Permutation tests: Are there differences in product liking?

Michael Meyners

Fachbereich Statistik, University of Dortmund, D-44221 Dortmund, Germany

E-mail: meyners@statistik.uni-dortmund.de

Fax: +49 (0)231 755 3454

Foreword

This paper was a contribution to a workshop on the analysis of sensory data held at the 5th Sensometrics Meeting 2000 in Columbia, Missouri. In a sensory study 28 grape/raspberry beverages were evaluated in a series of consumer tests to better understand the grape/raspberry beverage product category. The tests were conducted in two geographies, within each geography two sets of test with different sets of attributes and scales were considered. In each test, the evaluation of the 28 products was conducted in 4 days. For the first test additional classification information was collected. Furthermore a trained sensory panel evaluated the products for 52 attributes. The questions considered within the workshop were the following:

- Are the consumers uniform in their liking and disliking of products?
- Do the consumers in the two geographies behave similarly?
- Do the consumers in the two tests behave similarly?
- What are the relationships between the sensory attributes and liking?
- What are the relationships between the sensory attributes and consumer attribute ratings?

In this paper we confine ourselves to the first three questions given here.

Abstract

The target of our considerations is whether or not we can find significant differences between subgroups of consumers with respect to given hedonic variables. For this purpose a STATIS-consensus is computed for each group and the dissimilarities between groups are judged with the help of the RV-coefficient. Since the distribution of this coefficient is unknown and we do not make any assumptions on the distribution of the error terms, a permutation test is performed. This provides a simple possibility to test for significance of the dissimilarities in question. Some pre-treatment of the data is necessary to perform this statistical test, afterwards subgroups according to the two sets of consumer tests, the different geographies and some of the classification variables within consumer test 1 are considered. Whenever we find significant dissimilarities a graphical representation of the respective consensuses is provided to interpret the differences.

Introduction

A main interest in sensory analysis is to optimise the products according to consumer preferences. Thus an important question is whether or not subgroups of consumers differ from each other at all with respect to the given hedonic variables. If so, it might be necessary to develop different variations of a product for these subgroups. Hence we confine ourselves to a consideration of the hedonic variables given within the data, assuming that these variables represent the preferences of the assessors for the different products. The data set contains the judgements of 692 assessors on 28 grape/raspberry beverages on several variables. Two consumer tests with different variables were considered, attended by 342 and 350 assessors, respectively. Furthermore two different geographies have been chosen in which these tests took place. Finally for the assessors of the first test also classification variables have been collected which account e.g. for age and education.

The first question is whether or not there are differences in product liking (as it is represented by the hedonic variables) between the two sets of consumer tests as well as between the given geographies. Further consideration is given to the consumers of the first test according to the provided classification information. Several ways to divide these consumers into two subgroups have been carried out. Each time we consider the question whether or not the dissimilarities between the two subsets are significant. The division has been carried out according to the variables *number in household*, *age*, *education*, *martial status* and *employment*. Note that we consider the respective statistical tests only in a descriptive manner, since otherwise we would have to take a multiple test problem into account and adjust the level for each test.

A model to describe sensory profiling data has been given by Meyners, Kunert and Qannari (2000). Even though this model is rather simple, it is hard to find an analytical statistical test for dissimilarities between subgroups of consumers as well as products, even if we assume a normal distribution of the random errors. Thus it seems reasonable to consider permutation tests to examine these dissimilarities. A permutation test for product differences within a sensory framework has already been given by Wakeling, Raats and MacFie (1992). Within this work we confine ourselves to the dissimilarities between the consumers.

Pre-treatment of the data

In what follows we consider the RV-coefficient as well as the STATIS-method (cf. Schlich, 1996). Both demand complete data sets, so we had to pre-treat the data. In a first step we substituted each missing value by the overall mean for this variable. There are a lot of different possibilities how to substitute these, e.g. by zeros or by an individual mean (if not all values are missing). Anyway, since the number of missing values was quite small compared to the total number of observations, it should not heavily influence the results. Thus we are not going into details here.

