UNIFORMITY AND CONSTANCY OF WINE TASTERS EVALUATING THE SAME WINES ON TWO DIFFERENT OCCASIONS

M. B. LIMA *, A. P. BELCHIOR ** and G. F. ESTABROOK ***

* Estação Agronómica Nacional. 2780 OEIRAS. Portugal.
** Estação Vitivinícola Nacional. Dois Portos. 2575 RUNA. Portugal.
*** Herbarium and Biology Department. The University of Michigan. Ann Arbor, MI 48109 USA.

SUMMARY

Five wines were evaluated by a jury of ten winetasters on two occasions one month apart. On each occasion each winetaster evaluated the wines in two different ways: individually using the Bucaresti scores (contest, 1968); and by ranking the overall similarity of each of the ten pairs of wines from the most similar pair to the least similar pair. These data were then suject to two kinds of analyses: ordination by MDSCALE of the winetasters at the two times, to seek those who comprise the uniform core of the jury; and by calculating direct measures of constancy for each winetaster, namely correlations, distances, and statistical measures of the relationship between his evaluations at one time and his evaluations at the other. The ordinations reveal a core of three or four winetasters, with the others clustered around them. The direct measures of constancy revealed basically four types of winetasters: those whose opinions remained close and were highly correlated through time for both evaluation tasks; those whose opinions did not remain close but were highly correlated through time for both evaluation tasks; those whose opinions remained close and were highly correlated through time for the evaluation of individual wines, but ranked the overall similarity of pairs of wines very differently at different times; and finally, those whose opinions were neither close nor correlated for either of the evaluation tasks. Both tasks made different but valid contributions to the evaluation of wine, when performed with constancy. The three winetasters whose direct measures of constancy were best for both tasks occurred in the uniform core of the jury, as revealed by the ordinations.

INTRODUCTION

The reason for evaluating wine tasting and wine tasters is to confirm the accuracy and reproducibility that is desirable in any measuring procedure, and to minimize the internal and external influences to which the procedure might be subject.

Flanzy (1982) states that wine tasting can be characterized by three parameters: ability to preceive the tastes; accuracy; and constancy. Advocates of winetasting consider the human capacity for preceiving and distinguishing tastes to be very great. The constancy of human perception may be subject to greater variation.

Great winetasters generally achieve their recognition as a result of hard work, much experience, but rarely by comparing their opinions with more objective measures. Peynaud (1980) advocated that winetasters in the same jury all have extensive experience with the same method. This rarely happens in practice because many winetasting juries are improvised.

Amerine *et al.* (1969) were pioneers in analysis of taste and aroma. Recently, many attempts have been made to standardize terminology (see references in Noble *et al.*, 1987), and to classify winetasters and winetasting schools (Gervois and Sauvageot, 1984).

We seek objective criteria to define demarcated wine regions in Portugal, and to maintain the quality and reputation of Portuguese wines. For this we need to train competent winetasters, and to evaluate their competence objectively. Here we present methods for objectively measuring the uniformity of winetasting juries, and the constancy of winetasters, exemplified with data gathered from experiments in which the winetasters of the jury of the Estação Vitivinícola Nacional at Dois Portos, Portugal, generously and corageously participated.

MATERIALS, METHODS AND RESULTS

In January 1983, a group of professional winetasters at the Estação Vitivinícola Nacional, Dois Portos, Portugal, evaluated 5 wines. One month later, the same winetasters evaluated the same wines. By comparing the evaluations of the same wines by the same winetasters at two times one month apart, the constancy of the winetasters can be objectively measured.

To preserve anonymity, the winetasters are referred to by the letters A through J, followed by the numbers 1 or 2 to indicate their evaluation at the first or second evaluation date.

The wines evaluated were:

1. Adega Cooperativa de Torres Vedras, do ano (new);
2. Adega Cooperativa da Carvoeira, do ano (new);
3. Carvalho Ribeiro & Ferreira, garrafeira 1970 (13 yrs old);
4. Adega Cooperativa de Dois Portos, do ano (new);
5. Romeira, 1974 (9 yrs old).

