magnitude. Hence, for the purposes of further discussion of the results, unless specific references to results based on medians are made, comments apply to both arithmetic mean and median-based data.

Differences between the varieties were sufficiently consistent to allow characterisation on the basis of either size or shape, irrespective of orientation group. Using all descriptors in each subgroup {SIZE} and {SHAPE}, it was possible to obtain characterisations which were either totally error-free or which had low self-classification errors of the order of 1-4% amongst a total of 100 samples.

Restriction of canonical discriminant analysis to only those descriptors which met the criteria of normality and homogeneity of variance provided further evidence of the "unsuitability" of data from profile groups C and D. The practical difficulties associated with the gathering of data from these two orientations of the seeds have already been described (Purchase, 1990). Numerically, those descriptors in each subgroup which comply with the basic assumptions of the method of analysis may be inadequate for the purposes of giving a clearly disjointed distribution of the groups of variety samples within discriminant space.

Profiles A and B proved the most "useful" of the four seed orientations since most descriptors within each subgroup complied with assumptions of both normality and homogeneity of variance. Flawless variety characterisation was obtained in both profiles A and B using arithmetic mean data for each descriptor subgroup. Characterisations based on medians had higher incidence of self-classification errors; using descriptors based on size, characterisation was error-free in profile B only. Hence, when considering possible future applications of these results, if the full probabilistic potential of this method of numerical analysis is required, it may be preferable to use arithmetic means rather than medians. This may be particularly relevant in any "end-use"

which involves economic decision-making on the basis of such classifications.

Evaluating the potential of this technique in its application to characterisation of barley varieties, data from profile A provides the most "accessible" information in the sense that the definition of each descriptor may be related to approximately analogous features on individual barley grains. Combining two different views of the grains, i.e., data from profiles A and B, in the same canonical discriminant analysis may be worthy of future investigation, since it is likely that the analysis method will place different weights on measurements from each profile group.

It was possible to achieve near optimum variety characterisations using reduced numbers of input variables from either descriptor subgroup, {SIZE} or {SHAPE}. Using only one variable from each subgroup, a generally disjointed distribution of groups of variety samples could be achieved in which most of the self-classification errors could be attributed to incorrect assignments of the three smaller varieties, *Halcyon, Pipkin* and *Maris Otter*. Using arithmetic means, the descriptors with the highest individual probabilities of correct sample identification were v30 and v60 from subgroups {SIZE} and {SHAPE}; for medians, the corresponding descriptors were v14 (perimeter) and v60. In very general terms, v30 is related to the point of inflexion of the dorsal hull of the grain just beyond the lemma base and above the point of insertion of the lodicules on the ventral surface. V60 is a related but size-invariant descriptor which provides an assessment of the feature relative to grain length. Using arithmetic means, error-free characterisation of *Igri* was possible on the basis of either v30 or v60 taken singly.

Submission of an additional descriptors to canonical discriminant analysis from each subgroup brought about considerable reduction in the self-classification error percentages. Measurements of size, such as length (v2), perimeter (v14) and v30 were adequate to give flawless characterisation of varieties *Igri* and *Panda* using arithmetic means; using medians, one more descriptor, v6 (germ length) was required to achieve the same result. Perfect charactersation of these two varieties on the basis of shape required more "information"; using arithmetic means, v37 (relative germ length) was needed in addition to the combination v44,v53,v56,v57 and v60 which was adequate for separation of these two varieties on the basis of medians. (V44,v53,v56 and v57 are all descriptors which measure aspects of shape of the "embryo-end" of the grain as viewed by the camera.) Still more information was required to characterise the remaining three varieties Halcyon, Pipkin and Maris Otter. The "closeness" of these three varieties in the genetic sense is evident at the biochemical level by their PAGE groupings (see table 1) which reflects the degree of "relatedness" between them: *Halcyon* and *Pipkin* share a common parent in *Maris Otter*. Morphologically, differences between these three varieties are very slight; existing taxonomic methods for their identification, (relying on such features as rachilla hair length, strength of rachilla hairs, lodicule size, degree of nerve

pigmentation and aleurone colour) are based on relative differences between the varieties and require an experienced "eye" in order to use them effectively.

Since Wilks' Λ ranges from 0 to 1, the incidence of a null value for this statistic may be taken an indication of optimal dispersion of the variety samples within discriminant space. On this basis, minimum subsets of descriptors from each subgroup {SIZE} and {SHAPE} can be identified for which associated self-classification errors involve, at most, a few samples of particular varieties, as follows:

arithmetic means:
{SIZE}
v1,v2,v6,v7,v13,v14,v27,v30,v31,v68,v69
(Self-classification errors associated with 2 samples of *Halcyon* only.)

{SHAPE}
v32,v33,v35,v36,v37,v38,v39,v42,v44,v44,v45,v46,v48,v51,v52,v53,
v56,v57,v60,v65,v66
(Self-classification errors associated with 2 samples of *Halcyon* and 1 sample of *Maris Otter*).

medians:
{SIZE}
v1,v2,v6,v7,v8,v13,v14,v27,v30,v31,v68,v69
(Self-classification errors associated with 2 samples each of varieties *Halcyon* and *Maris Otter*.)

{SHAPE}
v32,v33,v34,v35,v36,v37,v38,v39,v42,v44,v45,v46,v47,v48,v49,v50,
v51,v52,v53,v54,v55,v56,v57,v58,v59,v60,v61,v62,v63,v64,v65,v66
(Self-classification errors associated with 3 samples of *Halcyon* and 1 sample of *Maris Otter*.)

When used for wheat, the descriptors of size and shape may be used to reconstruct the image of a particular grain; though certain specific features may be meaningless in terms of the morphology of barley grains, the descriptors do provide predictable measurements of convex polygons which approximate the size and shape of each grain within the variety samples. It is therefore possible to determine in which part of the barley grain the most important differences occur.

Certain "basic" overall measurements of gross morphology appear to be important. Within the {SIZE} and {SHAPE} descriptor subsets given above, measurements of area (v1), length (v2), and perimeter (v14) are "matched" by indications of overall circularity (shape factor, v32) and rectangularity (aspect ratio, v33). All five descriptors have relatively high values for p(correct identification), (see tables 11 and 12), implying that there are important differences between the five varieties in terms of overall size and shape alone.More specific descriptors, such as v30 and v60 have been discussed above, indicating the relative

"importance" of the shape of the "embryo-end" of the grain, as do descriptors such as v6 (germ length), v7 (germ angle), v8 and other measurements of shape in this region, such as v39 and v44.

Similarly, descriptors v27, v56 and v57 indicate the importance of size and shape in the grain's ventral hull just beyond the lemma base. Curvature of the ventral hull is largely the determinant of values of descriptors such as v13 (high point), v51, v52, v68 and v69.

Considering the possible future development of a range of barley specific measurements for use in conjunction with the existing analyser, the areas of the grain profile mentioned above seem to be worthy of further investigation, from the point of view of either adjusting the software definitions to "follow" the barley profile more specifically, or in the development of new descriptors.

Significantly, variety characterisations based on either full descriptor subgroups or reduced subsets drawn from these subgroups define classifiers which are generally adequate for the classification of samples taken from subsequent seasons' harvests. This suggests that the differences between varieties are relatively stable from year to year, sufficiently so that they might permit classifications to be made from a reference database comprising accumulated data from, say, previous years' variety samples. Generally, the results appear to suggest that the greater the "time-difference" between samples used in characterisation and those which are to be classified by "rules" developed from that characterisation, the poorer the subsequent classifier performance. Improvement in classifier performance achieved by "mixing" data from two years within characterisation may represent wider sampling of the total variation in the system represented by measurements from two years' variety samples.

There was some evidence to suggest that "updating" a training set to include samples from two different years had potential for improving classifier performance in terms of reduction on the incidence of misclassification error. However, investigation of the possible differences in the dispersion of the variety means between two years gave results which implied that variation was statistically significant. Multivariate tests of the significance of dispersion of the vectors containing means for each variety between pairs of years (1988 and 1989, 1988 and 1990 and 1989 and 1990) gave results which were very highly significant (p <0.001 in each comparison) within each descriptor subgroup {SIZE} and {SHAPE}. Interpretation of the practical implications of these results is difficult in view of the relative paucity of data from 1989 and 1990 compared with that available for 1988, though the tests used equal numbers of samples from each year. It is possible that they reflect differences in location and the interaction of climate with crop growth during the growing season.

Given the specificity of many of the descriptors, the results discussed above serve best as examples of the possible use of

multivariate variety characterisation and subsequent classification. The scope of this study was limited to the use of binary images only; this may not represent the optimum method of variety identification as it ignores those surface features currently used in existing taxonomic methods. Such features may serve to simplify the accuracy and speed of identification by acting both as "sieves" to divide the varieties into groups or as highly specific descriptors for the characterisation of one particular variety alone. This study has indicated particular regions of the grain (basal dorsal and ventral hulls) which may be worthy of further immediate research with a view to producing barley-specific descriptors for use with the existing prototype wheat grain image analyser. The results have shown that it is possible, for example, to obtain error-free characterisation of Igri on the basis of single descriptors, such as v30 or v60, which refer to specific features found in the outline shape of wheat grains. "Tailored" descriptors may demonstrate that it is possible to obtain similar, error-free characterisations for other varieties, thus opening up potential applications of this technique in the area of variety characterisation for DUS purposes.

CONCLUSIONS

The unsuitability of the apparatus is due to the curvature of both dorsal and ventral hulls of individual barley grains. Essentially, lacking a flattened surface, barley grains pivot about their point of contact with the sample presentation bar; the outcome of this is described as (*i*) lack of uniform "attitude" within a sample of grains and (*ii*) the occurrence of oblique, rather than lateral, grain profiles. Both (*i*) and (*ii*) introduce additional variation in the system; non-uniform attitude imposes bias on the arithmetic mean scores calculated for each descriptor; oblique profiles produce spurious measurements of both size and shape.