In a second step it appeared that there are some errors in the data set. We found some assessors for which two observations were given for one of the samples, whereas another sample was missing for those persons. This presumably dues to mistakes in the input of the data sets. Since we were not able to reconstruct the true assignment we reassigned the observations in the following manner: We looked for the sample number missing and compared this number with the one for which two observations appeared. The observation that appears first in the data set was then assigned to the sample with the smaller number. Again this should not heavily influence the results, since this case occurred for only 4 assessors, namely those with *ID number* 2119, 2192, 3111 and 5149.

STATIS method – RV-coefficient – Permutation tests

The STATIS method is used to calculate an overall consensus of the hedonic variables from different assessors. Other methods exist, e.g. GPA (cf. Dijksterhuis, 1996), which is not useful for our purpose since it takes too much time to calculate the consensus for this large data set to be feasible for the permutation test.

Roughly spoken, the consensus calculated with the help of STATIS is a weighted mean of all association matrices $X_iX_i^T$, where X_i is the mean centred data matrix of assessor i. The weights are determined in such a way that an assessor who agrees well with the other ones gets a larger weight than one who does not agree. For details, see e.g. Schlich (1996).

To calculate the weights the so-called RV-coefficient is used which provides a measure of similarity between two matrices X and Y. It is calculated according to

$$\mathrm{RV}(X, Y) = \frac{tr(XX^TYY^T)}{\sqrt{tr(XX^TXX^T)\; tr(YY^TYY^T)}}.$$

It can be shown that it is equivalent to Pearson’s correlation coefficient after rearranging the association matrices $X_iX_i^T$ into vectors. Thus a large value of RV indicates high similarity, whereas a small value indicates larger differences. The RV-coefficient will be used as the test statistic for the permutation test. To formalise this somehow, in what follows we develop the idea of the permutation test for this particular case.

We assume two subgroups of consumers A and B, say. In the considered application these groups are subgroups due to the classification variables, the different geographies or the different consumer tests. We would like to show that there are non-negligible dissimilarities between the groups. Thus our null hypothesis claims that there are no differences. In a first step we need a measure of similarity, which is derived as follows: For each group we calculate the STATIS consensus of the respective assessors which gives matrices C_{A} and C_{B} for group A and B, respectively. Afterwards the RV-coefficient for these matrices is computed. A large value indicates similarity, whereas a small one indicates larger differences. Since the distribution of the RV-coefficient is unknown we do not know a critical value to statistically test for similarity. Anyway, if the null hypothesis is true, it should make no difference whether a particular person belongs to group A or B. Thus the observed value of the test statistic should not be significantly smaller than an usual value for the other possible assignments to the groups (in this framework a significantly large value does not have any useful meaning, see also the end of the following section). Therefore we randomly permute the assignment to the subgroups, in which it is necessary to retain the structure of the original data, i.e. in this case to derive subgroups of the same size as given in the data set. For this re-assignment, we recalculate the RV-coefficient. This is carried out $n = 100$ -times and we compare the RV-coefficient of the original assignment with those of the permutations. With it, the number of smaller values within permutations is counted. If we find only few smaller values we conclude that the observed small value could not be due to chance, i.e. we claim significant dissimilarities between groups A and B. Considering e.g. a 5%-level, we state significance whenever we do not observe more than 5 smaller values, whereas in case of a 1%-level we state significance if we observe not more than one smaller value.

Note that we do not consider explicitly the p-value of the tests. The size of the corresponding confidence interval in general decreases with an increasing number of permutations, hence 100 is a rather small number. Nevertheless, independent of the number of permutations the proposed procedure provides a level-α-test. Of course the power of the test is strongly influenced by the total number of permutations, i.e. with a small number of permutations the probability to detect existing differences is smaller than with a large number. Thus for the interpretation of the results we can state significant differences whenever we observe a sufficient small number of RV-coefficients that are smaller than the observed value. Otherwise, if we found too many smaller values to claim significance, we can not conclude that there are no differences since the power of the test with 100 permutations is rather small.

