J. Sci. Food Agric. 1979, **30**, 669–676

Use of Canonical Variate Analysis in the Differentiation of Swede Cultivars by Gas–Liquid Chromatography of Volatile Hydrolysis Products

Rosemary A. Cole and Kathleen Phelps

National Vegetable Research Station, Wellesbourne, Warwickshire

(*Manuscript received 31 January 1979*)

Swede cultivars can be differentiated by their volatile hydrolysis products obtained after maceration. Multivariate statistical techniques however, are required to interpret the data because of the interactions between volatiles. Canonical variate analysis allowed a chemical interpretation to be placed on the effects of storage and provided a basis for differentiating between cultivars. The relative chemical configuration of the cultivars is displayed graphically by plotting the cultivar means relative to the first two canonical variates.

1. Introduction

When cruciferous material is macerated, released thioglucosidases and glucosinolates come into contact and, in the presence of moisture, hydrolysis occurs giving rise to many volatile products, including isothiocyanates and nitriles. Both glucosinolates and the volatile hydrolysis products of cruciferous plants can affect insect behaviour,[1–3] the volatile hydrolysis products probably attracting the insects to the host-plant.[4] The marked qualitative and quantitative differences in the glucosinolates in different cultivars of crucifers[5] may account for their differing susceptibilities to attack by some insect pests. Swedes are frequently stored both for human and animal consumption and the compositions of some cultivars may change more than others in storage. At Wellesbourne, swedes used to rear insects need to be stored for many months and changes in their composition during storage affect the productivity of the insect cultures. The compositions of 14 swede cultivars were therefore determined to detect those that changed most when they had been kept in a cold store for 3 months or left in the field to overwinter. Comparisons among the cultivars, both fresh (overwintered) and stored, were also of interest for insect-rearing purposes.

Although there are methods available to detect chemical changes, the results are often complex and hence difficult to interpret when comparing, for instance, cultivars. When there are only two or three volatile hydrolysis products of interest, the mean amounts or concentrations may provide adequate descriptions to distinguish cultivars.[5,6] However, when several chemicals are of interest, and especially if some interact, a method is needed to describe the overall configuration of the volatile hydrolysis products. Multivariate statistical techniques offer a powerful aid to the interpretation of such data. In effect they enable the mass of results representing the chemical configuration of a sample to be assigned to a single point in multi-dimensional space. MacFie, Gutteridge and Norris[7] discussed the use of several multivariate techniques to process peak heights from pyrolysis g.l.c. Of the several multivariate statistical techniques which they considered, canonical variate analysis was chosen here because it provided a method for distinguishing between groups (cultivars) and objects (swedes) using the criterion that members of the same group should be more similar to each other than to members of different groups. This paper reports the results obtained when the 14 cultivars were examined in this way, whether overwintered in the field or after storage.

0022–5142/79/0700–0669 $02.00

2. Method

2.1. Experimental

2.1.1. Plant material

Fourteen swede cultivars (Table 1) were grown in rows 55 cm apart from July 1977 to February 1978 in a field at Wellesbourne and 1 month after sowing the plants were thinned to 30 cm apart in the rows. No pesticides were used. Half of the swedes were harvested in November and stored until February at approximately 5 °C, the remainder being left in the field until February.

Table 1. Swede cultivars used in multivariate analysis

A.	Acme	H.	Mancunian
B.	Best of all	I.	Peerless
C.	Conqueror	J.	Scotia
D.	Danestone	K.	Seefelder
E.	Doon Major	L.	Victory
F.	Eclipse	M.	Vogessa
G.	Magnificient	N.	Wilhelmsburger

2.1.2. Sample preparation and chemical analysis

After washing and removing the leaves and rootlets, the swedes were weighed and sliced using a Moulinex slicer. A 10 g subsample was macerated with 10 ml distilled water and allowed to hydrolyse for 1 h at 20 °C. The hydrolysis products were extracted with CH_2Cl_2 and the fibrous material removed by centrifugation. Extracts were prepared for each of five fresh and five stored swedes for each of the cultivars. The compositions of the extracts were determined using a PYE 104 gas chromatograph, as reported previously, and the compounds were identified by g.c.–m.s.[8] Pentadecane was injected with each extract as an internal standard.