Though such operational difficulties preclude the routine use of the prototype analyser for obtaining data on shape and size in barley, measurements that were obtained indicated that variety characterisation and subsequent classification was achievable using two particular pattern recognition techniques.

Clustering could be used as a numerical "sieve" to divide the varieties into groups which could then either be subjected to canonical discriminant analysis or further clustering. For varieties such as *Igri* and *Panda*, where there are obvious differences in size and shape, clustering could be used to characterise each variety as a unique group of samples within feature space. Clustering alone was not an adequate method of characterising the three very similar varieties *Halcyon*, *Pipkin* and *Maris Otter*; in order to distinguish between these, canonical discriminant analysis was required.

Of the two pattern recognition techniques, canonical discriminant analysis appeared to be the most "useful"; it could be applied with equal success to all five varieties to give flawless variety characterisation on the basis if either size or shape. Even restricting the descriptors used to characterise the varieties to those which fulfilled the assumptions underlying the numerical analysis method, it was possible to obtain error-free variety characterisations. The numerical "rules" obtained from characterisation could be used to classify unknown variety samples taken from subsequent years with low incidence of misclassification error.

This project has been restricted to the use of sample means and medians to characterise varieties; the question of identifying varieties from single grains, though not examined, may be of significance in possible future applications of these techniques, since it seems that the most likely "end-use" of this technology would be in the detection of individual seed contaminants. The main problems readily identifiable in the extension of this work, in which characterisation is based on variety samples, to that of characterisations based on individual seeds may be stated as follows:

(*i*) testing the data for conformity to the assumptions behind canonical discriminate analysis.

While "large" numbers of grains from each variety may meet both

criteria of normality and homogeneity of variance, there may be problems with "small" numbers which may not be solved by simple data transformation. If the end-use is to be in the area of contaminant detection, then these two criteria must be met in order to realise the full, probabilistic power of canonical discriminant analysis. If they cannot be met, then it will be necessary to turn to other clustering algorithms in order to achieve characterisation.

(*ii*) dealing with the increase in variance introduced when using individual grain data within the models.

"Blurring" of the boundaries between groups of individual grains representing each variety may occur,such that, even if the underlying criteria for canonical discriminant analysis are met, the dividing "lines" between varieties may become dividing "zones". This may mean that individual grains lying within the boundary zones would have approximately equal probabilities of membership of two or more groups on the basis of Mahalanobis distances and thus could not be classified with any degree of certainty. Given an extremely large sample of individual grains, it may be possible to use other numerical methods to define confidence "shells" about the average location of each variety within discriminant space and to use the boundaries of different "shells" to meet the level of accuracy required in any classification procedure.

The problems of single-grain identification are at their most complex in situations in which classifiers are required for use in the sense of a traditional botanical key; by contrast, in quality control it may be sufficient to identify a certain number of individual grains in a sample as "not x" without needing to identify them further.

Careful development of models which are reliable classifiers for given varieties is fundamental to both types of application of this technology. This necessitates considerable sampling effort in the first instance to ensure that as complete a range as possible of the variance within a variety is included. In possible extensions of this work to the problems of single grain identification, descriptors are needed which are relatively constant within-varieties compared to differences between-varieties; ideally, descriptors which are diagnostic of particular varieties should be sought.

In the search for diagnostic features which can be exploited in this technology, two significant areas for future research can be identified: one is based upon continued use of outline shape descriptors and the other upon the use of surface features used in existing taxonomic methods.

Development of barley-specific descriptors, based upon results presented above which suggested that differences in the size and shape of the grain's ventral and dorsal basal hulls may be important, may prove beneficial in the provision of descriptors which will at least be able to characterise variety samples on the basis of a single feature; this has immediate potential in the

possible use of this technology in DUS applications.

Restricted to binary images only, the possible implementation of grey-level processing to bring surface features of the grain "on-line" in an identification scheme is, at this stage, speculative. Technology currently available on the market would permit rapid implementation of initial studies in this particular area; switching to grey-level processing may have greater potential for eventual commercial development and exploitation in that it is likely to lead to the establishment of variety characterisations based upon single, highly specific characters. These may yield classifications which use "presence/absence" of particular features (*e.g.*, blue aleurone present in *Halcyon*, but not in either *Pipkin* or *Maris Otter*; nerve pigment in *Maris Otter* characteristically in three "neat" lines.)

In summary, therefore, despite the general unsuitability of the prototype for data acquisition in barley, the results obtained in the course of this study have demonstrated the obvious potential of image analysis methods for the characterisation and classification of barley varieties. Although the results obtained are generally encouraging,in that they demonstrate the clear potential of this technique, areas requiring significant future research effort have been indicated as a pre-requisite for further development; it seems very likely that there will be a considerable amount of further work which must be done in order to produce a system which would be both usable in industry and viable commercially.

ACKNOWLEDGEMENTS

Thanks are due to Twyford Seeds Ltd. for allowing use of the prototype wheat grain image analyser throughout this study.

I would like to thank Dr.P.D. Keefe for his technical advice and encouragement throughout this project.

REFERENCES

Dunn, G. and Everitt, B.S., (1982). An introduction to mathematical taxonomy. *Cambridge University Press.*

Keefe, P.D. and Draper, S.R., (1986). The measurement of new characters for cultivar identification in wheat using machine vision. *Seed Science and Technology, 14, 715-724.*

Kendall, M., (1975). Multivariate Analysis. *Charles Griffin and Company Ltd., London.*

Lubischew, A.A., (1962). On the use of discriminant functions in taxonomy. *Biometrics, 21, 491-505.*

Purchase, L.V., (1990). Rapid identification of barley varieties by machine vision examination of seed. *First interim report for two year project commencing March 1989, reference number 0062/2/87, dated April 1990: Home Grown Cereals Authority, London.*

Sneath, P.H.A. and Sokal, R.R., (1973). Numerical Taxonomy. *W.H.Freeman and Co.,San Francisco.*

Sokal, P.H.A. and Rohlf, F.J., (1981). Biometry. (2nd.edition) *W.H.Freeman and Co.,San Francisco.*

Wilkinson, L. (1988) . SYSTAT: The system for statistics. *Evanston, IL: SYSTAT Inc., 1988.*

variety:	code:[1]	abbreviation:[2]	uses:[3]	PAGE group:[4]
Halcyon	hcn	h	S/IOB	10.10
Igri	igr	i	G/wf	2.7
Pipkin	kin	k	S/IOB	10.13
Maris Otter	mot	m	S/IOB	10.10
Panda	pda	p	G/wf	10.3

notes:

[1]codes used in text to identify varieties.

[2]abbreviated codes used as labels in discriminant analysis.

[3]uses: S variety fully recommended, special use
G variety fully recommended, general use
wf winter feed variety
IOB accepted by the Institute of Brewers as a malting variety.

[4]varieties on the UK National List,January 1989, classified according to hordein electropherograms obtained by the use of the standard International Seed Testing Association (ISTA) PAGE (polyacrylamide gel electrophoresis) method.

table 1 :The five target varieties used in the study, indicating recommended uses and PAGE groups.

Identifier used in text:	n^2	N^1	Year	no samples taken per seed lot 1-10 for each variety: 1	2	3	4	5	6	7	8	9	10
B_{88}	20	100	1988	2	2	2	2	2	2	2	2	2	2
C_{88}	30	150	1988	3	3	3	3	3	3	3	3	3	3
X_{89}	15	75	1989	1	2	1	2	1	2	1	2	1	2
Y_{89}	10	50	1989	-	1	-	1	-	2	1	2	1	2
X_{90}	15	75	1990	1	2	1	2	1	2	1	2	1	2
Y_{90}	10	50	1990	-	1	-	1	-	2	1	2	1	2
combined samples:													
C_{88+89}	35	175	1988	3	3	3	3	3	3	3	3	3	3
			1989	1	1	1	1	1	-	-	-	-	-
C_{88+90}	35	175	1988	3	3	3	3	3	3	3	3	3	3
			1990	1	1	1	1	1	-	-	-	-	-
$C_{88}+ X_{89}$	45	225	1988	3	3	3	3	3	3	3	3	3	3
			1989	1	2	1	2	1	2	1	2	1	2

[1] N denotes total number of cases in sample series.

[2] n denotes number of cases per variety within that series.

combined samples denotes series drawn from two years' samples

table 2 :Details of the seed-lot composition of 1988, 1989 and 1990 training and test sets for the five target varieties *Halcyon, Igri, Pipkin, Maris Otter* and *Panda*, showing abbreviations used throughout the text.

identifier	description	notes[1]
v1	area	
v2	length	grain length (mm)
v3	height	grain height (mm)
v4	brush height	height of brush above stage
v5	germ height	height of the bottom of the scutellum above the stage
v6	germ length	scutellum length
v7	germ angle	tangent of an angle subtended to the horizontal by a line drawn through the germ region of the grain
v10	foot length	length of contacting surface between grain and stage
v13	high point	horizontal distance from the front of the grain to the highest point on the grain
v14	perimeter	
v15	dorsal tangent	tangent of an angle subtended to the horizontal by a line drawn along the dorsal area of the grain
v32	shape factor	$(area*4*\pi)/(perimeter)^2$
v33	aspect ratio	(height/length)
v34	Q	(area/(height*length))
v35	relative brush ht	v4/v3
v36	relative germ ht	v5/v3
v37	relative germ len	v6/v2
v38	horizontal axis	v36/v35

[1] references are to structures specific to wheat.

table 3 :Description of public domain measurements referred to in text (Keefe and Draper, 1986).

ventral furrow	lemma base / rachilla	awn base /palea base	PROFILE CODE
down	left	right	A
down	right	left	B
up	left	right	C
up	right	left	D

"down", "up", "left", "right" refer to relative positions of specified features as viewed on camera monitor screen.

table 4 :Explanation of difference between profiles A,B,C and D in terms of relative positions of grain features.