Wines were presented anonymously, with fixed but arbitrary designations, and simultaneously to each taster, who made two distinct evaluations. First, using the standard Bucaresti form, the wines were scored from 1 = bad to 5 = excellent in each of five categories: color, clarity, aroma before tasting, sensations in the mouth, and aroma during tasting. These data are presented in Table I. Second, using a new format formerly unknown to these winetasters, each pair of wines was assessed subjectively for its overall similarity and all the possible ten pairs of wines ranked without ties from most similar (1) to least similar (10). These data are presented in Table II.

In January, winetasters I and J did not take part, and winetaster H completed only the Bucaresti form but did not rank the similarities of pairs. In February, all ten winetasters performed both tasks.

The data were analysed for two basic purposes:

1. To discover the group of winetasters who comprise a unified school of opinion, for each of the two evaluation methods; and
2. To evaluate the constancy over the time period of one month of each of the winetasters.

To meet the first purpose, two dimensional multidimensional scaling (2D-MDS) was used (Kruskal, 1964a, b). Using a computer, Euclidean distances between each pair of the 18 evaluations based on Bucaresti forms (Table I) were calculated as the square root of the sum of the squares of the 25 differences

between each score at one time and its corresponding score at the other time, and the 18 evaluations then arranged in two dimensional space so that the distance between pairs in that space is maximally correlated with their Euclidean distances. To speed up the calculations, which are done in a systematic trial-and-error manner, spatial rearrangements started from an

TABLE I

Bucaresti scores for five wines, five categories, and two time periods

Pontuações da ficha de Bucareste dos cinco vinhos, nas cinco características, em dois períodos

Wine		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	F2	G1	G2	H1	H2	I2	J2
Color	1	3	3	5	4	4	4	4	4	4	4	4	4	3	3	4	4	3	3
	2	3	3	5	4	4	4	4	4	4	4	4	4	3	3	4	4	3	3
	3	4	5	5	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	4	2	2	3	2	3	3	2	4	3	3	3	3	3	2	4	4	3	3
	5	3	4	5	4	4	4	4	5	4	4	4	4	3	3	4	4	4	3
Clarity	1	4	4	4	4	4	4	4	3	5	5	4	4	4	4	4	4	3	3
	2	4	4	5	4	4	4	4	3	5	5	4	4	4	4	4	4	3	3
	3	4	4	5	5	4	4	4	3	5	5	4	4	4	4	4	4	3	3
	4	3	4	2	2	3	2	2	3	2	3	3	4	3	4	4	4	3	2
	5	4	4	5	4	4	4	4	4	4	4	3	4	4	4	4	4	3	3
Odour	1	3	3	3	3	3	2	3	1	4	3	3	4	3	2	2	3	2	2
	2	3	3	4	3	3	3	3	3	3	3	3	3	2	2	3	3	2	1
	3	4	4	5	4	4	4	4	4	5	4	4	2	4	4	4	3	5	4
	4	2	2	1	2	2	2	1	2	3	2	3	3	2	2	2	3	1	1
	5	3	4	4	3	3	3	2	4	3	4	2	4	4	3	2	3	4	3
Taste	1	3	3	3	3	3	2	2	1	3	2	4	4	2	2	3	3	1	2
	2	3	3	4	3	3	3	2	2	3	3	4	3	4	3	3	3	2	2
	3	4	4	4	3	4	4	3	4	4	5	4	3	4	4	4	4	4	3
	4	2	2	1	2	2	2	1	2	3	2	3	3	3	3	2	2	1	2
	5	4	4	3	2	4	3	3	3	4	3	3	4	4	4	2	3	4	3
Odour during Taste	1	3	3	3	2	3	3	2	1	3	2	3	3	3	2	3	3	2	2
	2	3	3	3	3	3	3	2	2	3	3	3	2	3	3	3	3	2	1
	3	4	4	4	3	4	4	3	3	4	4	4	3	3	4	4	4	4	4
	4	2	2	1	2	2	2	1	2	2	2	2	2	3	2	2	2	2	1
	5	3	4	3	2	3	3	3	3	3	4	3	3	4	3	2	3	4	4