2.1.3. Records

The heights of eight peaks of known identity were measured to the nearest millimetre, corrected for g.l.c. variation by the internal standard and the relative detector response and the amounts of volatile compounds from each swede were then calculated. The description of the swedes would be incomplete without a measure of weight because the amount of volatile chemicals produced on hydrolysis is a function of weight. As this function is not linear it is difficult to make a satisfactory simple correction for weight and therefore weight was included as an extra variate in the statistical analysis.

2.2. Statistical methods

Canonical variate analysis is a well-documented statistical technique, the mathematical details of which may be found in many multivariate analysis text books and computer program manuals.[9–12] The following is a very brief description of the relevant principles of the technique and its terminology.

Each unit of data was considered as a point in p-dimensional space; in this case each swede was considered as a point in nine-dimensional space (8 peak heights plus weight of swede). The data were assumed to fall naturally into g groups (g being the number of cultivars analysed). A multivariate analysis of variance was first done to test the null hypothesis of no differences between the cultivars. On rejection of this hypothesis, linear combinations of the original variates were calculated which maximised the ratio of between- to within-group variance. The analysis thus defined a new set of variates which emphasised the differences between the groups (cultivars) and the similarities within them.

The theoretical requirements for data which are to be subjected to multivariate analysis of variance are extensions of the homogeneity of variance requirements for univariate analysis. There

is the additional requirement that the correlation between the original variates must be the same for each unit of the data. Thus correlations between peaks 1 and 2 for the replicates for cultivar A must be similar to those for the replicates of cultivar B. This 'homogeneity of dispersion matrix' requirement is difficult to test but seems to be intuitively reasonable in the present context. The primary aim of the analysis is to reduce the dimensionality of the data so that most of the characteristics of the data-set can be expressed by a small number of new (canonical) variates. In this paper only the first two canonical variates from each analysis are quoted, the relevant equations for swede, s, being:

$$C_{1s} = d_{11}h_{1s} + d_{12}h_{2s} + \ldots d_{18}h_{8s} + d_{19}w_s \ldots \quad (1)$$

$$C_{2s} = d_{21}h_{1s} + d_{22}h_{2s} + \ldots d_{28}h_{8s} = d_{29}w_s \ldots \quad (2)$$

where C_{1s} and C_{2s} are the first and second canonical variate *scores* for swede, s; $h_{1s} \ldots h_{8s}$ are the standardised heights of peak 1 . . . 8 for swede, s; w_s is the weight of swede, s; $s = 1 \ldots n$ where n is the number of swedes; $d_{11} \ldots d_{19}$ are the calculated *loadings* for *canonical* variate 1; $d_{12} \ldots d_{29}$ are the calculated *loadings* for *canonical* variate 2. The loadings are such that C_{is}^1 are constrained to have unit variance.

In order to centre the plotted points around zero the quantity $C_{is}^1 = C_{is} - C_i$, equation (3) was calculated, C_i being the mean of the scores for canonical variate, i. The co-ordinates used for plotting the positions of the cultivars relative to the canonical axes are the mean values of C_{is} for the five replicates of each cultivar. The circular 95% confidence limits have radii of $\sqrt{5.99}/\sqrt{5}$ where $\sqrt{5.99}$ is the chi square on 2 degrees of freedom (because C_{is} have unit variance) and 5 is the number of replicates on which each mean is based.

3. Results and data analysis

3.1. Chemical composition

Table 2 lists the eight volatile hydrolysis products measured.