Profile	MEDIANS n=20[1] SIZE %	n=20[1] SHAPE %	ARITHMETIC MEANS n=20[1] SIZE %	n=20[1] SHAPE %
A	0	1	0	0
B	0	2	0	0
C	1	4	1	1
D	2	1	1	0

[1] n, number of samples for each variety = 20

table 5 :Assessing performance of the two classifiers, {SIZE} and {SHAPE} by the number of samples incorrectly assigned to each variety expressed as a percentage of the total number of samples used. Classifiers {SIZE} and {SHAPE} used all descriptors, 1-31+ 67-69 inclusive and 32-66 inclusive respectively.

Classifier	Profile	ARITHMETIC MEANS Wilks' $\Lambda^{1,2}$	Pillai trace $\nu^{1,2}$	Theta θ^{1}	MEDIANS Wilks' $\Lambda^{1,2}$	Pillai trace $\nu^{1,2}$	Theta θ^{1}
SIZE	A	26.047	12.982	0.989	19.124	11.979	0.981
	B	21.761	11.917	0.989	14.005	7.692	0.980
	C	16.755	9.348	0.986	9.864	6.250	0.969
	D	13.819	8.825	0.977	12.697	8.016	0.970
SHAPE	A	23.306	11.919	0.985	13.998	8.645	0.974
	B	19.904	11.245	0.988	12.300	7.029	0.973
	C	15.514	9.211	0.981	8.467	4.964	0.970
	D	12.056	8.145	0.968	10.461	7.811	0.957

[1] [2] all with corresponding probabilities of $p < 0.001$

F-approximations to the multivariate statistics shown.

table 6 :Arithmetic means and medians: values of multivariate statistics Λ, ν and θ "describing" the arrangement of groups of samples of each variety within the feature space defined by 34 and 35 descriptors within the two subgroups {SIZE} and {SHAPE} respectively.

Profile	ARITHMETIC MEANS {SIZE}	{SHAPE}
A	1,2,3,4,5,7,8,9,10,11,12,13, 14,15,16,17,18,19,20,21,22, 23,25,28,29,67,68	32,33,34,35,36,38,39,40,41, 42,43,44,45,47,50,52,53,55, 56,58,59,62,63,64,65
B	1,2,3,4,5,6,7,8,10,11,12,13, 14,15,16,17,18,19,20,21,22, 23,25,26,27,28,31,67,68,69	32,33,34,35,36,37,38,39,40, 41,43,44,45,46,47,48,49,51, 52,53,54,55,56,57,58,61,62, 63,64,65,66
C	1,2,3,4,5,10,11,12,13,15,17, 18,19,24,25,27,28,30,68,69	32,33,34,36,41,43,47,50,55, 58
D	1,2,3,10,14,18,25,28,68	32,33,34,41,43,47,48,55,58, 61,64

table 7 :Arithmetic means: descriptors within each subgroup which fulfil criteria of normality within varieties (tested by Kolmogorov Smirnov test, n =20 for each variety) and homogeneity of variances across all five varieties (Bartlett's test, n=20 for each variety) using arithmetic means as input variables.

Profile	MEDIANS {SIZE}	{SHAPE}
A	1,2,3,4,5,7,8,9,10,11,12,13, 14,15,16,17,18,19,20,21,22, 23,25,28,29,67,68,69	32,33,34,35,36,37,38,39,41, 42,43,44,45,46,49,50,51,52, 53,55,58,59,61,62,65
B	1,2,3,4,5,6,7,8,10,11,12,13, 14,15,16,17,18,19,20,21,22, 23,25,26,67,68,69	32,33,34,35,36,37,38,42,43, 44,46,47,48,49,50,51,52,53, 55,56,57,58,59,61,62,63,64, 65,66
C	3,8,10,11,12,13,15,17,18,19, 21,22,25,27,28,67,68,69	33,34,35,36,41,42,43,47,48, 50,57,58,60
D	1,2,3,6,8,10,13,18,21,25,26, 27,68	32,33,34,37,41,43,44,48,55, 56,60

table 8 :Medians: descriptors within each subgroup which fulfil criteria of normality within varieties (tested by Kolmogorov Smirnov test, n =20 for each variety) and homogeneity of variances across all five varieties (Bartlett's test, n=20 for each variety) using medians as input variables.

Profile	MEDIANS n=20[1] SIZE %	MEDIANS n=20[1] SHAPE %	ARITHMETIC MEANS n=20[1] SIZE %	ARITHMETIC MEANS n=20[1] SHAPE %
A	3	5	0	0
B	0	9	0	0
C	5	1	1	8
D	14	2	14	6

[1] n, number of samples for each variety = 20.

table 9 :Medians and arithmetic means: comparing the performance of the two classifier models {SIZE} and {SHAPE} when based upon input variables meeting criteria of univariate normality and homogeneity of variances.

		ARITHMETIC MEANS			MEDIANS		
Classifier	Profile	Wilks' $\Lambda^{1,2}$	Pillai trace $\nu^{1,2}$	Theta θ^{1}	Wilks' $\Lambda^{1,2}$	Pillai trace $\nu^{1,2}$	Theta θ^{1}
SIZE	A	23.280	12.697	0.983	20.478	11.804	0.975
	B	20.497	11.297	0.989	16.042	9.109	0.976
	C	27.718	13.305	0.980	15.635	2.674	0.961
	D	28.049	15.645	0.948	16.134	2.289	0.940
SHAPE	A	17.106	11.256	0.962	14.892	10.035	0.954
	B	17.134	9.893	0.981	13.112	7.894	0.968
	C	21.934	12.516	0.935	13.925	9.043	0.911
	D	17.436	11.655	0.881	18.872	13.972	0.907

[1] [2] all with corresponding probabilities of $p < 0.001$

F-approximations to the multivariate statistics shown.

table 10 :Arithmetic means and medians: values of multivariate statistics Λ, ν and θ "describing" the arrangement of groups of samples of each variety within the feature space defined by descriptors within the two subgroups {SIZE} and {SHAPE} respectively which meet the two criteria of (*i*) univariate normality within varieties and (*ii*) homogeneity of variance between five varieties.

ARITHMETIC MEANS			MEDIANS		
descriptor	F-ratio[1]	p^2	descriptor	F-ratio[1]	p^2
30	424.948	0.883	14	342.031	0.857
14	317.516	0.848	2	290.945	0.838
2	307.343	0.844	6	269.547	0.828
27	282.123	0.834	30	214.263	0.801
6	248.371	0.819	69	199.965	0.793
69	225.775	0.807	1	174.115	0.777
1	193.922	0.789	7	135.679	0.749
31	150.274	0.760	27	115.496	0.733
7	104.674	0.723	13	102.598	0.721
13	92.015	0.710	31	90.379	0.708
68	91.554	0.710	68	74.005	0.690
8	89.454	0.707	22	67.705	0.683
22	70.615	0.686	23	65.624	0.680
11	67.568	0.682	8	65.161	0.679
28	64.486	0.679	16	62.615	0.676
23	61.395	0.675	11	59.596	0.672
67	57.318	0.669	28	54.478	0.665
5	53.931	0.664	19	51.779	0.661
4	53.129	0.663	5	50.216	0.659
29	53.021	0.663	4	49.907	0.658
19	52.284	0.662	67	40.234	0.643
16	45.242	0.651	3	39.272	0.641
3	43.187	0.648	17	33.025	0.630
17	21.573	0.606	29	30.561	0.625
15	20.876	0.604	15	22.250	0.607
20	18.620	0.598	24	16.679	0.593
25	17.613	0.593	20	15.856	0.591
21	14.236	0.586	25	14.983	0.588
12	10.452	0.574	21	11.958	0.579
24	6.285	0.558	12	8.576	0.567
9	4.798	0.550	10	5.389	0.533
10	4.737	0.550	18	4.891	0.551
26	4.239	0.548	24	3.440	0.543
18	2.334	0.535	9	3.247	0.541

[1] Univariate F-ratios for each descriptor (d.f.=4,145)

[2] probabilities of correct identification using each descriptor singly

table 11 :Arithmetic means and medians: descriptors from the subgroup {SIZE} ordered by decreasing value of the univariate F-ratio based on 30 samples from each of the five target varieties.

ARITHMETIC MEANS			MEDIANS		
descriptor	F-ratio[1]	p^2	descriptor	F-ratio[1]	p^2
60	419.602	0.882	60	209.886	0.799
44	106.304	0.724	44	142.732	0.755
53	87.935	0.706	53	95.432	0.714
56	84.156	0.702	56	85.474	0.703
57	75.592	0.692	57	80.010	0.697
37	71.090	0.687	37	75.206	0.692
45	67.980	0.683	32	64.687	0.679
52	67.800	0.683	52	62.855	0.676
32	67.795	0.683	51	60.400	0.673
65	63.434	0.677	45	59.550	0.672
36	60.022	0.673	36	53.446	0.664
66	57.953	0.670	33	52.196	0.662
46	56.319	0.668	65	50.052	0.659
33	56.136	0.667	49	45.471	0.651
51	48.230	0.656	66	45.125	0.651
38	46.271	0.653	42	41.703	0.645
42	39.565	0.642	38	40.153	0.643
35	39.540	0.642	50	32.775	0.630
48	33.729	0.631	46	30.132	0.624
39	31.922	0.628	3	26.354	0.617
62	28.749	0.622	62	26.266	0.616
61	25.309	0.614	39	21.607	0.606
59	23.371	0.610	63	18.029	0.597
34	21.187	0.605	34	17.859	0.596
55	16.193	0.592	48	16.886	0.594
63	14.813	0.588	61	16.865	0.594
47	14.371	0.587	54	15.680	0.590
58	14.199	0.586	47	14.209	0.586
50	11.256	0.578	64	13.857	0.585
41	10.784	0.575	59	13.771	0.585
49	8.462	0.567	58	12.745	0.582
43	5.377	0.553	55	10.885	0.576
40	5.184	0.552	41	7.223	0.562
54	4.411	0.548	43	4.314	0.548
64	3.106	0.541	40	2.940	0.539