Comparison to the approach of Kazi-Aoual *et al.* (1995)

It seems reasonable to emphasise the difference between our method and the approach proposed by Kazi-Aoual, Hitier, Sabatier and Lebreton (1995). Consider two data matrices on the same k products. We might permute the rows of the matrices, i.e. we re-label the products at random. Kazi-Aoual *et al.* (1995) derive the exact values for the first three moments of the distribution of the corresponding RV-coefficients if all possible $k!$ permutations are considered. Schlich (1996) proposes to standardise the RV-coefficient by subtracting the mean and dividing by the standard deviation. Assuming the resulting value to be normal distributed, he states a significant similarity between the data matrices if the observed value is too large. Note that he implicitly assumes that there are apparent differences between the products, otherwise both matrices would contain only random numbers. Even if the subgroups which derive the matrices usually judge similar, in this case we will not obtain a significant result. However, this approach is reasonable to investigate on similarity of the matrices.

On the opposite, the method is not applicable to investigate on dissimilarity. In this case, there are different possibilities. The first one is that the assessors perceive the products different or use different sensory dimensions to describe the product characteristics. If the judgements are independent, within the permutations we expect the RV-coefficient neither to increase nor to decrease in general, i.e. we expect the observed value to be on average and in particular not to be significant small. The second possibility is that at least one matrix contains just random numbers, since the respective assessor(s) did not perceive any differences. Then the columns of the matrices are independent, and, as it was described in the first case, a permutation should not generally enlarge the RV-coefficient.

On the opposite, we propose to permute the assignments of the assessors to the subgroups. Hence we get totally changed association matrices, whereas in the approach of Kazi-Aoual *et al.* (1995) and Schlich (1996) we just permute their rows.

Consider there are differences between the subgroups with respect to the sensory dimensions used. Assigning the assessors at random to the subgroups, we usually have some assessors from each original subgroup in each new subgroup. Then the calculated consensus reflects to some extent the sensory dimensions of both subgroups. Hence we expect a larger RV-coefficient than in the original data. In case the assessors of one subgroup do not perceive any differences at all, the consensuses after reassigning the assessors reflect the dimensions of the other subgroup and therefore we expect an enlarged value of RV again. Only if the assessors behave similar, i.e. either they perceive the differences the same way or none of them perceives any differences at all, we expect the observed RV-coefficient to be on average.

The results of Kazi-Aoual *et al.* (1995) are not valid for our permutation procedure, which among other things can be seen directly from the number of permutations that is given by $\binom{M}{m_1}$, where m_1 is the number of assessors in the first (or equivalently second) subgroup and M is the total number of assessors. In general this is unequal to $k!$, in the case considered here, this is much larger than $k!$.