Table 2. Volatile compounds obtained from hydrolysis of swede cultivars

G.l.c. peak no.	Volatile hydrolysis product	Parent glucosinolate
1	1-Cyano-4-methylthiobutane	4-Methylthiobutyl glucosinolate
4	4-Methylthiobutylisothiocyanate	
2	2-Phenylproprionitrile	2-Phenethyl glucosinolate
5	2-Phenethylisothiocyanate	
3	1-Cyano-5-methythiopentane	5-Methylthiopentyl glucosinolate
6	5-Methylthiopentylisothiocyanate	
7	α-1-Cyano-2-hydroxy epithiobutane	Progoitrin
8	β-1-Cyano-2-hydroxy epithiobutane	

3.2. Statistical analysis

3.2.1. Analysis of variance

Table 3 shows the mean weights and the amounts of volatile hydrolysis products found in seven cultivars representing the range of composition encountered in both stored and fresh swedes. The mean composition of all 14 cultivars is also shown, together with standard errors derived conventionally from an analysis of variance.

The large standard errors associated with the data were not mainly attributable to the chemical analysis since replicate samples from single swedes of several varieties had been shown independently to result in standard errors of only ±10% of the mean values for the components. Some of the standard errors in Table 3 were larger than the cultivar means which suggested that a transformation of the data might be warranted, but the individual replicate values did not indicate any inherent

Table 3. The mean weights and volatile compounds in seven representative swede cultivars and the mean composition of the 14 cultivars analysed fresh and after 14 weeks storage at 5°C

Cultivar	Wt (kg)	Volatile compounds (mg per swede)							
		1	4	2	5	3	6	7	8
Fresh swedes									
H	0.48	0.37	0.99	1.17	6.23	2.31	3.78	0.22	0.24
A	0.40	0.84	0.78	2.02	5.47	5.41	2.80	0.45	0.46
B	0.38	0.41	0.74	1.64	5.15	2.82	1.83	0.37	0.37
I	0.42	0.26	0.45	1.50	4.35	3.08	2.01	0.52	0.49
K	0.46	0.99	0.19	2.76	3.55	3.02	0.65	0.48	0.48
N	0.39	0.70	0.46	2.51	2.79	2.83	1.68	0.24	0.25
M	0.43	1.27	0.54	0.90	1.24	0.02	0.02	1.18	1.22
Mean of 14 cultivars	0.43	0.62	0.72	1.48	4.14	2.69	2.08	0.45	0.45
Standard error of the mean (56 d.f.)	0.07	0.44	0.69	0.94	2.50	2.00	1.93	0.49	0.49
Stored swedes									
I	0.62	1.53	0.83	3.49	2.76	10.30	1.92	0.89	0.86
H	0.56	1.50	0.53	3.72	3.20	9.02	1.85	1.01	0.96
A	0.60	1.55	0.82	3.25	3.23	7.69	1.99	0.85	0.87
K	0.49	1.87	0.25	4.59	1.40	6.01	0.67	1.12	1.06
B	0.49	0.80	0.46	3.58	3.95	4.70	2.05	0.75	0.75
M	0.60	1.63	1.09	2.93	6.04	4.01	2.93	1.05	1.05
N	0.44	3.45	1.09	5.56	3.30	3.47	1.52	1.74	1.71
Mean of 14 cultivars	0.53	1.55	0.78	3.32	3.20	5.75	1.77	1.04	1.02
Standard error of the mean	0.03	1.12	0.66	1.96	1.50	3.01	1.14	0.62	0.62

non-normality and the lack of homogeneity of variance could be traced to the effect of a small number of exceptional cultivars. It was thus decided to leave the data untransformed and accept that some standard errors may not be reliable.

To aid visual comparisons, the cultivars are ranked in Table 3 according to the most dominant components, peaks 5 and 3 for the freshly harvested and stored swedes, respectively. Certain prominent features of the data were obvious, such as the relative increase in peak 3 during storage, other features were still obscure until a canonical variate analysis was subsequently performed.

