László Sipos[1], Márta Ladányi[2], Viktor Losó[3], Zoltán Kókai[1], Attila Gere[1]

Received: 2016. July – Accepted: 2016. September

Monitoring the agreement of sensory panelists

Keywords:
consensus, sum of ranking differences (SRD) method, Page test,
Cabilio-Peng multiple comparisons procedure

1. Summary

It is an important theoretical question of sensometrics, as well as a practical issue of accredited laboratories, how to monitor and analyze the development of the agreement/consensus of the sensory evaluation group (panel), regarding the series of measurements in a given period. For the evaluation of the consensus of the individual judgements, the sum of ranking differences (SRD) method is a good evaluation alternative. The difficulty in analyzing rank sum patterns lies in the fact that the pairwise significant difference method for rank sum data was developed only a few years ago. In the present work, a combination of methods is recommended for tracing the consensus of sensory panelists.

2. Introduction and literature review

When monitoring performance, what is tested is whether the sensory judge or the sensory evaluation group (panel) is capable of recognizing, identifying and measuring the given property, as well as using and interpreting it the same way as other judges or evaluation groups (panels) **[1]**. Testing and improvement of trained and expert panelists, and the measurement of their performance can typically be realized through a standardized, multi-stage system that is based on feedback and realized under standardized conditions, preferably with some kind of software support.

Performance evaluation methods of sensory tests are classified in the literature in several ways. The most common classification is based on the number of people performing the tests (a single panel member, a full panel or panels), but it can also be based on the mathematical method used (single or multivariate statistical methods). Sensory assessors are classified into three categories by the literature according to their training: naive / consumer panelists, trained panelists, expert panelists. The application of panelists with different training is necessary for different types of tasks, and performance evaluation methods are usually applied to tests performed by trained or expert panelists **[2], [3], [4].**

The solution of the quality management systems of testing laboratories accredited for sensory testing is the application of preventive and corrective actions, the task of which is the identification, prevention and elimination of non-conformities, i.e., of panelists evaluating differently from the group average. To do so, it is advisable to analyze the averages and medians of raw data on a box-and-whisker plot and, furthermore, to perform the most important descriptive statistics (minimum, maximum, standard deviation, kurtosis, skewness, range), in order to obtain a picture about the data structure and outliers. Graphical representation of the data can save time, and it also offers an effective way to investigate and evaluate the performance of the sensory panel.

Monitoring the performance characteristics of the individual panelists is a key factor in the quality of the results of the sensory panel, because without it the result can be incorrect or unreliable. Obviously, examination of the sensory panel or comparison of several panels can only take place the evaluation at the individual level. The performance of the individual panelists is usually characterized by three classic in-

[1] Szent István University, Faculty of Food Science, Department of Postharvest Science and Sensory Evaluation
[2] Szent István University, Faculty of Horticultural Sciences, Department of Mathematics and Informatics
[3] MIRELITE MIRSA Zrt.

dicators: the discrimination ability of the individual panelists, the repeating ability of the individual panelist, agreement of the individual panelist with the rest of the panel **[1].**

The **discrimination ability** in fact means an ability to make a distinction between the products, therefore, the ability to separate and/or distinguish is often used for characterization. The reasons for inadequate ability to separate might be due to the application of not suitable panelist selection methods, sensory fatigue, or inadequate sensory memory or concentration. The **repeating ability** of the panelist means that the same product is tested - typically on the next day – at the same time under identical conditions, and the results are compared to those of the previous day. The repeating ability of the judge is inversely proportional to the repeated results of testing the same samples, or to the standard deviation of the results of replicate samples. The repeating ability is significantly influenced by mental conditions, health status or the lack of motivation. **Agreement**/consensus is the ability of different panelists or evaluation groups (panels), based on which similar scores are assigned to samples of the same products. This means that a given panelist agrees with other members of the panel regarding the sensory property, within a given tolerance (small difference between the judges) **[5].**

According to standard *„MSZ ISO 11132 Sensory analysis. Methodology. Guidelines for monitoring the performance of a quantitative sensory panel"*, the evaluation group is not in agreement/consensus if one or more panelists do not agree with the rest of the panel. This can be concluded if a significantly different value is given by one of the panelists (*Cusum* analysis); if the standard deviation of the residues of a panelist is significantly higher than that of the panel, or if the correlation coefficient between the scores of the panelists and the average for the panel is very low or negative, the regression curve of the scores of the panelists differ significantly from 1, compared to the panel averages, and/or the intercept differs significantly from zero **[5], [6].** Often times, calculations using these methods are hard or relatively slow, therefore, it is necessary to examine the application possibilities of other methods or method combinations.

Numerous methods have been developed for the evaluation of individual judges and panels, and the manner of application is usually determined by the goal to be achieved and the available software. It was emphasized by Meullenet et al. **[7]** that univariate measurements are useful, however, a more comprehensive picture can be obtained by the multivariate analysis of descriptive data. The advantage of multivariate methods is that they are capable of simple graphic presentation of complex data sets.