Results

1. Different consumer tests

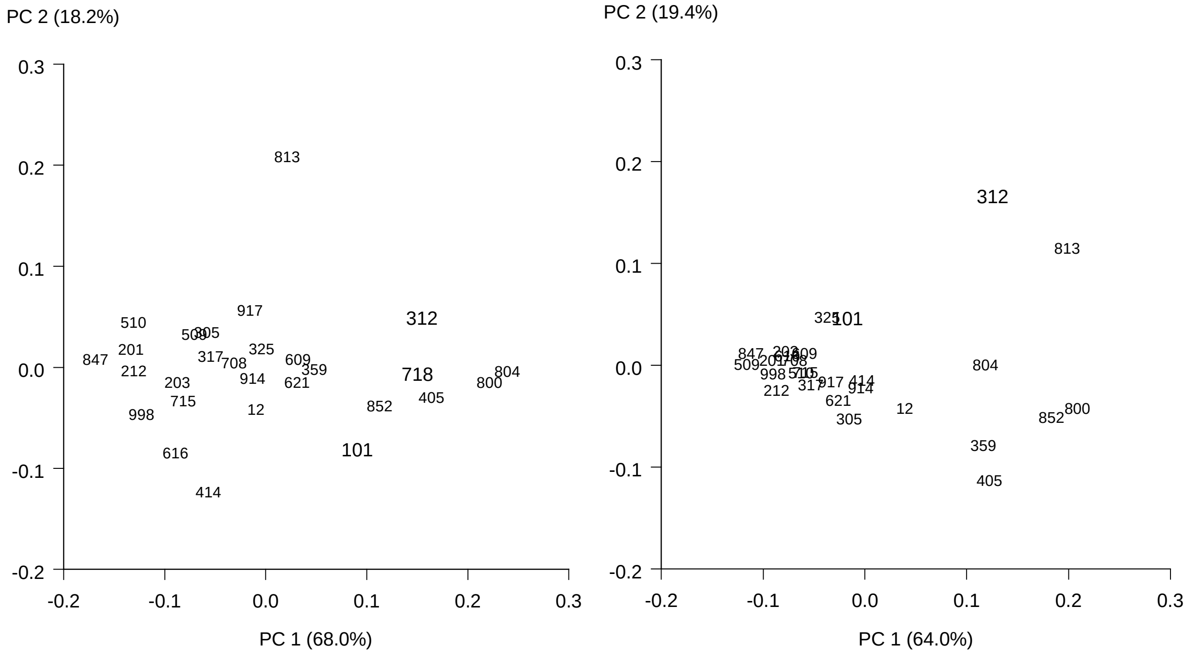
The first dissimilarities considered were those between the two consumer tests. Notice that we had more variables in the first test than in the second one, precisely we had 3 hedonic variables in the second test while there were 7 in the first one. From the latter we neglected *overall aftertaste*, since aftertaste was not considered at all in consumer test 2 and this might be an additional dimension that accounts itself for significant dissimilarities. Since also different numbers of attributes and with it different numbers of dimensions might influence the results, we decided to confine ourselves to the first three principal components of each assessor. For those of consumer test 2 this makes no difference at all, while in test 1 the dimension is reduced onto three main directions.

The data contained 342 and 350 assessors for consumer test 1 and 2, respectively. Thus we calculated the STATIS-consensus for both groups and from these consensuses the RV-coefficient. After that we carried out the permutations to derive the critical values. The results are represented in table 1.

ordered value no.	RV-coefficient
1	**0.8903**
2	0.8989
3	0.8998
4	0.9005
5	0.9026
6	0.9053
⋮	⋮
99	0.9715
100	0.9720
observed	0.6605

Table 1: Results (extract) of the permutation test for differences between the two consumer tests.

It can be seen that there is no permutation at all that gave a smaller RV-coefficient than the one observed with the original assignment. In fact, the difference to the smallest value found in the permutations is quite large and thus we state a highly significant dissimilarity between the hedonic assessments of the two consumer tests. To clarify what causes the differences we give the two-dimensional representation of each STATIS-consensus in figures 1 and 2.



Figures 1 and 2: First principal components of the STATIS-consensus for the consumers of tests 1 and 2. Note that in figure 2 sample 718 is omitted since it coincides with sample 101.

It can be seen that the groups mainly differ in their judgement of products 101, 312 and 718. While the assessors in consumer test 1 perceive 312 – at least in the second principal component – similar to the mean of the products, those in test 2 find larger differences from the other ones. In this consumer test it is judged similar to 813. The products 101 and 718 are situated in the big cluster of products for test 2, whereas they tend to be more similar to 405 or 852 in test 1. Related to a three dimensional consensus the contribution of the dimensions to the total variance is quite similar for the groups.