TABLE II

Rankings of the ten pairs of the five wines by winetasters at two different times one month apart

Ordenação dos dez pares dos cinco vinhos pelos provadores em duas datas separadas por um mês

Wine tasters	Pairs of wines									
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
A1	3	5	7	1	6	8	2	10	4	9
A2	1	5	8	3	6	7	4	10	2	9
B1	2	5	7	1	4	9	3	10	6	8
B2	1	8	5	3	7	6	2	10	9	4
C1	2	6	7	4	5	8	3	10	1	9
C2	5	6	7	4	3	8	1	10	2	9
D1	2	6	10	4	5	9	3	7	1	8
D2	5	8	2	7	4	6	3	10	1	9
E1	3	4	8	2	7	5	6	10	1	9
E2	2	7	6	4	5	8	3	10	1	9
F1	4	1	9	2	6	7	5	10	3	8
F2	2	6	9	1	4	7	3	8	5	10
G1	7	4	5	8	2	6	3	9	1	10
G2	2	9	4	7	8	3	5	10	1	6
H2	4	3	9	1	6	8	2	10	5	7
I2	3	7	2	6	8	4	5	10	1	9
J2	3	4	5	2	9	6	7	10	1	8

initial 10 dimensional arrangement based on Principal Components Analysis (Sneath and Sokal, 1971; Cabral *et al.*, 1977; Curvelo-Garcia *et al.*, 1988). The two dimensional arrangement determined by 2D-MDS is presented in Figure 1. It gives an idea of the overall distance relationships among the 18 evaluations. The 17 evaluation (Table II) were arranged in two dimensional space using 2D-MDS, but in this case Spearman rank correlations were used to represent similarities. The initial arrangement was based on Principal Coordinate Analysis (Gower, 1966, 1967; Ferraz and Lima, 1982). This ordination is presented in Figure 2.

The second purpose, to evaluate constancy of winetasters, was met by calculating for each winetaster six measures of constancy between his January evaluations and his February evaluations. Two measures were based on Bucaresti forms:

1. Euclidean distance in 25 dimensional space between January scores and February scores; and

2. Pearson correlation over 25 pairs of scores, the first from January and the second from February.
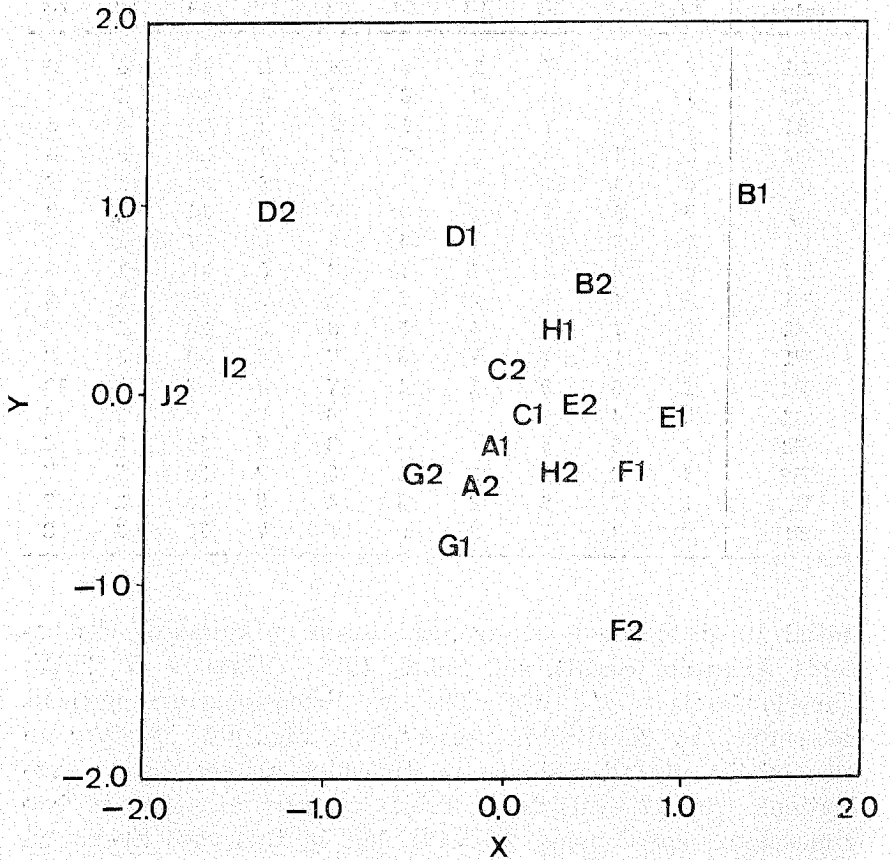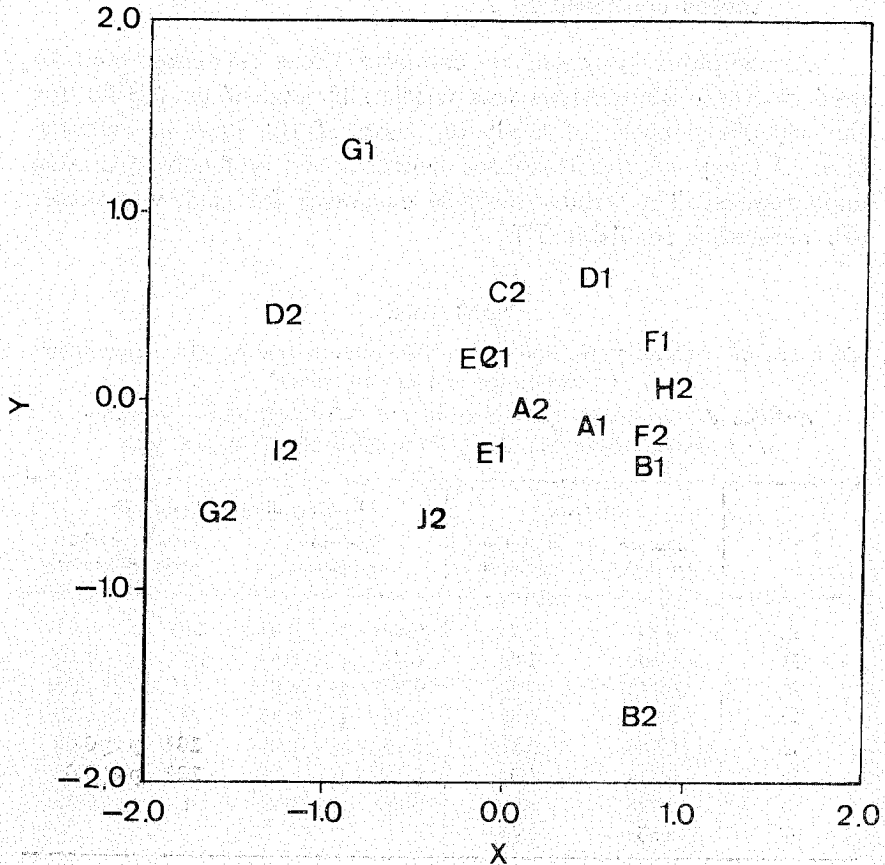

Fig. 1 — Two dimensional arrangement of the ten winetasters at two different times one month apart determined by multidimensional scaling (MDS), based on Euclidean distances between each pair of the 18 evaluations using Bucaresti forms.

Distribuição bidimensional dos dez provadores, em duas datas separadas de um mês, determinada por escalonamento multidimensional, baseada nas distâncias euclidianas entre cada par das 18 avaliações utilizando a ficha de Bucareste.