[1] Univariate F-ratios for each descriptor (d.f.=4,145)

[2] probabilities of correct identification using each descriptor singly

table 12 :Arithmetic means and medians: descriptors from the subgroup {SHAPE} ordered by decreasing value of the univariate F-ratio based on 30 samples from each of the five target varieties.

model number	descriptors included
1	v30
2	v2,v14,v30
3	v2,v6,v14,v27,v30
4	v1,v2,v6,v7,v14,v27,v30,v31,v69
5	v1,v2,v6,v7,v13,v14,v27,v30,v31,v68,v69
6	v1,v2,v6,v7,v8,v13,v14,v27,v30,v31,v68,v69
7	v1,v2,v6,v7,v8,v13,v14,v22,v27,v30,v31,v68,v69
8	v1,v2,v6,v7,v8,v11,v13,v14,v22,v23,v27,v28,v30, v31,v68,v69
9	v1,v2,v4,v5,v6,v7,v8,v11,v13,v14,v19,v22,v23,v27, v28,v29,v30,v31,v67,v68,v69
10	v1,v2,v3,v4,v5,v6,v7,v8,v11,v13,v14,v16,v19,v22,v23, v27,v28,v29,v30,v31,v67,v68,v69
11	v1,v2,v3,v4,v5,v6,v7,v8,v11,v13,v14,v15,v16,v17,v19, v22,v23,v27,v28,v29,v30,v31,v67,v68,v69
12	v1,v2,v3,v4,v5,v6,v7,v8,v12,v11,v13,v14,v15,v16,v17, v19,v20,v21,v22,v23,v25,v27,v28,v29,v30,v31,v67,v68, v69

table 13 :Arithmetic means: descriptors from the subgroup {SIZE} used as input variables for canonical discriminant analysis in each of the models indicated.

model number	descriptors included
1	v60
2	v44,v60
3	v44,v53,v56,v60
4	v37,v44,v53,v56,v57,v60
5	v32,v36,v37,v44,v45,v53,v56,v57,v60,v65
6	v32,v33,v36,v37,v44,v45,v46,v52,v53,v56,v57, v60,v65,v66
7	v32,v33,v36,v37,v38,v44,v45,v46,v51,v52,v53,v56, v57,v60,v65,v66
8	v32,v33,v35,v36,v37,v38,v39,v42,v44,v45,v46,v48, v51,v52,v53,v56,v57,v60,v65,v66
9	v32,v33,v34,v35,v36,v37,v38,v39,v42,v44,v45,v46, v48,v51,v52,v53,v56,v57,v59,v60,v61,v62,v65,v66
10	v32,v33,v34,v35,v36,v37,v38,v39,v41,v42,v44,v45, v46,v47,v48,v50,v51,v52,v53,v56,v57,v58,v59,v60, v61,v62,v65,v66

table 14 :Arithmetic means: descriptors from the subgroup {SHAPE} used as input variables for canonical discriminant analysis in each of the models indicated.

model number	descriptors included
1	v14
2	v2,v6,v14,v30
3	v1,v2,v6,v7,v13,v14,v27,v30,v69
4	v1,v2,v6,v7,v13,v14,v27,v30,v31,v69
5	v1,v2,v6,v7,v8,v13,v14,v27,v30,v31,v69
6	v1,v2,v6,v7,v8,v13,v14,v27,v30,v31,v68,v69
7	v1,v2,v6,v7,v8,v13,v14,v16,v22,v23,v27,v30,v31, v68,v69
8	v1,v2,v5,v6,v7,v8,v11,v13,v14,v16,v19,v22,v23,v27, v28,v30,v31,v68,v69
9	v1,v2,v4,v5,v6,v7,v8,v11,v13,v14,v16,v19,v22,v23, v27,v28,v30,v31,v67,v68,v69
10	v1,v2,v3,v4,v5,v6,v7,v8,v11,v13,v14,v16,v17,v19,v22, v23,v27,v28,v29,v30,v31,v67,v68,v69
11	v1,v2,v3,v4,v5,v6,v7,v8,v11,v13,v14,v15,v16,v17,v19, v22,v23,v27,v28,v29,v30,v31,v67,v68,v69
12	v1,v2,v3,v4,v5,v6,v7,v8,v11,v13,v14,v15,v16,v17,v19, v20,v21,v22,v23,v24,v25,v27,v28,v29,v30,v31,v67,v68, v69

table 15 :Medians: descriptors from the subgroup {SIZE} used as input variables for canonical discriminant analysis in each of the models indicated.

model number	descriptors included
1	v60
2	v44,v60
3	v44,v53,v60
4	v44,v53,v56,v57,v60
5	v37,v44,v53,v56,v57,v60
6	v32,v37,v44,v51,v52,v53,v56,v57,v60
7	v32,v33,v36,v37,v44,v45,v51,v52,v53,v56, v57,v60,v65,
8	v32,v33,v36,v37,v38,v42,v44,v45,v49,v51, v52,v53,v56,v57,v60,v65,v66
9	v32,v33,v36,v37,v38,v42,v44,v45,v46,v49, v50,v51,v52,v53,v56,v57,v60,v65,v66
10	v32,v33,v34,v35,v36,v37,v38,v39,v42,v44,v45, v46,v49,v50,v51,v52,v53,v56,v57,v60,v62,v65, v66
11	v32,v33,v34,v35,v36,v37,v38,v39,v42,v44,v45, v46,v47,v48,v49,v50,v51,v52,v53,v54,v55,v56, v57,v58,v59,v60,v61,v62,v63,v64,v65,v66

table 16 :Medians: descriptors from the subgroup {SHAPE} used as input variables for canonical discriminant analysis in each of the models indicated.

{SIZE} model reference number	ARITHMETIC MEANS Wilks' Λ	Pillai trace ν	Theta θ	{SIZE} model reference number	MEDIANS Wilks' Λ	Pillai trace ν	Theta θ
1*	-	-	-	1*	-	-	-
2	0.011	1.906	0.933	2	0.008	2.163	0.937
3	0.003	2.393	0.963	3	0.002	2.691	0.968
4	0.001	2.915	0.967	4	0.001	2.934	0.969
5	0.000	3.040	0.970	5	0.001	3.018	0.972
6	0.000	3.117	0.972	6	0.000	3.079	0.973
7	0.000	3.148	0.973	7	0.000	3.143	0.978
8	0.000	3.223	0.978	8	0.000	3.214	0.979
9	0.000	3.325	0.980	9	0.000	3.243	0.980
10	0.000	3.331	0.981	10	0.000	3.295	0.980
11	0.000	3.367	0.982	11	0.000	3.324	0.980
12	0.000	3.419	0.983	12	0.000	3.382	0.981

1* model 1, only a single descriptor used, hence no calculation of multivariate statistics.

table 17 :Multivariate test statistics for models 1-12, using serial combinations of descriptors from the descriptor subgroup {SIZE}. Values shown for both arithmetic mean and medians.

{SHAPE} model reference number	ARITHMETIC MEANS Wilks' Λ	Pillai trace ν	Theta θ	{SHAPE} model reference number	MEDIANS Wilks' Λ	Pillai trace ν	Theta θ
1*	-	-	-	1*	-	-	-
2	0.027	1.552	0.930	2	0.038	1.561	0.879
3	0.014	1.847	0.941	3	0.025	1.795	0.901
4	0.006	2.340	0.953	4	0.013	1.960	0.935
5	0.002	2.689	0.956	5	0.006	2.357	0.946
6	0.001	2.877	0.958	6	0.003	2.675	0.948
7	0.001	3.025	0.958	7	0.002	2.849	0.954
8	0.000	3.159	0.968	8	0.001	2.946	0.959
9	0.000	3.240	0.971	9	0.001	2.957	0.961
10	0.000	3.339	0.983	10	0.001	3.050	0.963
-	-	-	-	11	0.000	3.146	0.966

1* model 1, only a single descriptor used, hence no calculation of multivariate statistics.

table 18 :Multivariate test statistics for models 1-12, using serial combinations of descriptors from the descriptor subgroup {SHAPE}. Values shown for both arithmetic mean and medians.

model reference number	ARITHMETIC MEANS %error (n=30)[1]	MEDIANS %error (n=30)[1]
1	42.0	50.0
2	20.7	18.0
3	12.0	6.7
4	2.7	3.3
5	1.3	4.0
6	1.3	2.7
7	0.7	1.3
8	0.7	0.7
9	0.7	1.3
10	0.7	0.0
11	0.0	0.0
12	0.0	0.7

[1] 30 samples of each variety.

table 19 :Arithmetic means and medians: comparing the performance of serial combinations of descriptors from the subgroup {SIZE} in terms of their ability to self-classify samples of each variety correctly. Model reference numbers refer to those combinations of descriptors given in tables 13 and 15 for arithmetic means and medians respectively.

model reference number	ARITHMETIC MEANS %error (n=30)[1]	MEDIANS %error (n=30)[1]
1	40.7	40.7
2	23.3	22.7
3	18.0	22.0
4	9.3	18.0
5	6.7	8.0
6	4.0	2.7
7	3.3	4.0
8	2.0	3.3
9	1.3	4.0
10	1.3	3.3
11	-	2.7

[1] 30 samples of each variety.

table 20 :Arithmetic means and medians: comparing the performance of serial combinations of descriptors from the subgroup {SHAPE} in terms of their ability to self-classify samples of each variety correctly. Model reference numbers refer to those combinations of descriptors given in tables 14 and 16 for arithmetic means and medians respectively.