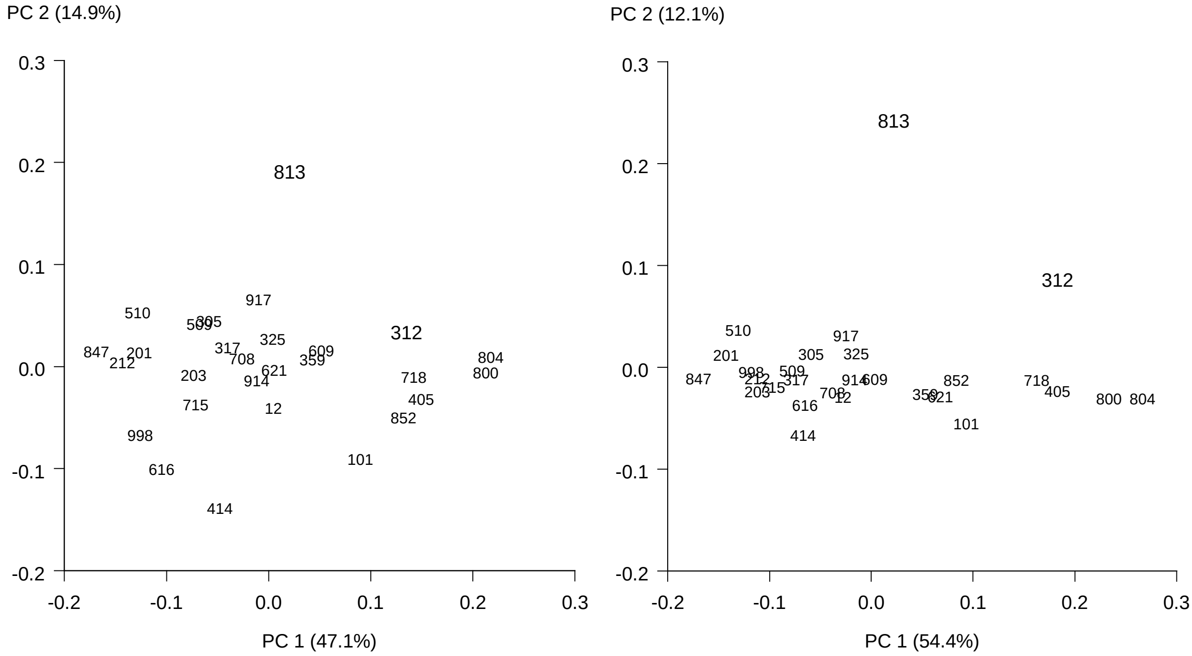
2. Different geographies

The geographies are considered separately for the different consumer tests, so we could use more dimensions for test 1 as we did in the last section. The procedure was the same as described before, and the results are given in table 2.

	RV-coefficients for	
ordered value no.	consumer test 1	consumer test 2
1	0.9047	**0.8179**
2	0.9074	0.8340
3	**0.9076**	0.8355
4	**0.9101**	0.8383
5	0.9111	0.8405
6	0.9112	0.8413
⋮	⋮	⋮
99	0.9420	0.9198
100	0.9440	0.9266
observed	0.9091	0.7783

Table 2: Results (extract) of the permutation tests for differences between the two geographies for both sets of consumer tests.

