J.Natn.Sci.Foundation Sri Lanka 2018 46 (1): 95 - 102
DOI: http://dx.doi.org/10.4038/jnsfsr.v46i1.8269

RESEARCH ARTICLE

Use of Mandel's bundle of lines model to improve agreement of a panel of tea tasters

T. Ursla S. Peiris[1*], Chamila K. Walgampaya[2], David L. Banks[3], R.O. Thattil[4] and I. Sarath B. Abeysinghe[5]

[1] *Biostatistics Unit, Faculty of Livestock Fisheries and Nutrition, Wayamba University of Sri Lanka, Makandura, Gonawila.*
[2] *Department of Engineering Mathematics, Faculty of Engineering, University of Peradeniya, Peradeniya.*
[3] *Department of Statistical Science, Duke University, Durham, North Carolina, USA.*
[4] *Department of Crop Science, Faculty of Agriculture, University of Peradeniya, Peradeniya.*
[5] *Tea Research Institute of Sri Lanka, Talawakelle.*

Revised: 10 October 2017; Accepted: 16 November 2017

Abstract: Sensory evaluation is of prime importance in the tea industry. The price of tea is predominantly determined by the tea taster's judgement. Hence, statisticians and sensory analysts have significant interest in improving the reliability of sensory panel evaluations. This study suggests a statistical de-biasing approach to improve the agreement of panels of tea tasters by using the Mandel's bundle of lines model. This paper analyses data from a panel of eight experts who evaluated 72 black tea samples collected from 13 different geographical regions of Sri Lanka over a one year (January 2010 – December 2010) period at monthly intervals. The results showed that in only 20 instances out of 72, the intra-class correlation coefficients were greater than 0.5, but after bias correction, 64 instances out of 72 have intra-class correlations greater than 0.5. Therefore, the Mandel's bundle of lines approach improves consensus among the experts by enforcing better statistical calibration.

Keywords: Agreement, Mandel's bundle of lines model, reliability, sensory evaluation, tasting panel.

INTRODUCTION

Sensory evaluation is a scientific method used to evoke, measure, analyse and interpret responses to products as perceived through the senses of sight, smell, touch, taste and hearing (Stone & Sidel, 1993). The importance of sensory analysis is vital in many domains, especially in quality assurance and product development in the food and beverage industry. In product development, sensory evaluations are of prime importance to identify desirable attributes, to explore specific characteristics of an ingredient, and to compare similarities/ differences in a range of products. Moreover, sensory evaluation plays a significant role in grading such products as tea according to international or national norms, or to assure the product is manufactured correctly. In addition, sensory evaluations act as a way to reduce risk and uncertainty in decision making.

Sensory evaluation is of critical importance in the tea industry, since it is the primary factor in determining the price of the product. As a result, tea tasters play a key role, and the industry must be able to rely upon their assessments. Tea tasters judge the quality of tea through three sensory methods: eye (sight), tongue (taste) and nose (smell). The quality parameters perceived by the eye and the tongue are collectively called plain black tea quality parameters and correspond to brightness, briskness, thickness and the colour of infusions. These attributes are due to catechines, theaflavins, thearubigines, residual chlorophylls and the caffeine present in black tea. The flavour of black tea is determined by the nose, and is due to volatile flavoury compounds (Pal *et al.*, 1998).

* Corresponding author (turslasp@gmail.com; https://orcid.org/0000-0002-2365-2541)

High performance levels by descriptive panels of assessors, and the quality of the data they provide, are essential for proper research and business decisions in the tea industry. A good sensory panel should provide results that are accurate, discriminating and precise (Kermit & Lengard, 2005). Hence, statisticians and sensory analysts are concerned about the reliability of the sensory panel. In blind studies, it has been observed that the tasters' scores differ significantly from taster to taster, even for the same sample of tea. This variation in sensory characteristics from the same batch can significantly distort industrial production, management decision making, and mislead scientific studies. Within the field of descriptive sensory analysis, it is a well-known fact that assessors give uneven results due to differences in motivation, sensitivity, and psychological response behaviours (Lundahl & McDaniel 1999, as cited in Kermit & Lengard, 2005; Alvarez & Blanco, 2000). Individuals within a panel and people from different cultures may have different thresholds of perception, and the product experience of the panel may lead to differences in the ability to discriminate among samples (McEwan *et al*., 2002).