Four measures were based on rankings of resemblances of pairs of wines:

1. Spearman rank correlation of January rankings with February rankings;
2. Significance of this Spearman rank correlation under the null hypothesis that resemblances of these pairs of

Fig. 2 — Two dimensional arrangement of the ten winetasters at two different times one month apart determined by multidimensional scaling (MDS), based on Spearman rank correlations between each pair of the 17 evaluations.

Distribuição bidimensional dos dez provadores, em duas datas separadas de um mês, determinada por escalonamento multidimensional, baseada no coeficiente de correlação de Spearman entre cada par das 17 avaliações.

wines were ranked at random and independently in the two time periods, determined by computer simulation of 5000 pairs of random rankings;

3. The sum of the absolute values of the 25 differences between each January score and its corresponding February score; and

4. Significance of the measure of 3. under the null hypotheses described in 2.

A computer program to compute these measures and to perform these simulations was written by one of us (GFE) for the specific purpose of analysing data of the kind in example here. A copy of the PASCAL source code is freely available upon request. The values of these measures for each winetaster are persented in Table III.

TABLE III

Measures of constancy for individual winetasters tasting the same wines at two times one month apart

Medidas de constância para provadores individualizados, provando os mesmos vinhos em duas datas separadas por um mês

Taster	Bucaresti		Rankings of similarities of pairs			
	Pearson correlation	Euclidean distance	Spearman correlation		Sum of absolute rank differences	
A	0.866	2.24	0.89*	$p < 0.01$	10*	$p < 0.01$
B	0.850	4.24	0.62*	$p < 0.03$	22	$p < 0.10$
C	0.889	2.00	0.89*	$p < 0.01$	8*	$p < 0.001$
D	0.598	4.80	0.36	$p < 0.15$	24	$p < 0.10$
E	0.756	3.32	0.76*	$p < 0.01$	16*	$p < 0.02$
F	0.274	4.00	0.72*	$p < 0.02$	16*	$p < 0.02$
G	0.728	3.00	0.28	$p < 0.25$	28	$p < 0.25$
H	0.852	2.45				

DISCUSSION

The ordination of the winetasters based on their Bucaresti scores, as shown in Figure 1, shows a close group of winetasters with others further away. C, A, E, H, and possibly G have both their evaluations in a close central group. F, and possibly B, have one but not both evaluations near the central group. Some

of the evaluations of I, J, D, B, and F, are clearly at variance with this central group. Although Figure 1 represents a combination of many factors, some general trends are discernable. Opinions in the upper left judged the wines to be of lower overall quality, and opinions to the lower right, and including, the line through B1 and A2 judged the wines to be of higher overall quality. Opinions about the odour and taste of wine 4, considered by all winetasters to have the lowest overall quality of the five, were somewhat higher in the central core than on the periphery.

The ordination of the winetasters based on rankings of overall similarity of pairs of wine, as shown in Figure 2, also reveals a central core of winetasters with others further away. However, there are some very significant differences, reflecting the differing capacities of the winetasters. Both evaluations of winetasters A, C, E, and F together with one but not both evaluations of B and D might be considered closely grouped. (Recall that H and J did not take part in the first evaluation). For this ranking data it is dificult to make explicit explanations of the arrangements. However, the group to the upper left considered the wine pair 1-5 to be quit different, while the group on the right side of the central core considered these two wines to be the most similar of all. Striking is the exit of G from the central core in Figure 1, based on individual evaluations, and the entrance of F into the central core of Figure 2, based on rankings of overall similarity.

Direct measures of the constancy of winetasters do not show so readily the central group of winetasters, but they allow for a direct measure of the constancy of each wine taster individually, apart from the similarity of his opinion from that of others. As can been seen in Table III, for Bucaresti scores all the winetasters had positive correlation between their scores in January and their scores in February. A high positive correlation means that the relative rankings of wines remained about the same even though their scores at the two dates may be quite different. Correlations above 0.800, as in A, B, C, and H, evidence very consistent relative scores. Correlations below 0.500, as in F, evidence less consistent scores between the two dates.