model	no. cases[1]	%error	allocation of %error amongst varieties (as % no.cases in error) hcn	igr	kin	mot	pda
1	63	42.0	39.7	0.0	19.0	31.7	9.5
2	31	20.7	45.2	0.0	32.3	22.5	0.0
3	18	12.0	50.0	0.0	5.6	44.4	0.0
4	4	2.7	100.0	0.0	0.0	0.0	0.0
5	2	1.3	100.0	0.0	0.0	0.0	0.0
6	2	1.3	100.0	0.0	0.0	0.0	0.0
7	1	0.7	100.0	0.0	0.0	0.0	0.0
8	1	0.7	100.0	0.0	0.0	0.0	0.0
9	1	0.7	100.0	0.0	0.0	0.0	0.0
10	1	0.7	100.0	0.0	0.0	0.0	0.0
11	0	0.0	0.0	0.0	0.0	0.0	0.0
12	0	0.0	0.0	0.0	0.0	0.0	0.0

[1] total number of cases in which variety incorrectly self-classified.

table 21 :Arithmetic means: showing the number of cases of each variety which were incorrectly self-classified using each serial combination of descriptors from the {SIZE} subgroup. Number of cases incorrectly assigned in each variety expressed as a percentage of the total number of incorrect assignations.

model	no. cases[1]	%error	allocation of %error amongst varieties (as % no.cases in error) hcn	igr	kin	mot	pda
1	75	50.0	17.3	13.3	36.0	17.3	16.0
2	27	18.0	33.3	3.7	11.1	51.9	0.0
3	10	6.7	50.0	0.0	10.0	40.0	0.0
4	5	3.3	60.0	0.0	0.0	60.0	0.0
5	6	4.0	50.0	0.0	0.0	50.0	0.0
6	4	2.7	50.0	0.0	0.0	50.0	0.0
7	2	1.3	100.0	0.0	0.0	0.0	0.0
8	1	0.7	100.0	0.0	0.0	0.0	0.0
9	2	1.3	50.0	0.0	0.0	50.0	0.0
10	0	0.0	0.0	0.0	0.0	0.0	0.0
11	0	0.0	0.0	0.0	0.0	0.0	0.0
12	1	0.7	100.0	0.0	0.0	0.0	0.0

[1] total number of cases in which variety incorrectly self-classified.

table 22 :Medians: showing the number of cases of each variety which were incorrectly self-classified using each serial combination of descriptors from the {SIZE} subgroup. Number of cases incorrectly assigned in each variety expressed as a percentage of the total number of incorrect assignations.

model	no. cases[1]	%error	allocation of %error amongst varieties (as % no.cases in error) hcn	igr	kin	mot	pda
1	61	40.7	37.7	0.0	9.8	32.8	19.7
2	35	23.3	31.4	0.0	20.0	42.9	5.7
3	27	18.0	33.3	0.0	22.2	40.7	3.7
4	14	9.3	57.1	0.0	7.1	35.7	0.0
5	10	6.7	80.0	0.0	0.0	20.0	0.0
6	6	4.0	83.3	0.0	0.0	16.7	0.0
7	5	3.3	80.0	0.0	0.0	20.0	0.0
8	3	2.0	66.7	0.0	0.0	33.3	0.0
9	2	1.3	100.0	0.0	0.0	0.0	0.0
10	2	1.3	100.0	0.0	0.0	0.0	0.0

[1] total number of cases in which variety incorrectly self-classified.

table 23 :Arithmetic means: showing the number of cases of each variety which were incorrectly self-classified using each serial combination of descriptors from the {SHAPE} subgroup. Number of cases incorrectly assigned for each variety expressed as a percentage of the total number of incorrect assignations.

model	no. cases[1]	%error	allocation of %error amongst varieties (as % no.cases in error) hcn	igr	kin	mot	pda
1	61	40.7	32.8	1.6	11.5	29.5	24.6
2	35	23.3	25.7	2.9	22.8	40.0	8.6
3	32	21.3	28.1	0.0	34.3	31.2	6.5
4	27	18.0	33.3	0.0	33.3	33.3	0.0
5	12	8.0	50.0	0.0	16.7	33.3	0.0
6	4	2.7	100.0	0.0	0.0	0.0	0.0
7	6	4.0	100.0	0.0	0.0	0.0	0.0
8	5	3.3	100.0	0.0	0.0	0.0	0.0
9	6	4.0	83.3	0.0	0.0	16.7	0.0
10	5	3.3	100.0	0.0	0.0	0.0	0.0
11	4	2.7	75.0	0.0	0.0	25.0	0.0

[1] total number of cases in which variety incorrectly self-classified.

table 24 :Medians: showing the number of cases of each variety which were incorrectly self-classified using each serial combination of descriptors from the {SHAPE} subgroup. Number of cases incorrectly assigned for each variety expressed as a percentage of the total number of incorrect assignations.

training set	N[1]	test set	n[2]	correctly classified as variety: SIZE hcn	igr	kin	mot	pda	%error[3]	correctly identified as variety: SHAPE hcn	igr	kin	mot	pda	%error[3]
C_{88}	150	X_{89}	75	15	15	14	15	15	1.3	14	15	15	13	15	4.0
C_{88}	150	X_{90}	75	15	15	15	11	15	5.3	15	15	15	11	15	5.3
C_{88+89}	175	Y_{89}	50	10	10	10	9	10	2.0	10	10	9	9	10	4.0
C_{88+90}	175	Y_{90}	50	10	10	10	9	10	2.0	10	10	10	9	10	2.0
$C_{88}+X_{89}$	225	X_{90}	75	15	15	12	13	15	6.7	15	15	12	12	15	8.0

[1]N= total number of samples within training set.

[2]n= total number of samples within test set.

[3]%error= number of misclassified test samples as % total number of test samples.

table 25 :Arithmetic means: classification of "unknown" varieties in test sets using samples from training sets shown to establish location of groups of variety samples in discriminant space. Training and test set codes are those given in table 2.
Figures given in the table show the number of samples of each variety correctly identified for {SIZE} and {SHAPE} descriptor subgroups.

training set	N[1]	test set	n[2]	correctly classified as variety: SIZE hcn	igr	kin	mot	pda	%error[3]	correctly identified as variety: SHAPE hcn	igr	kin	mot	pda	%error[3]
C_{88}	150	X_{89}	75	14	15	14	14	15	4.0	11	15	14	11	15	12.0
C_{88}	150	X_{90}	75	13	15	15	11	15	8.0	12	15	15	11	15	9.1
C_{88+89}	175	Y_{89}	50	10	10	9	10	10	2.0	10	10	9	8	10	6.0
C_{88+90}	175	Y_{90}	50	10	10	10	8	10	4.0	10	10	10	8	10	4.0
$C_{88}+X_{89}$	225	X_{90}	75	15	15	12	13	15	6.7	15	15	15	13	15	2.6

[1]N= total number of samples within training set.

[2]n= total number of samples within test set.

[3]%error=number of samples misclassified as % of total number of samples in test set.

table 26 :medians: classification of "unknown" varieties in test sets using samples from training sets shown to establish location of groups of variety samples in discriminant space. Training and test set codes are those given in table 2.
Figures given in the table show the number of samples of each variety correctly identified for {SIZE} and {SHAPE} descriptor subgroups.

training set	N[1]	n[2]	correctly classified as variety: SIZE hcn	igr	kin	mot	pda	%error[3]	SHAPE hcn	igr	kin	mot	pda	%error[3]
C_{88}	150	30	30	30	30	30	30	0.0	29	30	30	30	30	0.7
C_{88+89}	175	35	35	35	35	35	35	0.0	34	35	35	35	35	0.6
C_{88+90}	175	35	35	35	35	35	35	0.0	34	35	34	35	35	1.1
$C_{88}+X_{89}$	225	45	45	45	44	45	45	0.4	43	45	44	45	45	1.3

[1]N= total number of samples within training set.

[2]n= number of samples for each variety.

[3]% error=total number of samples assigned incorrectly as % of total number of samples.

table 27 :Arithmetic means: self-classification of varieties in training sets using samples from one year only (C_{88}) or from two different years. C_{88+89} comprises 30 samples of each variety from 1988 harvested stocks plus 5 samples of each variety from 1989 harvested stocks; similarly, C_{88+90} comprises samples from 1988 and 1990 in the same relative proportions. $C_{88}+X_{89}$ combines all available samples from 1988 and 1989 into one training set. Figures given in the table show the number of samples correctly self-classified within each training series for {SIZE} and {SHAPE} descriptor subgroups.

training set	N[1]	n[2]	correctly classified as variety: SIZE hcn	igr	kin	mot	pda	%error[3]	SHAPE hcn	igr	kin	mot	pda	%error[3]
C_{88}	150	30	29	30	30	30	30	0.7	28	30	30	29	30	2.0
C_{88+89}	175	35	35	35	35	35	35	0.0	31	35	35	33	35	3.4
C_{88+90}	175	35	34	35	35	35	35	0.6	32	35	35	34	35	2.3
$C_{88}+X_{89}$	225	45	45	45	44	45	45	0.4	41	45	45	45	45	1.8

[1]N= total number of samples within training set.

[2]n= number of samples for each variety.