Therefore, the reliability of this approach has been questioned by the scientists all over the world (Pal *et al*., 1998; Alvarez & Blanco, 2000; Vaamonde *et al*., 2000; McEwan *et al*., 2002; Latreille *et al*., 2006; Hodgson, 2008). Different methodologies based on descriptive statistics, analysis of variance, general linear models and mixed models have been previously proposed to study or improve the performance of a trained panel of sensory evaluators (Latreille *et al*., 2006). Due to its versatility, analysis of variance (ANOVA) has been one of the most frequently employed statistical tools to study differences between products. This standard univariate method is also used to separate the total variation of sensory data into sources that affect specific sensory responses (Kermit & Lengard, 2005).

Currently, there is no viable alternative to expert panels. Biochemical analysis of tea is expensive and the benchmarks needed for such evaluation are ultimately based on human judgment from expert panels. Therefore, there is a need for better methodology to establish the reliability and validity of sensory methods (Meiselman, 1993).

There have been several researches on replacing sensory analysis by instrumental methods like the electronic nose. According to a study done by Benedetti *et al*. (2004), the electronic nose was able to correctly classify all of the samples in their study, whereas the sensory panel was correct in only 50 % of the cases. Shen *et al*. (2001) claims that the electronic nose is able to measure changes in volatile compounds associated with oil oxidation and could be used to supplement data obtained from sensory evaluation panels. Compared with sensory tests, the electronic nose methodology is simple, rapid, and the literature suggests that it can be useful in discriminative testing. Some studies indicate that the electronic nose instrumentation can be used as a complementary discriminatory tool in quality control for food packaging (Deventer & Mallikarjunan, 2002). However, although instrumental assessments of other sensory properties (sight, taste) are becoming possible, it is unlikely that such devices will replace human assessments in tea industry, because sensory quality is a result of the interaction between the tea and the human sensation produced by certain stimuli from the tea cup (Alvarez & Blanco, 2000). In the tea industry, there are no physical and chemical methods that can replace sensory assessments performed by the panel of tasters.

Therefore, statistical methods are needed to improve the reliability and accuracy of sensory panel performances. It is challenging because there are no fixed reference standards to which the tasters can compare the teas. The reliability of tasting panels can be assessed using two indicators: an agreement indicator describing the consensus of ratings for the same product, and a reliability indicator for the discrimination ability of a panel (Vaamonde *et al*., 2000; Bi, 2003, as cited in Latreille *et al*., 2006). It is therefore indispensable to have quality control for the sensory data prior to the determination of the sensory quality of the food stuffs to ensure the reliability of the results of the analysis (Alvarez & Blanco, 2000).

The objective of the present study was to improve the performance, i.e., the agreement of a panel of tea tasters. The method for achieving this improvement is to correct for individual bias using the Mandel's bundle of lines model.

METHODOLOGY

The panel

Initially, a panel of 10 members was selected for the study. All panellists were senior members serving in brokering companies, exporting companies, and the Sri Lanka Tea Board, and had extensive experience in the field of sensory evaluation. At the beginning, panellists were briefed individually about the nature of the samples and the purpose of the research. A pilot study was carried out by sending the same sample four times at 2 wk

intervals to assess the consistency and the reliability of the selected panellists. All evaluations were done as blind evaluations; i.e., all samples were provided with a code concealing the true origin of the sample. At the end, an awareness programme was held at the Sri Lanka Tea Board and all the tasting attributes in the evaluation sheet and the scoring scale were discussed at length, in order to improve agreement. A common tasting session was also held for the panel members in which they could discuss their perceptions with each other and come to an agreement about the discriminating boundaries of different classes of attributes.

Data collection

Fifty-two black tea samples were collected from 13 different geographical areas of Sri Lanka. These teas have distinctive characteristics unique to their regions. Samples were collected for a one year (January 2010 – December 2010) period at monthly intervals. One sample collected from each region was duplicated to continually evaluate the consistency of individual panellists. Thus, each rater evaluated 65 (52 + 13) tea samples per month and 780 (65 × 12) samples during the study period. However, these regional tea samples were tasted at various sessions due to the practical limitations of having common tea tasting sessions. The following organoleptic attributes were considered for the study: colour, brightness, strength, flavour, aroma and quality. Evaluations were made according to a structured evaluation sheet. All attributes were categorised into five categories and ranks were assigned using a 1 to 5 scale.