The Euclidean distance between the scores at the two time periods decreases as consistency increases. A distance of 0 would

evidence identical scores, and a distance of 5.00 would occur if each score differed by 1 between the two dates. Distances of 3.00 and below as in winetasters A, C, G, and H, evidence very constant scores. Winetasters with large distance but high correlations, such as B, experienced a major change of scale between the two time periods, but expressed nearly the same relative opinions of the wines. Winetaster F, with a low correlation and a large distance changed his opinion between the two tasting dates in a more erratic way.

Ranking the similarities of the ten pairs of wines is a different kind of task, requiring subjective synthesis and some ability to consider many sensations at once. It is also a task that winetasters have not usually practiced. Wine tasters A, B, C, E, and F had significantly correlated rankings between the two tasting dates. Since the winetasters were obliged to rank all ten pairs of wines without ties, the posibilities for changes of scale from one date to the next are slight (unlike for the Bucaresti scores where scale changes can occur readily). The sum of the absolute differences in rank is somewhat like Euclidean distance for the Bucaresti scores in that it is a direct measure of the absolute constancy between the two time periods.

Winetaster D continued to be inconsistent. For wine taster B, as with the Bucaresti scores, the correlation is high but the distance between the ranks at the two times is great. B may represent a class of winetasters for whom the relationships among wines is more accessible than absolute individual evaluations.

Winetasters F and G represent interesting situations. F, who was quite inconsistent at scoring wines individually on Bucaresti forms, was significantly consistant at ranking wines for overall similarity. By contrast, G, who was reasonably consistent at scoring wines individually, worked as if at random when ranking overall similarity.

As before, winetasters A, C, and E continued to show high correlations and close absolute distances, with C the most consistent in both kinds of tasks. Their evaluations for both tasting dates fell in the coherent cluster in the ordination for both types of tasks. They would seem to constitute an important part of the core of winetasters at Dois Portos.

ACKNOWLEDGEMENTS

RESUMO

Uniformidade e constância de apreciação de provadores dos mesmos vinhos em duas ocasiões diferentes

Uma câmara de dez provadores procedeu à prova dos mesmos cinco vinhos, em duas sessões, com um mês de intervalo. Em cada sessão, cada provador avaliou os vinhos com dois sistemas diferentes: individualmente, usando uma ficha de prova (concurso de Bucareste, 1968); e ordenando a semelhança global entre cada um dos dez pares de vinhos, do par mais semelhante para o par menos semelhante. Os dados assim obtidos foram submetidos a dois tipos de análise: ordenação, por escalonamento multidimensional (MDSCALE), dos provadores nas duas sessões, para encontrar os que formam o núcleo central uniforme; e pelo cálculo directo de medidas de constância para cada provador, nomeadamente correlações, distâncias e outras medidas estatísticas, da relação das suas apreciações numa e noutra sessão.

As ordenações revelaram um núcleo de três ou quatro provadores envolvido pelos restantes. As medições directas de constância revelaram basicamente quatro tipos de provadores: aqueles cujas opiniões se mantiveram próximas e altamente correlacionadas em ambas as sessões nas duas datas diferentes; aqueles cujas opiniões não se mantiveram próximas mas foram altamente correlacionadas nas duas datas para ambas as sessões de prova; aqueles cujas opiniões se mantiveram próximas e altamente correlacionadas, em ambas as datas, na avaliação dos vinhos considerados individualmente, mas ordenaram as semelhanças globais dos pares de vinhos de modo muito diferente nas duas sessões de prova separadas por

um mês; e finalmente aqueles cujas opiniões não foram nem próximas nem correlacionadas em ambas as sessões de prova.

Ambos os critérios de avaliação conduziram a contribuições diferentes, mas válidas, para a avaliação de vinhos, sempre que executados com constância. Os três provadores que manifestaram as melhores medidas directas de constância para os dois tipos de prova encontram-se localizados no núcleo central da câmara de provadores, de acordo com o evidenciado pelas ordenações obtidas pelo método de escalonamento multidimensional (MDSCALE).