[3]%error=number of samples incorrectly assigned as % of total number of samples.

table 28 :medians: self-classification of varieties in training sets using samples from one year only (C_{88}) or from two different years. C_{88+89} comprises 30 samples of each variety from 1988 harvested stocks plus 5 samples of each variety from 1989 harvested stocks; similarly, C_{88+90} comprises samples from 1988 and 1990 in the same relative proportions. $C_{88}+X_{89}$ combines all available samples from 1988 and 1989 into one training set. Figures given in the table show the number of samples correctly self-classified within each training series for {SIZE} and {SHAPE} descriptor subgroups.

model reference number	ARITHMETIC MEANS %error (n=15)[1] 1989 %	1990 %	MEDIANS %error (n=15)[1] 1989 %	1990 %
1	41.3	37.3	74.7	62.7
2	19.3	46.7	41.3	48.0
3	33.3	42.7	16.0	29.3
4	10.7	25.3	13.3	26.7
5	5.3	10.7	14.7	21.3
6	4.0	10.7	10.7	21.3
7	4.0	16.0	10.7	18.7
8	6.7	12.0	8.0	10.7
9	6.7	10.7	10.7	13.3
10	8.0	8.0	12.0	10.7
11	5.3	10.7	6.7	10.7
12	4.0	4.0	6.7	8.0

[1] 15 samples of each variety.

table 29 :Arithmetic means and medians: comparing the performance of serial combinations of descriptors from the subgroup {SIZE} in terms of their ability to classify 15 "unknown" samples of each variety taken from 1989 and 1990 harvested stocks.(Sample series X_{89} and X_{90} respectively.)
Classification "rules" developed from 30 samples of each variety drawn from 1988 harvested stocks in samples series C_{88}.
Model reference numbers refer to those combinations of descriptors given in tables 13 and 15 for arithmetic means and medians respectively.

model reference number	ARITHMETIC MEANS %error (n=15)[1] 1989 %	1990 %	MEDIANS %error (n=15)[1] 1989 %	1990 %
1	46.7	38.7	38.7	36.0
2	26.7	38.7	32.0	38.7
3	28.0	40.0	25.3	44.0
4	16.0	45.3	26.7	52.0
5	24.0	33.3	20.0	42.7
6	25.3	33.3	20.0	18.7
7	20.0	18.7	20.0	17.3
8	16.0	17.3	16.0	14.7
9	6.7	8.0	16.0	16.0
10	6.7	6.7	14.7	14.7
11	-	-	13.3	8.0

[1] 15 samples of each variety.

table 30 :Arithmetic means and medians: comparing the performance of serial combinations of descriptors from the subgroup {SHAPE} in terms of their ability to classify 15 "unknown" samples of each variety taken from 1989 and 1990 harvested stocks.(Sample series X_{89} and X_{90} respectively.)
Classification "rules" developed from 30 samples of each variety drawn from 1988 harvested stocks in samples series C_{88}.
Model reference numbers refer to those combinations of descriptors given in tables 14 and 16 for arithmetic means and medians respectively.

model	no. cases[1]	%error[2]	1989 samples:arithmetic means %contribution of variety error to total error. hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]	no. cases[1]	%error[2]	1990 samples: arithmetic means %contribution of variety error to total error. hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]
1	31	41.3	41.9	6.5	3.2	29.0	19.4	28	37.3	46.4	7.1	7.1	28.6	10.7
2	29	19.3	51.7	6.9	17.2	24.1	0.0	35	46.7	42.9	5.6	42.9	0.0	8.6
3	25	33.3	48.0	8.0	16.0	28.0	0.0	32	42.7	46.9	6.2	46.9	0.0	0.0
4	8	10.7	12.5	25.0	12.5	50.0	0.0	19	25.3	68.4	10.0	10.5	21.1	0.0
5	4	5.3	50.0	25.0	25.0	0.0	0.0	8	10.7	50.0	0.0	12.5	37.5	0.0
6	3	4.0	66.7	0.0	33.3	0.0	0.0	8	10.7	50.0	0.0	12.5	37.5	0.0
7	3	4.0	66.7	0.0	33.3	0.0	0.0	12	16.0	41.7	0.0	8.3	50.0	0.0
8	5	6.7	60.0	0.0	20.0	20.0	0.0	9	12.0	44.4	0.0	0.0	55.6	0.0
9	5	6.7	80.0	0.0	0.0	20.0	0.0	8	10.7	50.0	0.0	0.0	50.0	0.0
10	6	8.0	66.7	0.0	16.7	16.7	0.0	6	8.0	50.0	0.0	0.0	50.0	0.0
11	4	5.3	75.0	0.0	25.0	0.0	0.0	8	10.7	50.0	0.0	0.0	50.0	0.0
12	3	4.0	66.7	0.0	33.3	0.0	0.0	3	4.0	33.3	0.0	0.0	66.7	0.0

[1] total number of cases in which variety incorrectly classified.

[2] %error: (total number of cases incorrectly classified / total number of cases) * 100.

[3] number of cases of each variety incorrectly classified / total number of cases incorrectly classified)

table 31 :Arithmetic means: showing the number of cases of each variety which were incorrectly classified using each serial combination of descriptors from the {SIZE} subgroup. Classification "rules" based on 30 samples of each variety drawn from 1988 stocks in sample series C88.
15 samples from each variety drawn from 1989 and 1990 stocks (sample series X89 and X90 respectively) not included in the establishment of classification "rules".
Model numbers refer to those given in table 13.

model	no. cases[1]	%error[2]	1989 samples: arithmetic means %contribution of variety error to total error. hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]	no. cases[1]	%error[2]	1990 samples: arithmetic means %contribution of variety error to total error. hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]
1	35	46.7	37.1	5.7	2.9	31.4	22.9	29	38.7	48.3	6.9	3.4	31.0	10.3
2	20	26.7	35.0	10.0	0.0	55.0	0.0	29	38.7	51.7	6.9	3.4	31.0	6.9
3	21	28.0	47.6	4.8	4.8	42.8	0.0	30	40.0	50.0	3.3	0.0	43.3	3.3
4	12	16.0	8.3	8.3	8.3	75.0	0.0	34	45.3	41.2	0.0	32.4	26.5	0.0
5	18	24.0	38.9	5.6	11.1	44.4	0.0	25	33.3	60.0	0.0	4.0	36.0	0.0
6	19	25.3	52.6	0.0	5.3	42.1	0.0	25	33.3	56.0	0.0	0.0	44.0	0.0
7	15	20.0	13.3	6.7	6.7	73.3	0.0	14	18.7	28.6	0.0	7.1	64.3	0.0
8	12	16.0	16.7	8.3	8.3	66.7	0.0	13	17.3	30.7	0.0	7.7	61.5	0.0
9	5	6.7	40.0	0.0	20.0	40.0	0.0	6	8.0	16.7	0.0	16.7	66.6	0.0
10	5	6.7	40.0	0.0	20.0	40.0	0.0	5	6.7	0.0	0.0	20.0	80.0	0.0

[1] total number of cases in which variety incorrectly classified.

[2] %error: (total number of cases incorrectly classified / total number of cases) * 100.

[3] number of cases of each variety incorrectly classified / total number of cases incorrectly classified)

table 32 :Arithmetic means: showing the number of cases of each variety which were incorrectly classified using each serial combination of descriptors from the {SHAPE} subgroup. Classification "rules" based on 30 samples of each variety drawn from 1988 stocks in sample series C_{88}.
15 samples from each variety drawn from 1989 and 1990 stocks (sample series X_{89} and X_{90} respectively) not included in the establishment of classification "rules".
Model numbers refer to those given in table 15.

	1989 samples: medians %contribution of variety error to total error.							1990 samples: medians %contribution of variety error to total error.						
model	no. cases[1]	%error[2]	hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]	no. cases[1]	%error[2]	hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]
1	56	74.7	21.4	14.3	26.8	25.0	12.5	47	62.7	23.4	8.3	27.7	21.3	23.4
2	31	41.3	29.0	9.7	16.1	45.2	0.0	36	48.0	41.7	5.6	33.3	11.1	8.3
3	12	16.0	33.3	25.0	8.3	33.3	0.0	22	29.3	59.1	9.1	0.0	31.8	0.0
4	10	13.3	40.0	30.0	10.0	20.0	0.0	20	26.7	65.0	10.0	0.0	23.0	0.0
5	11	14.7	45.5	27.3	9.1	18.2	0.0	16	21.3	75.0	6.3	0.0	18.7	0.0
6	8	10.7	62.5	0.0	12.5	25.0	0.0	16	21.3	75.0	0.0	0.0	25.0	0.0
7	8	10.7	50.0	0.0	12.5	37.5	0.0	14	18.7	64.3	0.0	0.0	35.7	0.0
8	6	8.0	83.3	0.0	16.7	0.0	0.0	8	10.7	62.5	0.0	0.0	37.5	0.0
9	8	10.7	87.5	0.0	0.0	12.5	0.0	10	13.3	60.0	0.0	0.0	40.0	0.0
10	9	12.0	88.9	0.0	11.1	0.0	0.0	8	10.7	50.0	0.0	0.0	50.0	0.0
11	5	6.7	80.0	0.0	20.0	0.0	0.0	8	10.7	50.0	0.0	0.0	50.0	0.0
12	5	6.7	60.0	0.0	20.0	20.0	0.0	6	8.0	33.3	0.0	0.0	66.7	0.0

[1] total number of cases in which variety incorrectly classified.

[2] %error: (total number of cases incorrectly classified / total number of cases) * 100.

[3] number of cases of each variety incorrectly classified / total number of cases incorrectly classified)

table 33 : medians: showing the number of cases of each variety which were incorrectly classified using each serial combination of descriptors from the {SIZE} subgroup. Classification "rules" based on 30 samples of each variety drawn from 1988 stocks in sample series C_{88}.
15 samples from each variety drawn from 1989 and 1990 stocks (sample series X_{89} and X_{90} respectively) not included in the establishment of classification "rules".
Model numbers refer to those given in table 14.