Intra-class correlation coefficient (ICC)

Despite the training sessions that each panel member underwent, the reliability of the collected data may suffer both from individual assessor errors and panel agreement errors. The first step in the analysis of sensory data is therefore to identify individual assessors who perform abnormally or inconsistently, and have their data for the actual attribute(s) re-evaluated in further analysis (Kermit & Lengard, 2005)

The intra-class correlation coefficient (ICC) is a measure of the reliability of measurements or ratings, when two or more panellists rate a common set of samples. In this study ICC was used in two ways. One was to measure the absolute agreement: each sample was rated by a different and random subset of the panel of experts. The second use was to measure consistency: each sample was rated by the same experts. The interpretation of the ICCs is that the proportion of relevant variance that is associated with differences among the measured samples or experts. The equation of computing ICC is given below (Nichols, 1998),

$$ICC = \frac{S^2_{(b)}}{S^2_{(b)}+S^2_{(w)}}$$

where $S^2_{(w)}$ = pooled variance within subjects, $S^2_{(b)}$ = variance of the trait between subjects, $S^2_{(b)} + S^2_{(w)}$ = total variance of the ratings (i.e. the variance for all ratings, regardless of whether they are for the same sample).

Single measure ICC values were considered in this study to reflect the absolute agreement. The choice is based on the form of the ICC. The form reflects whether the reliability is to be calculated on a single measurement or by taking the average of two or more measurements taken by different panellists. In this study, the form of the ICC used was for the reliability calculated on a single measurement. The structure of the data for reliability analysis was as: $N = 65$ different regional tea samples (cases or rows), which are the objects being measured, and $k = 8$ raters (variables or columns), which denote the different measurements of the cases or objects. Although the cases have been selected systematically, as a large number of samples were selected from fairly different geographical areas, it was assumed to be a fixed factor and the factor 'taster' was assumed to be a random factor. The ICC estimates are based on mean squares obtained by applying ANOVA models to these data (Nichols, 1998). Two-way mixed effect model was used to calculate mean squares. In this study 'rater' effect was considered as random and the 'measure' effect was considered as fixed.

Mandel's bundle of lines (MBL) model

Mandel (1969) presented a method for the analysis of data representing functions of two variables when the response can be tabulated in a rectangular array. The procedure is based on a partition of the row by column interaction effects into a sum of terms, each of which is the product of a row factor by a column factor. As a result, one factor can be represented as a bundle of lines over the different levels of the other factor. Therefore, the behaviour of the levels of the second factor can be explained using slopes and intercepts.

A brief description of the basic MBL model is given below.

If u and v are two independent variables having levels m and n, respectively, the response variable y

which is dependent on u and v can be expressed as,

$$y = \varphi(u, v) + \text{experimental error} \quad ...(1)$$

Let i represent a row and j represent a column in a tabular arrangement of two explanatory variables u (row) and v (column) so that

$$y_{ij} = \varphi(u_i, v_j) + \varepsilon_{ij} \quad ...(2)$$

If η_{ij} is the expected value of y_{ij},

$$\eta_{ij} = \varphi(u_i, v_j) \quad ...(3)$$

Assume that the $\varphi(u_i, v_j)$ is of the particular type

$$\eta_{ij} = \varphi(u_i, v_j) = f(u_i) + g(u_i) * h(v_j) \quad ...(4)$$

If the average of the η_{ij} in column j is denoted by x_j

$$x_j = \bar{f} + \bar{g} * h(v_j) \quad ...(5)$$

where $\bar{f}$ and $\bar{g}$ are the averages of $f(u_i)$ and $g(u_i)$ over *i*.

Eliminating $h(v_j)$ in equation (4) by using equation (5)

$$h(v_j) = \frac{(x_j - \bar{f})}{\bar{g}}$$

$$\eta_{ij} = f(u_i) + g(u_i) * \frac{(x_j - \bar{f})}{\bar{g}}$$

By simplifying the above, following equation is resulted

$$\eta_{ij} = \left[f(u_i) - \frac{\bar{f}}{\bar{g}} g(u_i)\right] + \left[\frac{g(u_i)}{\bar{g}}\right] x_j \quad ...(6)$$

Here for any given *i*, $f(u_i)$ and $g(u_i)$ are constants. Thus, a plot of η_{ij} *versus* x_j will yield a straight line for any given *i*. Therefore, a linear relationship can be established for each level of the row factor. If graphically represented it looks like a bundle of straight lines. Thus, by the slope and intercept of each straight line it explains the behaviour of the levels of the row factor.