RÉSUMÉ

Uniformité et constance d'apreciation de dégustation des mêmes vins en deux occasions differentes

Un jury, de dix membres, a dégoûté les mesmes cinq vins, en deux séances separés d'un mois. En chaque séance chaque dégustateur à évalué les vins avec deux différents systèmes: par utilisation d'une fiche (concours de Bucaresti, 1968); et ordonnant les pairs, de les dix pairs possibles de vins, par son critère de ressemblance, du pair le plus semblable au moins pareil. Ces donnés ont été soumis a deux types d'analyse: rangement des dégustateurs dans les deux séances par le cadrage multidimensionel (MDSCALE), pour trouver ceux qui forment le noyau central du jury; et par le calcul direct de la constance pour chaque dégustateur, nommement, corrélation, distances et autres mesures statistiques, de la relation entre son critère dans l'une et l'autre séances. Les rangements ont révélé un noyau de trois ou quatre dégustateurs avec l'entourage des autres. Les mesures directes de constance, on révélé basiquement quatre types de dégustateurs: ceux qui mantiennent ses avis hautement corrélationés pendant tout le temps dans les deux séances; ceux qui ne mantiennent pas ses avis mais sont hautement corrélationés dans les deux séances; ceux qui mantiennent ses avis et sont hautement corrélationés dans les deux séances pour l'évaluation de chaque vin individuellement, mais rangent les similitudes des pairs de vins de façon trés différent entre les deux séances; et par fin, ceux pour lequels leurs avis ne sont pas congruents ni corrélationés pour quelqu'une des évaluations.

Les deux systèmes d'évaluation des vins on des contributions différents mais valides pour l'expertise du vin, quand ils sont effectués avec constance. Les trois dégustateurs qui on manifesté les meilleurs mesures directes de constance pour les deux syestèms ils sont dans le noyau central du jury, d'accord avec ce qui a été mis en evidence par les rangements, obtenus par la méthode de cadrage multidimensionel (MDSCALE).

LITERATURE CITED

Amerine, M. A.; E. B. Roessler and F. Filipello
  1959 Modern sensory methods of evaluating wine. *Hilgardia*, 28: 447-567.
Cabral, J. M. P.; A. H. Carvalho and M. B. Lima
  1977 Aplicação de métodos de taxonomia numérica na classificação de
        águas minerais de Portugal Continental. *Comunic. Serv. Geol. Port.*,
        61: 343-363.
Ferraz, J. F. P. and M. B. Lima
  1982 Utilização de métodos de taxonomia numérica no estudo das rela-
        ções fenéticas de isolamentos de Colletotrichum e de Gloeosporium.
        Garcia de Orta, 9: 197-212.
Flanzy, C.
  1982 La degustation: util de la recherche? *Rev. Française d'Oenologie*,
        88: 69-75.
Gervais, P. and F. Sauvageot
  1984 Un example de batteria d'epreuves permettant d'evaluer de denrees
        alimentaires. Lebnsm. — Wiss. u. — *Technol.*, 17: 69-73.
Gower, J. C.
  1966 Some distance properties of latent root vector methods used in
        multivariate analysis. *Biometrika*, 53: 325-338.
  1967 Multivariate analysis and multidimensional geometry. *Statistician*,
        17: 13-28.
Kruskal, J. B
  1964a Multidimensional scaling by optimizing goodness of fit to a nonme-
        tric hypothesis. *Psychometrica*, 29: 1-27.
  1964b Nonmetric multidimensional scaling: a numerical method. *Psycho-
        metrica* 29: 115-129.
Noble, A. C.; R. A. Arnold; J. Buechsenstein; E. J. Leach; J. O. Schmidt
and P. M. Stern
  1987 Modification of a standardized system of wine aroma terminology.
        Am. J. Enol. Vitic., 38: 143-146.
Peynaud, E.
1980 Le gout du vin. Dunod, Paris.
Sneath, P. H. A. and R. R. Sokal
  1971 Numerical taxonomy — The principles and practice of numerical
        classification. San W. A. Freeman, 573 pp.