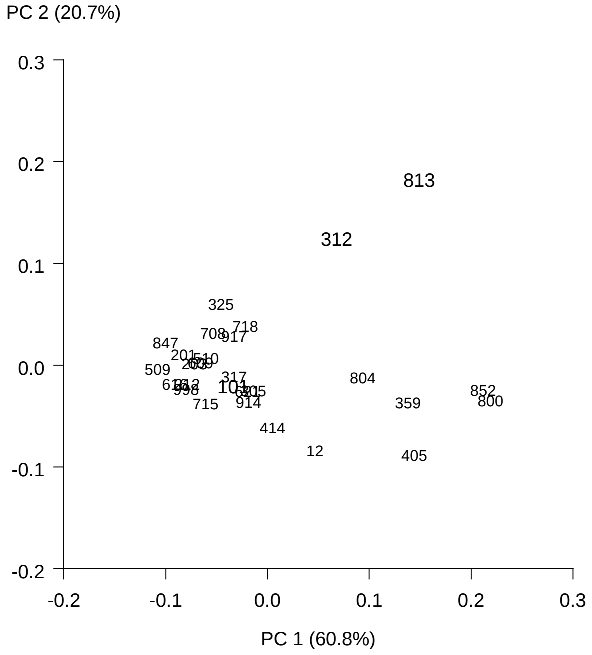
From the results we can also state dissimilarities according to the hedonic variables between the two geographies. For consumer test 1 we observed 3 permutations with a smaller value of RV, anyway, to a 5%-level we could also claim differences. For test 2 we did not observe any smaller value than the one that has been derived from the original data, thus we also state significant dissimilarities between the two geographies.



Figures 3 and 4: First principal components of the STATIS-consensus for the consumers of test 1 with geographies 1 and 2.

From figures 3 and 4 which represent the consensuses for the different geographies within consumer test 1 it seems that these dissimilarities are not quite large. Anyway, this is right with respect to the first dimension, in which we can only see minor changes. But in the second dimension we find a general difference. Despite the little difference in the assessment of product 312, we find the consumers in geography 2 giving just negligible differences in this dimension despite for product 813. Due to this product, the variation explained by this dimension is approximately as large as for geography 1. Hence we have, despite the products 813 and with concessions 312, a rather one-dimensional consensus for geography 2, while there are larger dissimilarities in the second dimension for geography 1.

Figures 5 and 6 represent the respective consensuses for the second consumer test. Here again we find all products being judged similar, despite differences for products 101, 312 and 813. These products are presumed to account for the significant dissimilarities between the geographies within test 2.



Figures 5 and 6: First principal components of the STATIS-consensus for the consumers of test 2 with geographies 1 and 2.

3. *Different classification*

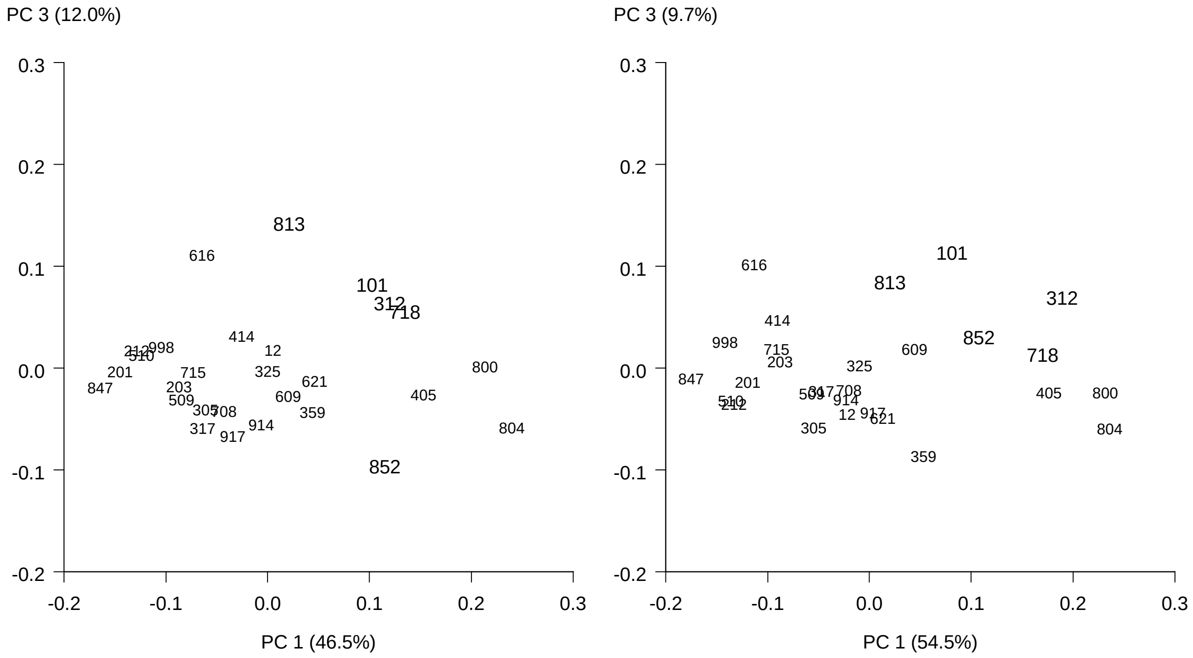
For the consumers in test 1 classification variables were provided. We used these variables to consider product dissimilarities between subgroups with respect to the hedonic variables. The following classification variables have been considered: *number in household*, *age*, *education*, *martial status* and *employment*. The *number of children* has been neglected, since it should give similar results as the *number in household*. Most of these variables had more than two levels while we could only compare two subsets at a time, so we divided the assessors in two groups for each variable. The variable *number in household* was divided into ‘smaller households’ (1-3 persons, 44.1% of the assessors) and ‘larger households’ (more then 3 persons). According to the variable *age* the assessors have been divided into ‘younger’ (younger than 45, 60.1%) and ‘older’, while for *education* we assigned ‘less educated’ to those who did not at least complete college (40.7%) and ‘more educated’ for those who did so. Within *martial status* we had no need to reassign values, since here only two values occurred, namely ‘married’ (34.9%) or ‘not married’. For the variable *employment* we found no arguments to decide whether ‘employed half time’ was closer to ‘unemployed’ (33.9%) or to ‘employed’ (49.0%). Since only 17.1% of the assessors belonged to that group, we neglected these and confined ourselves to the other ones.