Application of Mandel's bundle of lines model for the present research

The research problem presented in this paper fulfils the prerequisites of Mandel's bundle of lines model, i.e. the response variable (tasting scores) is a function of two variables (taster and geographical areas) and the response can be arranged in a rectangular array. The rectangular array of variables is shown in Table 1.

Table 1: Representation of tasting scores (y_{ij}) in a rectangular array of taster (u_i) and geographical origins (v_j)

Taster	v_1	v_2	v_3	.....	v_{65}	
u_1	y_{11}					$y_{(1.)}$
u_2						$y_{(2.)}$
u_3						
⋮						
u_8						$y_{(8.)}$
$y_{(.j)}$	$y_{.1}$	$y_{.2}$			$y_{.65}$	

* i = 1,2,……,8 j = 1,2,……,65

As Mandel explains in his method (equation 6), the expected ratings of $y_{ij}(i.e.\eta_{ij})$ can be expressed as a linear function of its average response $\bar{y}_{.j}$ $(i.e.x_j)$. The slope and the intercept explain the behaviour of the taster and is unique to the particular taster.

Proposed debiasing approach using Mandel's bundle of lines model

The proposed debiasing approach can be executed in 3 steps. Step 1 was intent to find out a reliable estimate for x_j. Step 2 was executed to obtain a robust estimate for η_{ij}. During step 3, ratings adjusted for rater bias were computed using Mandel's bundle of lines model.

Step 1: Approach to find out a reliable estimate for x_j

As there are no set standards for tea sensory attributes, the best estimate for the true score of any attribute is the average of ratings of the panel of tasters. However, if tasters are not in agreement or the variability of scores is high, this is not a robust estimate. Therefore, average of the most consistent sub panel of tasters was used as the truth. In order to find out the most consistent sub panel of tasters, a tea sample from each geographical area was duplicated and the duplicate samples were evaluated by each taster for all organoleptic attributes for all 12 months. As a rule of thumb, tasters who reported ≤ 25 % of inconsistencies out of the total duplicate samples they evaluated, were considered as consistent tasters. The average score of the tasters who fulfilled the stipulated requirement was considered as a robust estimate for x_j in equation (6).

Step 2: Approach to obtain a robust estimate for η_{ij}

After calculating x_j (average score of most consistent tasters), ratings given by the rest of the panel members were subsequently regressed against x_j by using equation (7). Here, x_j was considered as the response

variable and the observed rating was considered as an explanatory variable. To avoid confusions, same symbols as in equation (6) were used.

$$x_j = \alpha + \beta y_{ij} + \varepsilon_j \quad ...(7)$$

y_{ij} – observed rating of i^{th} taster for j^{th} sample j = 1 to 65

x_j – average score of most consistent tasters for j^{th} sample j = 1 to 65

The predicted values in equation (7) are the best estimates of the values the taster would have reported if well calibrated (η_{ij}).

Step 3: Fitting Mandel's bundle of lines model to adjust ratings for rater bias.

$$\eta_{ij} = \alpha_i + \beta_i x_j + \varepsilon_{ij} \quad ...(8)$$

where i is the specific taster, j is a specific tea sample or geographical area, and x_j is the average rating of the most consistent tasters for the j^{th} tea sample. The η_{ij} values are the expected ratings that taster i gives to tea sample j on a particular attribute (computed predicted values of step 2). Here the error term is allowed to be different for different raters as done in Pal *et al*. (1998). The predicted values of equation (8), i.e. $\hat{\eta}_{ij}$ are the values for each taster after controlling for the bias in their slopes and intercepts.

RESULTS AND DISCUSSION

Intra-class correlation coefficients (ICCs) for the selected organoleptic attributes; colour (C), brightness (B), strength (S), flavour (F), aroma (A) and quality (Q) among the eight tasters are shown in Table 2.

According to Table 1, ICC values for the attribute colour ranges from 0.39 (September and October) to 0.8 (May). For the attribute brightness it varies from 0.21 (December) to 0.52 (May). Absolute agreement for the organoleptic attribute strength ranges from 0.05 (October) to 0.45 (May). Ranges of ICC values for the attributes flavour, aroma and quality are 0.29 (October) to 0.77 (May), 0.33 (August) to 0.68 (May) and 0.31 (October) to 0.8 (May), respectively.