3.2.2. Multivariate analyses

Three separate multivariate analyses of variance were carried out on the data. The first incorporated data from both the fresh and the stored swedes, while the second and third were based on the results for either the fresh or the stored swedes, respectively. All of the analyses indicated that there were differences between cultivars, but a canonical variate analysis was required to locate the differing components. Table 4 shows the loadings and the percentage of the variance accounted for by the

Table 4. Loadings and percentage variance accounted for by first two canonical variates: loading associated with peaks (μg) and weight of swede (kg)

	i	di9	di1	di4	di2	di5	di3	di6	di7	di8	Variance accounted for (%)
Combined analysis	1	+2.36	+0.58	−0.45	+0.18	−0.33	+0.08	+0.00	+0.70	−0.65	36.6
	2	−5.21	+0.39	−0.34	+0.50	+0.02	−0.31	+0.42	+1.90	−1.55	22.6
Fresh analysis	1	−8.88	−2.04	+1.18	+0.40	+0.16	+0.54	−0.37	+9.07	−8.92	48.6
	2	+8.80	−1.31	+0.08	−0.17	+0.10	−0.21	+0.04	+0.95	−0.97	16.5
Stored analysis	1	−6.79	+0.12	−0.37	+0.39	−0.11	−0.20	+0.41	+2.07	−1.52	39.8
	2	+0.87	−0.29	+0.33	−0.04	+0.42	−0.15	+0.00	−1.44	+1.47	20.7

The loadings (di1 . . . di9) are as in equations (1) and (2).

Table 5. Contributions of each basic variate to first canonical variate scores (combined analysis)

Cultivar treatment		Wt	1	4	2	5	3	6	7	8
H	Fresh	+1.13	+0.22	−0.45	+0.21	−2.07	+0.19	0.01	+0.15	−0.15
	Stored	+1.32	+0.87	−0.24	+0.66	−1.06	+0.75	0.01	+0.70	−0.62
K	Fresh	+1.09	+0.59	−0.09	+0.49	−1.18	+0.25	0.00	+0.33	−0.31
	Stored	+1.16	+1.09	−0.11	+0.82	−0.47	+0.50	0.00	+0.79	−0.69
M	Fresh	+1.02	+0.74	−0.24	+0.16	−0.41	0.00	0.00	+0.82	−0.79
	Stored	+1.42	+0.95	−0.49	+0.52	−2.01	+0.34	0.01	+0.73	−0.68
Mean	Fresh	1.02	+0.36	−0.33	+0.26	−1.37	+0.22	0.01	+0.31	−0.29
	Stored	1.25	+0.90	−0.35	+0.59	−1.06	+0.48	0.01	+0.73	−0.66
Difference, stored-fresh		0.23	+0.54	−0.02	+0.33	+0.31	+0.26	0.00	+0.42	−0.37

first two canonical variates in each of the analyses. Table 5 was constructed using the loadings obtained from the first analysis as an example of the type of table which is helpful when interpreting canonical variate scores. Similar tables could be constructed to investigate the differences between particular cultivars in the fresh and stored analyses. Each element in the table is a product of a loading for a variate and the appropriate cultivar mean [equations (1) and (2)]. The sum of these products was adjusted to zero by equation (3) to calculate the cultivar means plotted relative to the first canonical axis in Figure 1. This axis discriminated so clearly between the fresh and stored

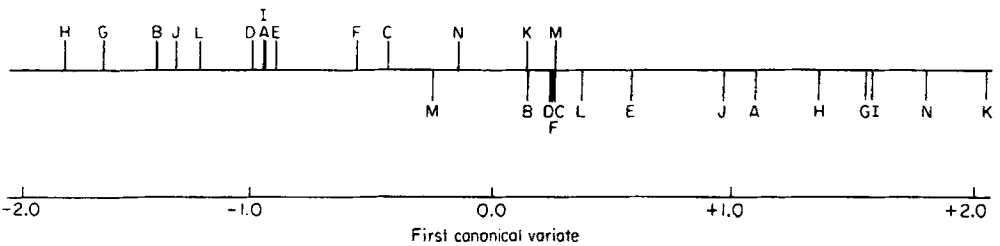

Figure 1. The distribution of cultivar means relative to the first canonical axis (fresh and stored swedes combined).

swedes that all further work was carried out on the two sets of data independently. The cultivar means are plotted relative to the canonical axes obtained from the second and third analyses in Figures 2 and 3.