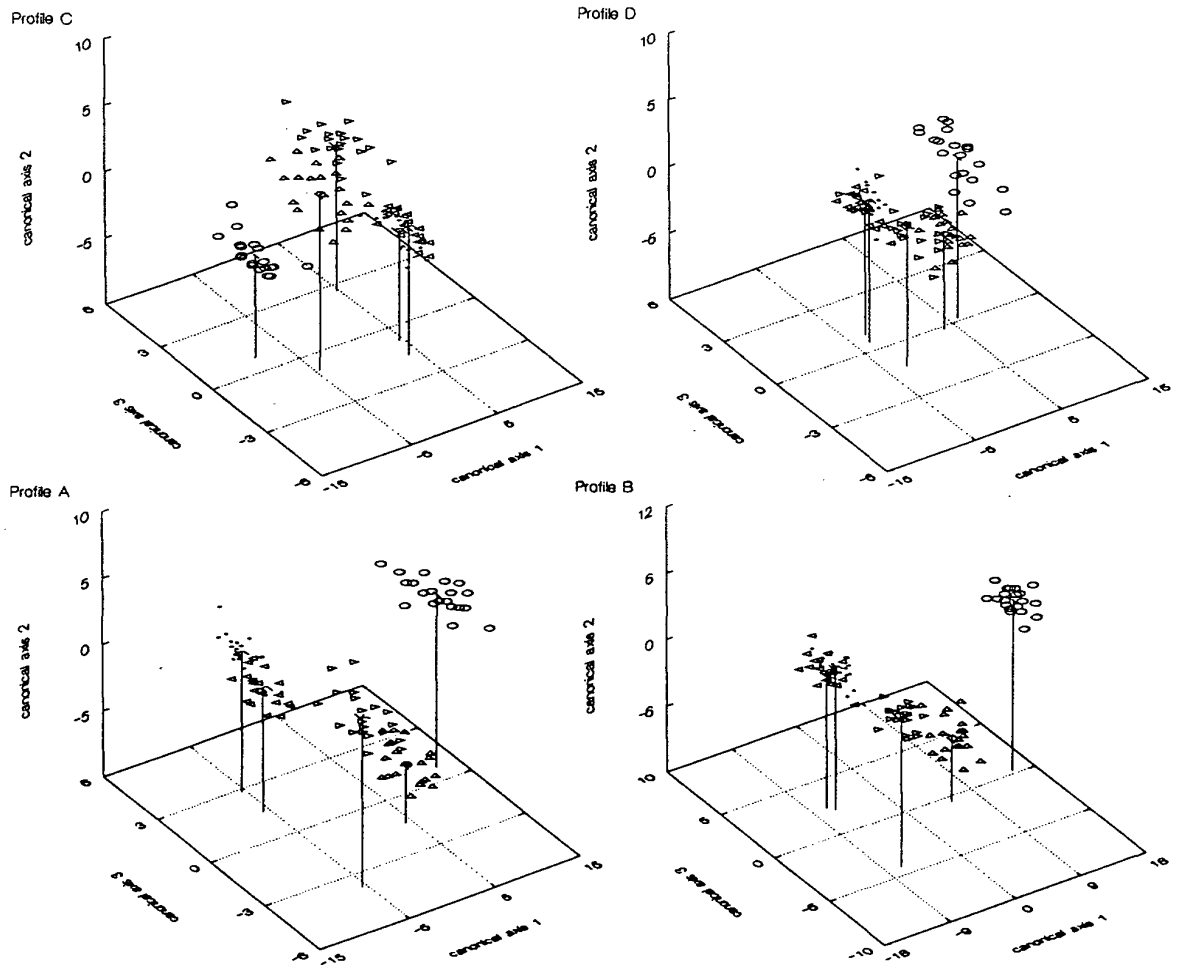
model	1989 samples: medians %contribution of variety error to total error.							1990 samples: medians %contribution of variety error to total error.						
	no. cases[1]	%error[2]	hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]	no. cases[1]	%error[2]	hcn[3]	igr[3]	kin[3]	mot[3]	pda[3]
1	29	38.7	51.7	10.3	6.9	10.3	20.7	27	36.0	55.6	7.4	18.5	18.5	0.0
2	24	32.0	16.7	12.5	20.8	50.0	0.0	29	38.7	51.7	6.9	6.9	34.5	0.0
3	17	25.3	23.5	11.8	0.0	64.7	0.0	30	44.0	50.0	6.7	6.7	36.6	0.0
4	20	26.7	25.0	10.0	10.0	55.0	0.0	39	52.0	38.5	5.1	12.8	33.3	10.3
5	15	20.0	13.3	20.0	6.7	60.0	0.0	32	42.0	34.4	0.0	28.1	37.5	0.0
6	15	20.0	20.0	20.0	0.0	60.0	0.0	14	18.7	28.6	14.3	0.0	57.1	0.0
7	15	20.0	20.0	13.3	6.7	60.0	0.0	13	17.3	23.1	0.0	0.0	76.9	0.0
8	12	16.0	33.3	0.0	8.3	58.3	0.0	11	14.7	27.3	0.0	0.0	72.7	0.0
9	12	16.0	25.0	8.3	8.3	58.3	0.0	12	16.0	33.3	0.0	0.0	66.7	0.0
10	11	14.7	36.4	0.0	9.1	54.5	0.0	11	14.7	36.4	0.0	0.0	63.6	0.0
11	10	13.3	40.0	0.0	10.0	50.0	0.0	6	8.0	50.0	0.0	0.0	50.0	0.0

[1] total number of cases in which variety incorrectly classified.

[2] %error: (total number of cases incorrectly classified / total number of cases) * 100.

[3] number of cases of each variety incorrectly classified / total number of cases incorrectly classified)

table 34 :medians: showing the number of cases of each variety which were incorrectly classified using each serial combination of descriptors from the {SHAPE} subgroup. Classification "rules" based on 30 samples of each variety drawn from 1988 stocks in sample series C_{88}.
15 samples from each variety drawn from 1989 and 1990 stocks (sample series X_{89} and X_{90} respectively) not included in the establishment of classification "rules".
Model numbers refer to those given in table 16.



KEY:

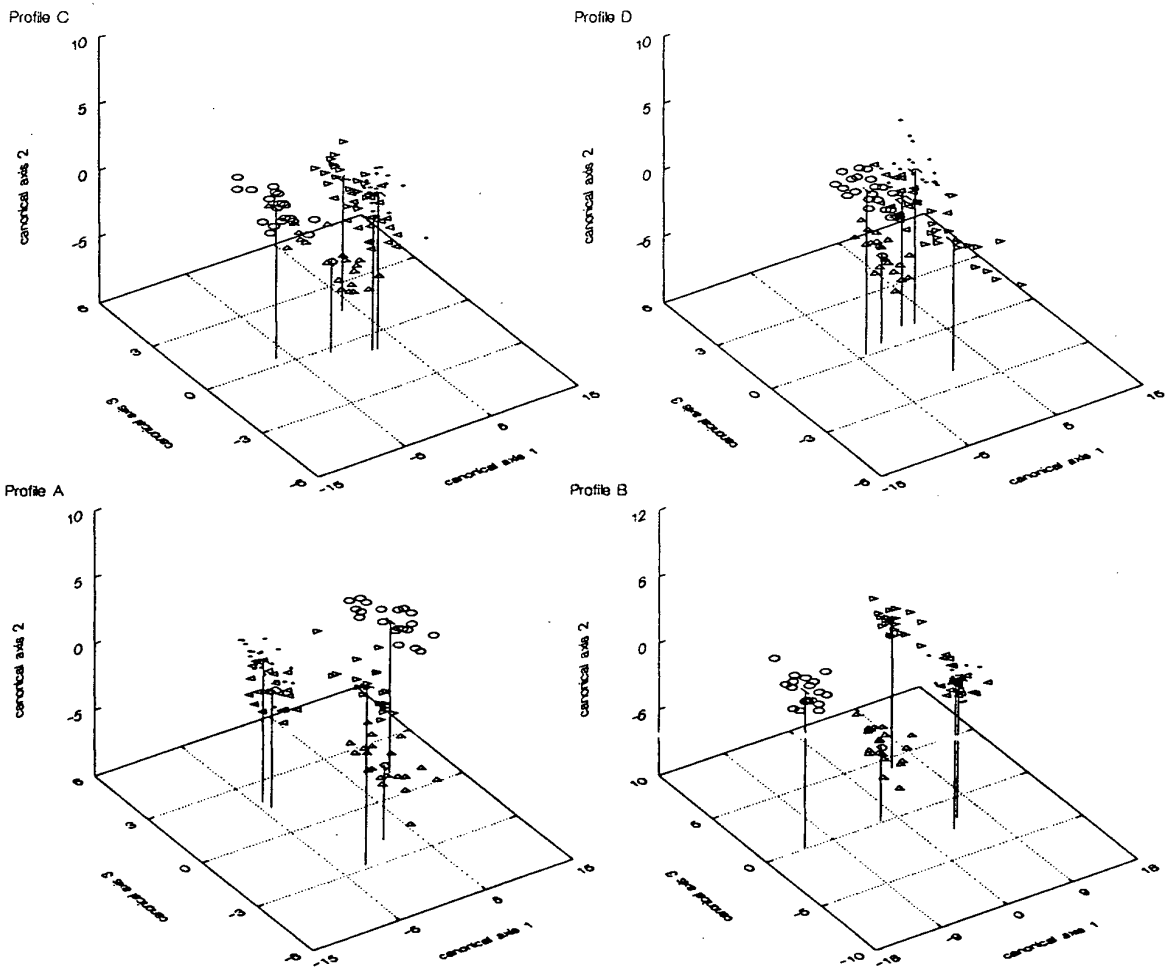
variety:	symbols:	
Panda	◁	P/p
MOtter	▽	M/m
Pipkin	△	K/k
Igri	O	I/i
Halcyon	.	H/h

centroids shown as points anchored to xy facet

Figure 1:

1988 arithmetic means. 20 cases per variety: distribution of varieties in discriminant space

classifier based on measurements of size complying with assumptions of normality and homogeneity of variances



KEY:

variety:	symbols:	
Panda	◁	P/p
McCiter	▽	M/m.
Pipkin	△	K/k
Igri	O	I/i
Halcyon	.	H/h

centroids shown as points anchored to xy facet

Figure 2:

1988 arithmetic means. 20 cases per variety: distribution of varieties in discriminant space

classifier based on measurements of shape complying with assumptions of normality and homogeneity of variances