Table 2: ICC values for absolute agreement for six organoleptic attributes at different months

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C	0.59	0.58	0.57	0.67	0.80	0.56	0.62	0.60	0.39	0.39	0.52	0.54
B	0.40	0.35	0.36	0.34	0.52	0.41	0.38	0.30	0.25	0.27	0.24	0.21
S	0.21	0.18	0.23	0.13	0.45	0.13	0.19	0.13	0.13	0.05	0.06	0.11
F	0.39	0.41	0.53	0.34	0.77	0.50	0.51	0.44	0.42	0.29	0.35	0.30
A	0.45	0.39	0.45	0.36	0.68	0.50	0.59	0.33	0.35	0.35	0.34	0.41
Q	0.42	0.39	0.51	0.38	0.80	0.49	0.50	0.33	0.39	0.31	0.33	0.33

C – colour; B – brightness; S – strength; F – flavour; A – aroma; Q - quality

All absolute agreement coefficients are, however, statistically significant, although only 20 coefficients out of 72 are greater than 0.5. For the attribute colour, except for September and October, all the months showed coefficients greater than 0.5. Only the coefficient for the month of May is greater than 0.5 for the attribute brightness. For the attribute strength, all agreement coefficients across the year are less than 0.5. In the attributes flavour and quality there are four instances (March and May – July) in which the coefficients exceeded 0.5. For aroma, May – July had more than a 0.5 agreement coefficient. For the months May – July, except for brightness and strength, all resulting coefficients are 0.5 or higher. As there is no set standard for the regional teas, this is a significant problem in the industry. In particular, this situation negatively affects the accuracy and precision of researches done in the tea industry, as most of these studies involve organoleptic measurements.

Other approaches used to assess reliability of panel of tasters

Pal *et al*. (1998) used a mixed effects model to estimate the true quality using a maximum likelihood technique. This requires a special software to make estimates, therefore, the method is not user friendly. According to Latreille *et al*. (2006), the mixed linear model approach has been used to estimate the reliability of a sensory panel.

Compared with Pal *et al*. (1998), Latreille *et al*. (2006) allow the term session in the model and all interaction terms. These model equations are given below:

$$y_{ij} = \mu + \varepsilon_i + \upsilon_{ij}$$

Pal *et al*. (1998)

Where y_{ij} = score given by the j^{th} taster for the i^{th} sample, μ = unknown basic quality, ε_i = deviation from basic quality due to sampling, and υ_{ij} = error associated with i^{th} sample due to j^{th} taster. In contrast, the more complex model is

$$y_{ijk} = \delta + \alpha_j + s_i + b_k + a_{ji} + c_{jk} + d_{ik} + \varepsilon_{ijk}$$

Latreille *et al*. (2006)

Where y_{ijk} = value of an attribute given by i^{th} judge for j^{th} product at k^{th} session, δ = intercept, α_j = product fixed effect, s_i = judge random effect, b_k = session random effect, a_{ji} = product * judge interaction random effect, c_{jk} = product * session interaction random effect, d_{ik} = judge * session interaction random effect, and ε_{ijk} = measurement error.

According to the approach suggested by Latreille *et al*. (2006), the nature of the possible differences are identified using the product * judge interaction effect estimates. This method is well adopted when the number of products tested or the number of sessions is small (less than five). When the number of products tested or the number of sessions increases, writing the contrasts becomes laborious, and the results become difficult to interpret. In addition, Hodgson (2008) mentioned that out of 65 panels of wine tasting experts, 24 panels resulted in a significant taster effect (37 %) and they suggest using ANOVA to test the taster effect. Therefore, it would be better for sensory analysts if there are simpler models to assess the reliability of a panel of tasters.

Kermit and Lengard (2005) also presented an ANOVA model to estimate the effect due to expert disagreement. They wrote

$$y_{ijkm}^{full} = \mu_k + \alpha_{ik} + \beta_{jk} + (\alpha\beta)_{ijk} + \varepsilon_{ijkm}^{full}$$

Kermit and Lengard (2005)

Where μ_k is the grand mean for attribute k and α_{ik} is the main effect contributed by assessor i for this attribute. The main effect from product j for the k^{th} attribute is represented by β_{jk}. The interaction effect $(\alpha\beta)_{ijk}$ provides the differences between assessors in measuring differences between products. The error term $\varepsilon_{ijkm}^{full}$ represents the residual variation due to replicates, and the superscript is included to indicate a full ANOVA model for further use in this analysis.