3. Objective

To evaluate the consensus of the different panelists, the *sum of ranking differences (SRD)* method is a proven alternative [8,9,10,11]. According to the SRD method, the performance of the individual judge is compared to the average performance of the evaluation group. A clear hierarchy is determined by the SRD method, from the panelist closest to the consensus, agreement or average of the evaluation group (the best) to the one farthest from it, based on which a selection can be performed or individual development recommendations can be made.

Tracing the consensus of panelists, or combined analysis of the evaluation of several different products has not been solved so far, using this method. One of the reasons for this is that the method was only developed and programmed a few years ago. The principle of the sum of ranking differences (SRD) method was laid down by Héberger **[12]**, while its validation and software implementation was carried out by Héberger and Kollár-Hunek **[13]**. The other difficulty in analyzing rank sum patterns lies in the fact that the pairwise significant difference method for rank sum data was developed only a few years ago **[14]**. In the present work, the objective was to trace the agreement/consensus of sensory panelists using a combination of methods.

4. Materials and methods

Data was generated with the help of a simulation: random rearrangement of 1, 2, 3....12 discrete SRD values of 20 evaluations of 12 judges in a for cycle (R-project 2.15.2) **[15]**. The part of the program script is shown below:

```
x<-c(1:12)
res<-matrix(nrow=20, ncol=12)

for (i in 1:20) {
rank<-sample(x, 12, replace=FALSE)
res[i,]<-rank
}
t<-t(res)
print(t)
```

The steps proposed for pattern analysis are the followings. Simulated rank numbers are summed row by row for each panelist and, based on these, the panelists are ranked again. In the table, the panelist that can be characterized by the lowest rank sum will be first, because his results will be closest to the panel consensus. The consensus of the panel will be reduced according to the values of the panelists who are characterized by higher rank sum values.

The sequence produced above is analyzed as a trend using the Page test. If the trend appears to be significant, multiple comparisons of the judges in pairs are performed by a specially developed Cabilio-Peng (Normal) method **[14]**. The Page test and the multiple comparisons in pairs were conducted using the XL-Stat software **[16]**.

5. Results

The method combination proposed for pattern analysis is illustrated by the following example. The starting matrix of the 1, 2, 3....12 discrete SRD values of 20 evaluations of 12 panelists was created by the simulation of random rankings. Following this, rank numbers were summed row by row for each panelist and, based on these, the panelists were ranked again, as shown in ***Table 1***.

The hypothetical ranking produced above was analyzed as a trend using a Page test. By the composite analysis of evaluations 1 to 20, the ranking could be confirmed. The calculated *p*-value was lower than the pre-determined type I error (α=0.05), and so the null hypothesis was rejected (H_0: identical treatments), and the alternative hypothesis was accepted (H_1: different treatments). The risk of rejecting the null hypothesis when it is true, is very low (<0.08%.) Results of the Page test are shown in ***Table 2***.

Multiple pairwise comparison of the panelists in pairs were performed by a specially developed Cabilio-Peng (Normal) method **[14]**. Based on the results of the multiple pairwise comparisons, panelists could be classified into 3 homogeneous groups (A, B, C). Overall results of the 20 simulated evaluations showed that panelist b8, who could be characterized by the lowest rank sum, was closest to the panel consensus. The consensus of the panel is reduced, according to their values, by panelists who are characterized by rank sums higher than that of panelist b8, and this is shown in ***Table 3***.

The best result was achieved by panelits b8, whose performance was the most typical to the panel performance. His result was significantly better (α=0.05) than those of panelists b11, b2, b10, b12, b7, b4 and b6. In terms of panel consensus, the second best performer was b3, differing significantly from b7, b4 and b6 (α=0.05). There was no significant difference between the other panelists, as shown by the pairwise comparison matrix in ***Table 3***.

Proposed steps of the new method combination and their software solutions:

1. Consensus analysis of the individual evaluations using the SRD software with ties (http://aki.ttk.mta.hu/srd/).
2. Ranking of the panelists based on the SRD values (Excel).
3. Overall rank sums have to be determined again after each new evaluation, and then the panelists have to be ranked again (Excel).
4. Examination of the assumed order using the Page test (XL-Stat).
5. If the assumed order is significant, then calculation of the pairwise significant differences using the Cabilio-Peng method (XL-Stat).