KEY:

variety:	symbols:	
Panda	◁	P/p
M.Otter	▽	M/m
Pipkin	△	K/k
Igri	○	I/i
Halcyon	·	H/h

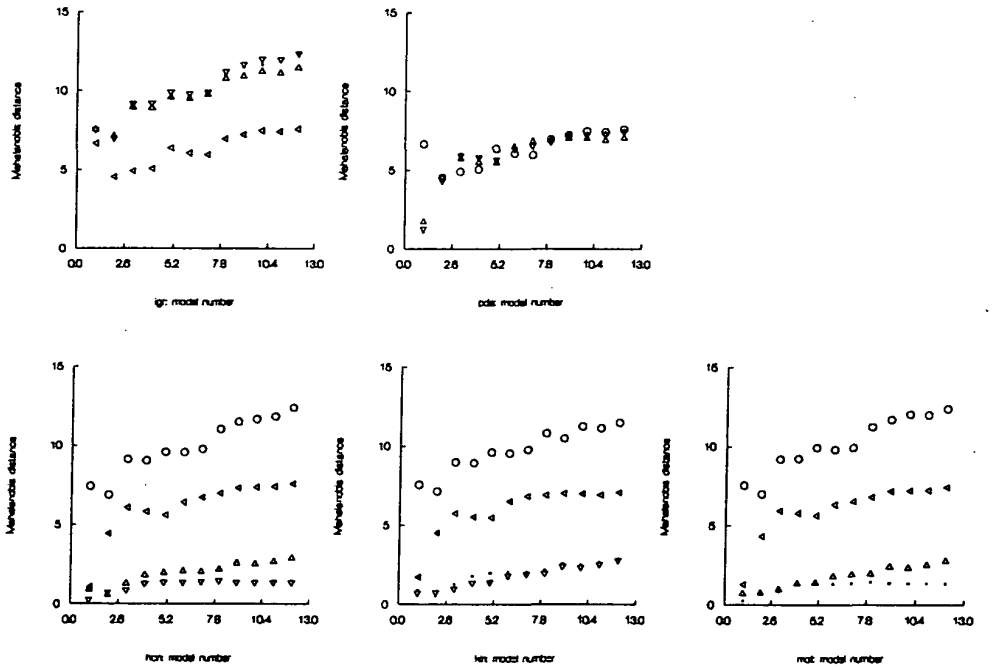


Figure 3:

1988 arithmetic means. 30 cases per variety: distances between varieties

classifier based on measurements of size: distance = Mahalanobis distance

KEY:

variety:	symbols:	
Panda	◁	P/p
M.Otter	▽	M/m
Pipkin	△	K/k
Igri	O	I/i
Halcyon	.	H/h

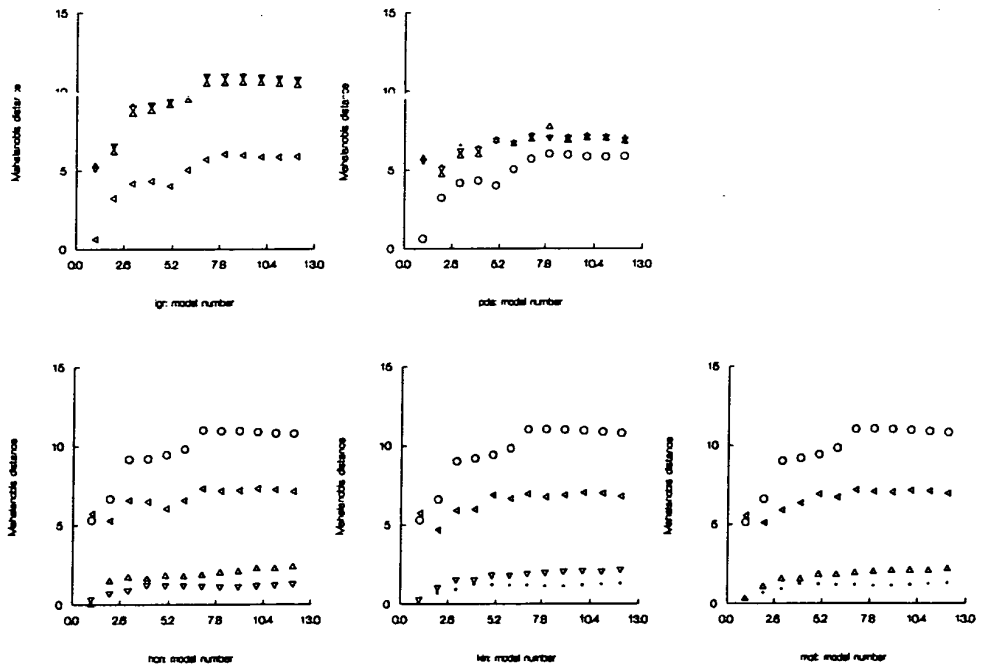


Figure 4:

1988 medians. 30 cases per variety: distances between varieties

classifier based on measurements of size: distance = Mahalanobis distance

KEY:

variety:	symbols:	
Panda	◁	P/p
M.Otter	▽	M/m
Pipkin	△	K/k
Igri	O	I/i
Halcyon	•	H/h

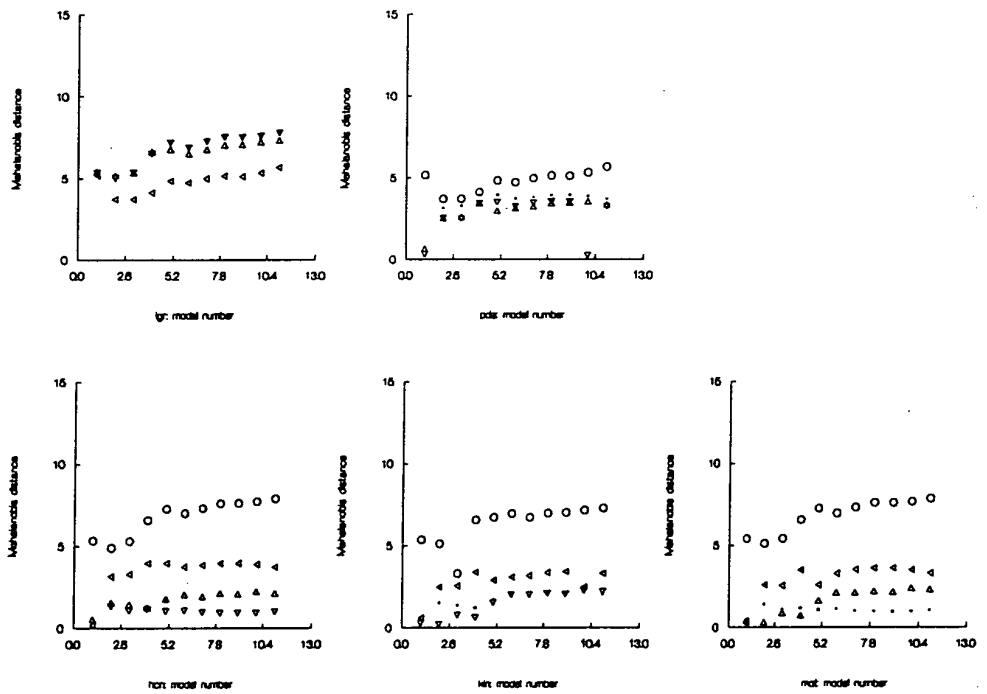


Figure 5:

1988 medians. 30 cases per variety: distances between varieties

classifier based on measurements of shape: distance = Mahalanobis distance

KEY:

variety:	symbols:	
Panda	◁	P/p
M.Otter	▽	M/m
Pipkin	△	K/k
Igri	○	I/i
Halcyon	·	H/h

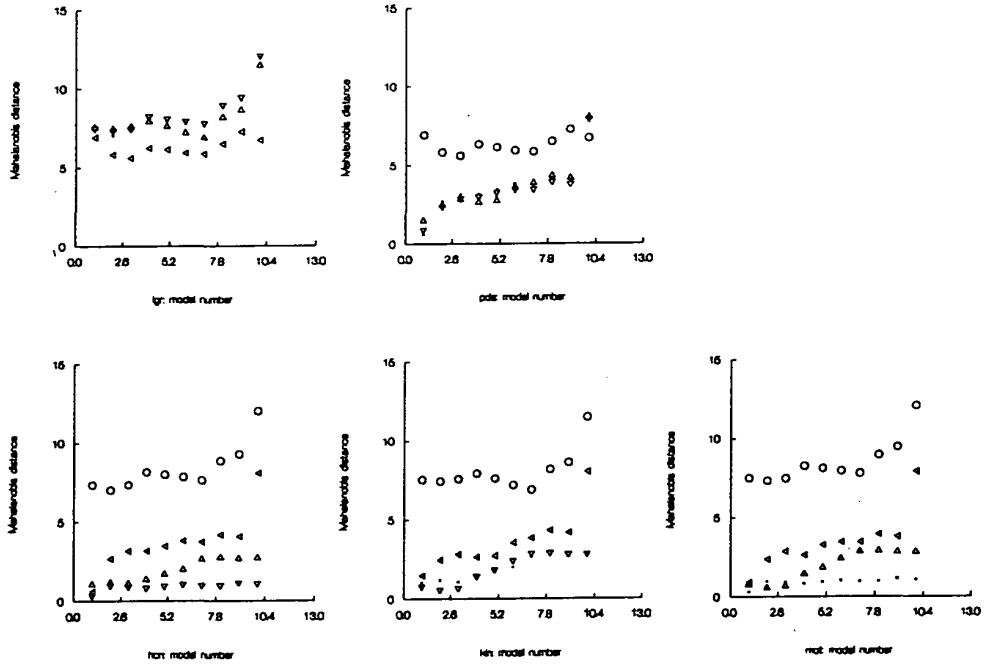


Figure 6:

1988 arithmetic means. 30 cases per variety: distances between varieties

classifier based on measurements of shape: distance = Mahalanobis distance