Kermit and Lengard (2005) have used consonance analysis and principal component analysis (PCA) to study the level of agreement within a panel. However, it does not allow them to identify the nature of the assessors' errors.

Mandel's bundle of lines approach

Compared to the generally complicated approaches previously used in literature to improve the reliability of organoleptic evaluations by panels of tasters, the authors of this study have used a simple approach. Mandel's bundle of lines allows to debias the personal differences from taster to taster and improve the agreement among members in the panel.

Results for step 1: Out of 78 tasting instances (13 months and 6 organoleptic attributes) the minimum inconsistencies have been reported by the taster T_2 (9 instances), while the maximum inconsistencies have been reported by the taster T_5 (39 instances). According to the results in Table 3, only three tasters (T_2, T_4 and T_6) out of the eight members have met the selection criteria (≤ 25 % inconsistencies).

Therefore, the average score of the most consistent three tasters were computed as the x_j value for each geographical origin, attribute and month.

Mutual agreement of three consistent panellists for six tasting attributes across months has been indicated in Table 4. It shows that 35 instances out of 72 have ICC values that are greater than 0.5. Agreement for the attributes brightness and strength are still poor.

Results for step 2: The results confirmed the linear relationship proposed in equation (7). Out of 564 models (8 tasters × 12 months × 6 attributes and with 12 missing values), 516 have proved significant linearity at α = 0.05. As a percentage it is 91.5 %. By further elaboration of regression analysis results it was evident that the Pearson correlation coefficients ranges from 0.39 – 0.94, 0.25 – 0.91, 0.25 – 0.90, 0.25 – 0.93, 0.26 – 0.94 and 0.23 – 0.92 for attributes colour, brightness, strength, flavour, aroma and quality, respectively. However, a comparatively low frequency of significant correlation coefficients was resulted in the attribute strength. Only 66 correlation coefficients were significant out of 96 models. But as a percentage, ~70 % of models showed significant linearity.

Table 3: Relative frequencies of inconsistencies reported for each taster

Taster	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
Number of inconsistencies reported	33	9	23	20	39	17	27	34
Total number of evaluations / tastings	72	78	71	78	78	78	78	78
Percentage of inconsistent responses	46	12	32	26	50	22	35	44

(Selection criteria: ≤ 25 % of inconsistencies)

Table 4: Intra-class correlation coefficients for most consistent tasters

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C	0.50	0.64	0.66	0.65	0.64	0.53	0.60	0.68	0.52	0.53	0.48	0.55
B	0.37	0.40	0.46	0.37	0.40	0.48	0.54	0.25	0.24	0.31	0.35	0.36
S	0.17	0.04	0.24	0.21	0.16	-0.04	0.16	0.21	0.16	-0.02	0.06	0.00
F	0.59	0.47	0.62	0.37	0.49	0.58	0.70	0.59	0.42	0.48	0.67	0.58
A	0.57	0.50	0.60	0.44	0.51	0.66	0.75	0.56	0.47	0.53	0.63	0.59
Q	0.56	0.46	0.55	0.51	0.43	0.57	0.61	0.47	0.41	0.45	0.45	0.46

C - colour; B - brightness; S - strength; F - flavour; A - aroma; Q - quality

For attributes colour, brightness, flavour, aroma and quality, respectively 98 %, 88 %, 91.5 %, 96 % and 95 % of fitted models confirmed linearity. Therefore, predicted values of equation (7) can be reasonably considered as the robust estimates for η_{ij} .

Results for step 3: Similar results as in step 2 were shown for step 3. Therefore, regression analysis results of equation (8) proved that the response of different tasters could be able to represent as a bundle of straight lines over different geographical origins (tea sample). Thus, the predicted values of equation (8) could be considered as ratings adjusted for rater bias.

Agreement after bias correction using Mandel's bundle of lines model

ICCs calculated after correction of bias using Mandel's bundle of lines model are shown in Table 5. According to these results, 64 instances out of 72 have absolute agreement greater than 0.5. For the attribute colour, ICC values varied from 0.77 to 0.95. The range of ICCs for the attribute brightness varied from 0.23 to 0.86. For strength it ranged from 0.14 to 0.67. The ranges resulted for attributes flavour, aroma and quality are 0.38 to 0.89, 0.36 to 0.87, and 0.43 to 0.88, respectively. This is a significant improvement over the previous ICC scores.