The advantage of the method is that tracing of the panel consensus can be achieved easily, since all one has to do is attach new results to the previous results (as a new column), and then perform the running of the program on the combined results of evaluations 1-2, 1-2-3, 1-2-3-4, 1-2-3-4-5, ... It is important that rank sums have to be determined again for each panelists after each new evaluation, and panelists have to be ranked again according to this, before this ranking is analyzed by the Page test. Considering its principle, rather the existence of an assumed trend is investigated by the Page test, and so it does not deal with pairwise comparisons. The result if this procedure will be positive if there is an identifiable trend when looking at the complete data set. Ranges within the set that behave differently are not taken into account with considerable weight **[17].**

6. Conclusions

Sensometrics and the tracing of the consensus of sensory panelists are important practical questions of accredited laboratories. In this connection, it is of key importance to monitor and analyze the development of the agreement/consensus of the sensory evaluation group (panel), regarding the series of measurements in a given period **[18], [19].** In summary, it can be stated that tracing and analyzing panel consensus can be achieved successfully by a combination of the Page test and the Cabilio-Peng method on the rank pattern of SRD values. With its help, it can be answered, based on the results of several completely different tests, which panelists performed best from a panel consensus point of view, and which panelists differ from each other significantly or not significantly. Tracing panel consensus can be achieved effectively by overall analysis of the individual evaluation results.

7. Acknowledgement

Supported by the ÚNKP-16-4 New National Excellence Program of the Ministry of Human Capacities.

8. References

[1] Lawless, H. T., Heymann, H. (2010): Sensory Evaluation of Food, 2st ed. Chapman and Hall, New York, NY. 243-246.

[2] Molnár, P. (1991): Élelmiszerek érzékszervi vizsgálata. Budapest: Akadémiai Kiadó, 11–204.

[3] Kókai, Z. (2003): Az almafajták érzékszervi bírálata. Doktori értekezés. Budapest: Budapesti Közgazdaságtudományi és Államigazgatási Egyetem, 35-59.

[4] MSZ ISO 6658:2007 Érzékszervi vizsgálat. Módszertan. Általános útmutató.

[5] MSZ ISO 11132:2013 Érzékszervi vizsgálatok. Módszertan. Általános irányelvek a leíró vizsgálatot végző bírálóbizottság teljesítményének mérése

[6] Naes, T., Brockhoff, P. B. Tomic, O. (2010): Statistics for sensory and consumer science. Wiley, Chicester. 1-287.

[7] Meullenet, J-F., Xiong, R., Findlay, C. F. (2007): Multivariate and Probabilistic Analyses of Sensory Science Problems. Wiley-Blackwell, New York, NY. pp. 27–47.

[8] Sipos, L., Kovács, Z., Szöllösi, D., Kókai, Z., Dalmadi, I., Fekete, A. (2011): Comparison of novel sensory panel performance evaluation techniques with e-nose analysis integration. Journal of Chemometrics, 25:(5) pp. 275-286.

[9] Sipos, L., Gere, A., Szabó, A., Kovács, S, Kókai, Z. (2013): Multivariate Methods For Assessing Sensory Panel Performance. In: Héberger K (szerk.) Programme and Book of Abstracts of CC 2013 - Conferentia Chemometrica. Konferencia helye, ideje: Sopron, Magyarország, 2013.09.08-2013.09.11. Budapest: Hungarian Chemical Society, p. 6. (ISBN:978-963-9970-38-0)

[10] Kollár-Hunek, Klára; Héberger, Károly (2013a): Method and model comparison by sum of ranking differences in cases of repeated observations (ties) Chemometrics and Intelligent Laboratory Systems, 127, 15 2013, 139-146.

[11] Kollár-Hunek, Klára; Héberger, Károly (2014): Erratum to “Method and model comparison by sum of ranking differences in cases of repeated observations (ties)” Chemometrics and Intelligent Laboratory Systems, 132, 15, 179-180.

[12] Héberger, K. (2010): Sum of ranking differences compares methods or models fairly. Trend. Anal. Chem. 29, 101–109.

[13] Héberger, K., Kollár-Hunek, K. (2011): Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. Journal of Chemometrics, 25, (4) 151–158.

[14] Cabilio, P. Peng, J. (2008): Multiple rank-based testing for ordered alternatives with incomplete data. Statistics and Probability Letters, 78, 2609-2613.

[15] R-project 2.15.2, The R Foundation for Statistical Computing, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Augasse 2-6, 1090 Vienna, Austria.

[16] XL-Stat 2012.6.2 manual. (Addinsoft, 28 West 27th Street, Suite 503, New York, NY 10001, USA)

[17] Sipos, L. (2009): Ásványvízfogyasztási szokások elemzése és ásványvizek érzékszervi vizsgálata. PhD értekezés. Budapesti Corvinus Egyetem. Döntéstámogató Rendszerek Doktori Iskola. 96-101, 179-184.

[18] EA-4/09 (European co-operation for Accreditation, Accreditation for Sensory Testing Laboratories)

[19] MSZ EN ISO/IEC 17025:2005 Vizsgáló- és kalibrálólaboratóriumok felkészültségének általános követelményei (ISO/IEC 17025:2005)