We do not state the following tests as some kind of real statistical testing, since then we would have to respect for the multiple test problem. We rather use these tests in a descriptive manner to look for possibly conspicuous results. The results of the permutation tests are given in table 3, from which it can be found that there seem to be dissimilarities only for the classification according to the variable *age*.

	RV-coefficients for subgroups according to variable				
ordered value no.	*number in household*	*age*	*education*	*marital status*	*employment*
1	0.9006	0.8963	0.8892	0.8852	0.8878
2	0.9079	0.9068	0.8940	0.8882	0.8931
3	0.9117	0.9068	0.9110	0.8890	0.8936
4	0.9119	0.9081	0.9122	0.9041	0.8952
5	0.9124	**0.9108**	0.9138	0.9065	0.8969
6	0.9126	**0.9127**	0.9143	0.9068	0.8974
⋮	⋮	⋮	⋮	⋮	⋮
56	0.9324	0.9287	0.9319	0.9272	**0.9197**
57	0.9324	0.9290	0.9324	0.9273	**0.9199**
⋮	⋮	⋮	⋮	⋮	⋮
86	0.9397	0.9374	0.9414	**0.9367**	0.9275
87	**0.9400**	0.9378	0.9416	**0.9372**	0.9276
88	**0.9406**	0.9380	0.9419	0.9381	0.9281
⋮	⋮	⋮	⋮	⋮	⋮
94	0.9441	0.9393	0.9458	0.9404	0.9324
95	0.9444	0.9399	0.9458	0.9417	0.9326
96	0.9448	0.9418	**0.9464**	0.9419	0.9331
97	0.9464	0.9422	**0.9481**	0.9425	0.9353
98	0.9470	0.9443	0.9497	0.9428	0.9359
99	0.9479	0.9446	0.9502	0.9445	0.9362
100	0.9483	0.9450	0.9515	0.9482	0.9366
observed	0.9405	0.9126	0.9480	0.9371	0.9198

Table 3: Results (extract) of the permutation tests for differences between the subgroups of the assessors within test 1. The subgroups have been assigned according to the given classification variables.