3.3. Effect of storage

In general fresh swedes had negative scores and stored swedes had positive scores (Figure 1). The contribution of each basic variate to the first canonical variate score for three of the cultivars (Table 5), illustrated how a chemical interpretation can be put on the results. The final line of the table shows how the individual components of the scores altered after storage. All of the variates played a part in describing the changes which occurred on storage except peaks 4 and 6 which are both methylthioalkyl isothiocyanates.

The individual cultivar means revealed the differing behaviour of the cultivars on storage (Figure 1). Cultivar H changed markedly in relation to the other cultivars mainly due to a particularly large decrease in peak 5. Cultivar G (details not shown) changed similarly indicating that these two cultivars may not be particularly suitable for storage for insect-rearing purposes. Conversely cultivars F, C and M changed little on storage, cultivar M becoming more like the other cultivars when fresh (Table 3, Figure 1) due mainly to an atypical increase in peak 5 and a smaller than average increase in peaks 1 and 2 (Table 5). Cultivar K is included in Table 5 as being one which exhibits a typical change on storage.

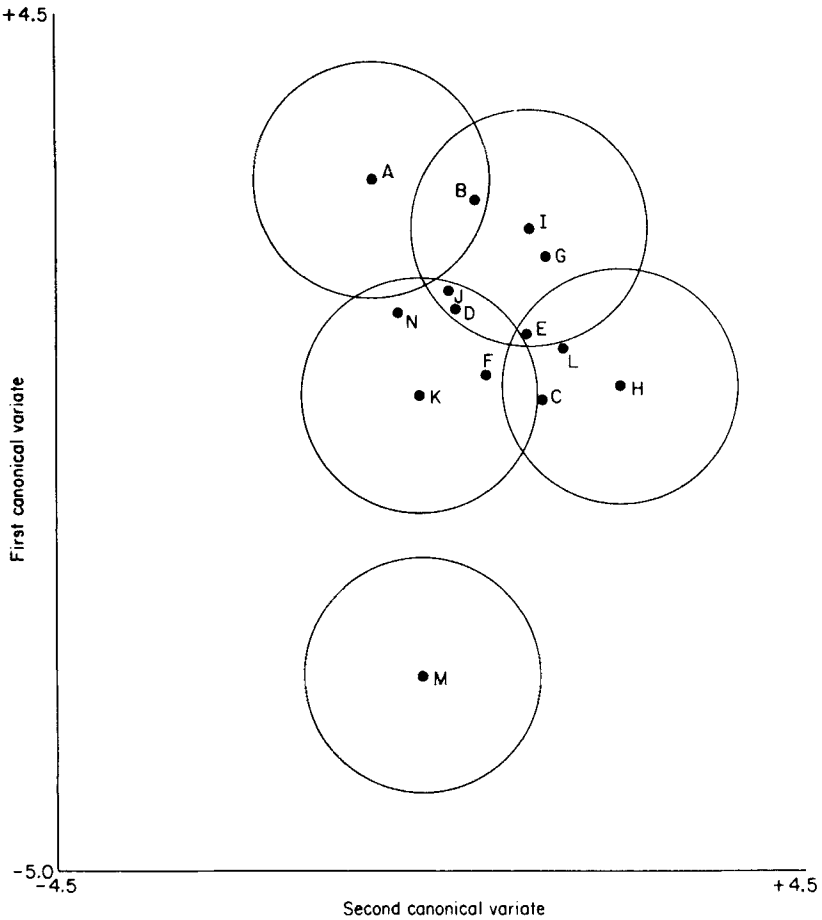

Figure 2. Plot of cultivar means and their confidence limits relative to the first two canonical axes (fresh swede).