Plotting the first two principal components of the STATIS-consensuses of both subgroups we were not able to see any interesting differences. Thus we decided to plot the first and the third principal component in figures 7 and 8, from which it can be seen that there are some dissimilarities for products 101, 312, 718, 813 and 852. Even though they seem not to be very large, it suffices for a rather small RV-coefficient. We do not state significance for this descriptive result, but it might be a topic for some future research to reconsider differences according to the age of the assessors. Note that this would require additional data to avoid running into a multiple test problem and misinterpretations.



Figures 7 and 8: First and third principal component of the STATIS-consensus for the consumers of test 1 with *age* smaller than 45 respectively greater or equal to 45. The second PC accounts for 14.4% respectively 13.4% of the variation.

Anyway, there is another classification that is a little bit conspicuous. For the variable *education* a very high RV-coefficient has been observed. We had just 4 permutations that gave a larger value, hence we might conclude that we have 'significant similarity'. We do not state this since it does not seem to have any useful meaning in this framework, furthermore this is something that we expect to happen in 1 out of 20 cases if there are in fact no differences. Hence we advise against an overinterpretation of this value.

Conclusion

We considered dissimilarities within several hedonic variables between different subsets of assessors which evaluated 28 grape/raspberry beverages. Fur this purpose a permutation test based on the RV-coefficient has been carried out to test for these differences. Dissimilarities have been found between the two different consumer tests as well as between the geographies within each of these tests. These dissimilarities are mainly due to three products, namely 101, 312 and 813, which was stated from a graphical representation of the respective STATIS-consensuses. Furthermore several classification variables have been used to divide the assessors of consumer test 1 into subgroups. To avoid a multiple test problem, these tests were only considered in a descriptive manner. From this analysis it seems that the age of the consumers might be a classification variable that provides dissimilarities, which might be the topic of some future research on similar or identical (but new!) data. For the other variables no differences could be found.

Another topic of interest is whether or not the significant result of the permutation test for consumer tests 1 and 2 might be due to different dimensionalities between the assessments within these tests. It also seems sensible to relate the results of our analysis to the descriptive data of the sensory panel on several attributes for the products. This would help to interpret the differences and to judge which attributes cause the differences between the consumers' assessments. From these considerations suggestions might be derived on how to develop a product for some specific subgroup.

Finally note that even though a large RV-coefficient points to similarity between the matrices, this does not necessarily rule out systematic differences between the respective subgroups. In our analysis we observed values of about 0.9 which nevertheless were found to be significant small. The absolute value of the RV-coefficient depends on several circumstances like e. g. number of products and attributes or the differences between the samples. Thus a permutation test is always sensible to examine whether or not the subsets are similar.

Acknowledgement

The author is grateful to the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivariate data structures") for the financial support of this work.

References

- Dijksterhuis, G. (1996). Procrustes Analysis in Sensory Research. In: T. Næs and E. Risvik, *Multivariate Analysis of Data in Sensory Science* (pp. 185-219).
- Kazi-Aoual, F., Hitier, S., Sabatier, R. and Lebreton, J.-D. (1995). Refined Approximations to Permutation Tests for Multivariate Inference. *Computational Statistics and Data Analysis 20,* 643-656.
- Meyners, M., Kunert, J. and Qannari, E. M. (2000). Comparing Generalized Procrustes Analysis and STATIS. *Food Quality and Preference 11,* 77-83.
- Schlich, P. (1996). Defining and Validating Assessor Compromises about Product Distances and Attribute Correlations. In: T. Næs and E. Risvik, *Multivariate Analysis of Data in Sensory Science* (pp. 259-306).
- Wakeling, I. N., Raats, M. M. and MacFie, H. J. H. (1992). A New Significance Test for Consensus in Generalized Procrustes Analysis. *Journal of Sensory Studies 7,* 91-96.