3.4. Effect of cultivars

3.4.1. Fresh swede analysis

A plot of the means of the fresh swede data relative to the first two canonical axes (Figure 2) showed that the 95% confidence regions for groups A, H, I, K and M overlapped very little permitting satisfactory discrimination between these five cultivars on the basis of their volatile hydrolysis products. Canonical variate 1 reflected mainly the contributions of peaks 3 and 6, both hydrolysis products of the same glucosinolate precursor, 5-methylthiopentyl glucosinolate. Cultivar M takes an extremely negative value due to its lack of peaks 3 and 6. Variate 2 appeared to be mainly related to the size of the swedes. Even when weight was omitted from the analysis the order of the cultivars relative to the second axis is very similar to that of the weights.

3.4.2. Stored swede analysis

The 95% confidence regions for the stored swedes (Figure 3) distinguished satisfactorily between cultivars B, I, K, M and N. Canonical variate 1 for stored swede data was closely related to the weight of the swedes while canonical variate 2 was probably a measure of variation in peak 5.

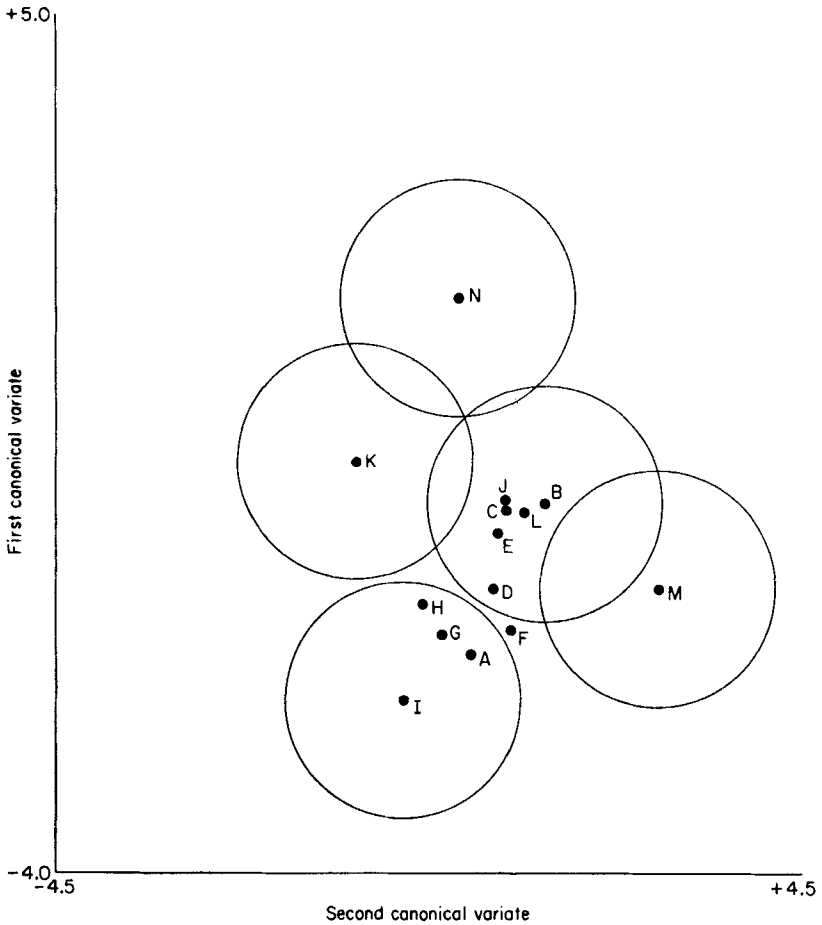

Figure 3. Plot of cultivar means and their confidence limits relative to the first two canonical axes (stored swede)

4. Discussion

The relationship between the canonical variates and the original peaks can be measured by calculating correlation coefficients. A more detailed interpretation in relation to meaningful chemical variation is dependent on the loadings. These indicate the magnitude of the contribution each chemical makes to the final distribution of cultivars and, as the sign of the loading can be either positive or negative, the effect of this contribution on the final distribution. A great advantage of canonical variate analysis over other multivariate techniques, like principal component analysis, is that the basic variates may be different types of measurements, since changing the scale of any one of them does not affect the scores. A disadvantage is that the loadings cannot be interpreted in a straightforward manner. Although all the peak heights were measured in terms of milligrams per swede the magnitudes of the heights were very different and the loadings are not directly comparable; small peaks such as 7 and 8 would always tend to have high loadings. Another complication in interpreting loadings arises when the basic variates are correlated. The loadings are such that two completely correlated variates make no more contribution to the analysis than one alone would have done; peaks 7 and 8 are extremely highly correlated because they are derived from the same parent glucosinolate and are isomeric forms of the same compound. It can be seen from Table 4 that their loadings effectively cancel each other out. An interesting feature of these analyses was the

pairing of the other peaks; the loadings for volatile compounds with the same parent glucosinolate almost always had opposite signs. This means that these peaks modified the effect of each other as might have been expected from the chemistry; the analyses picked out this feature although the interrelationship of the peaks was not obvious from the raw data.

Univariate analysis of variance was unable to differentiate between cultivars because of the interactions between volatile hydrolysis products and the number of compounds involved (Table 3). Canonical variate analysis enabled the composition of the cultivars to be distinguished and the cultivars to be grouped despite the variability and the small amount of replication of the basic data. The exploratory use of the technique has yielded useful information and further experiments using more replication and controlled growing conditions should provide more robust conclusions.

As is often the case with multivariate techniques the conclusions reached seemed obvious in retrospect. However, the analysis did pick out features of the data which the authors had not noticed previously and the two-dimensional plots provide excellent summaries of the data.

The technique was particularly useful here in showing the changes which took place on storage of the different cultivars. It may also be used to relate chemical data to the susceptibility of swede cultivars to insect pests. A detailed understanding of the interrelations of the chemical factors involved may allow initial screening of cultivars to be carried out chemically rather than by less easily regulated field experiments.

Acknowledgements

We thank Dr S. Finch for provision of the swede cultivars and Dr H. J. H. MacFie of the Meat Research Institute for helpful discussion on statistical techniques.

References

1. Thorsteinson, A. J. *Canad. J. Zool.* 1953, **31**, 52.
2. Coaker, T. H. *Proc. 5th Br. Insectic. Fungic. Conf. 1969* 1970, **3**, 704.
3. David, W. A. L.; Gardiner, B. O. C. *Entomol. Exp. Applic.* 1966, **9**, 247.
4. Finch, S. *Proc. 4th Int. Insect/Host Plant Symp: Entomol. Exp. Applic.* 1978, **24**, 350.
5. Van Etten, C. H.; Daxenbichler, M. E.; Williams, P. H.; Kwolek, W. F. *J. Agric. Food Chem.* 1976, **24**, 452.
6. Buttery, R. G.; Guadagni, D. G.; Ling, L. C.; Seiferl, R. M.; Lipton, W. *J. Agric. Food Chem.* 1976, **24**, 829.
7. MacFie, H. J. H.; Gutteridge, C. S.; Norris, J. R. *J. Gen. Microbiol.* 1977, **101**, 135.
8. Cole, R. A. *Phytochemistry* 1978, **17**, 1563.
9. Blackith, R. E.; Reyment, R. A. *Multivariate Morphometrics* Academic Press, London, 1971.
10. Marriott, R. H. C. *The Interpretation of Multiple Observations* Academic Press, London, 1974.
11. Seal, H. L. *Multivariate Statistical Analysis for Biologists* Wiley, London, 1966.
12. *Genstat: a general statistical program* Statistics Department, Rothamsted Experimental Station, Harpenden, Hertfordshire, 1977.