Table 5: ICC values for each month after bias correction

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C	0.94	0.89	0.87	0.92	0.47	0.92	0.95	0.97	0.94	0.77	0.89	0.94
B	0.86	0.80	0.73	0.79	0.23	0.66	0.73	0.74	0.62	0.62	0.57	0.52
S	0.67	0.48	0.63	0.51	0.14	0.50	0.57	0.37	0.57	0.45	0.31	0.28
F	0.72	0.74	0.89	0.76	0.38	0.66	0.87	0.76	0.84	0.69	0.66	0.69
A	0.81	0.73	0.77	0.79	0.36	0.80	0.87	0.65	0.71	0.71	0.62	0.68
Q	0.78	0.79	0.88	0.59	0.43	0.68	0.83	0.70	0.80	0.69	0.66	0.64

C - colour; B - brightness; S - strength; F - flavour; A - aroma; Q - quality

CONCLUSION

The consensus among panellists for all organoleptic attributes (i.e., colour, brightness, strength, flavour, aroma and quality) across months is generally poor in rating regional teas. Out of all the sensory attributes studied, disagreement on strength and brightness is most prominent. Using Mandel's bundle of lines approach, one can characterise the behaviour of each taster by estimating slopes and intercepts of each taster's line. After debiasing tasters using the fitted bundle of lines method, the consensus of panellists for all attributes improves significantly. Therefore, the Mandel's bundle of lines approach can be successfully used as a simple, user-friendly approach to calibrate tasters and to improve agreement among panels of tasters who rate regional teas.

Acknowledgement

Authors would like to greatly acknowledge the Tea Research Institute of Sri Lanka for financial support. Thanks are also expressed to the eight tea tasting experts for the sensory testing and Mr. H.K.K. Deshapriya, former Technical Officer, Biometry Unit, Tea Research Institute of Sri Lanka for his assistance in collection and preparation of regional tea samples for sensory evaluation.

REFERENCES

1. Álvarez P. & Blanco M.A. (2000). Reliability of the sensory analysis data of a panel of tasters. *Journal of Science for Food and Agriculture* **80**(3): 409 – 418.
2. Benedetti S., Pompei C. & Mannino S. (2004). Comparison of an electronic nose with the sensory evaluation of food products by "triangle test". *Electroanalysis* **16**: 1801 – 1805. DOI: https://doi.org/10.1002/elan.200303036
3. Deventer D. & Mallikarjunan P. (2002). Comparative performance analysis of three electronic nose systems using different sensor technologies in odor analysis of retained solvents on printed packaging. *Journal of Food Science* **67**(8): 3170 – 3183. DOI: https://doi.org/10.1111/j.1365-2621.2002.tb08878.x
4. Hodgson R.T. (2008). An examination of judge reliability at a major US wine competition. *Journal of Wine Economics* **3**(02): 105 – 113. DOI: https://doi.org/10.1017/S1931436100001152
5. Kermit M. & Lengard V. (2005). Assessing the performance of a sensory panel-panellist monitoring and tracking. *Journal of Chemometrics* **19**(3): 154 – 161.
6. Latreille J., Mauger E., Ambroisine L., Tenenhaus M., Vincent M., Navarro S. & Guinot C. (2006). Measurement of the reliability of sensory panel performances. *Food Quality and Preference* **17**(5): 369 – 375. DOI: https://doi.org/10.1016/j.foodqual.2005.04.010
7. Mandel J. (1969). The partitioning of interaction in analysis of variance. *Journal of Research of the National Bureau of Standards, Series B* **73**: 309 – 328. DOI: https://doi.org/10.6028/jres.073B.031
8. McEwan J.A., Hunter E.A., van Gemert L.J. & Lea P. (2002). Proficiency testing for sensory profile panels: measuring panel performance. *Food Quality and Preference* **13**(3): 181 – 190.
9. Meiselman H.L. (1993). Critical evaluation of sensory techniques. *Food Quality and Preference* **4**(1): 33 – 40.
10. Nichols D.P. (1998). *Choosing an Intraclass Correlation Coefficient*. Available at *http://www.ats.ucla.edu/stat/spss/library/whichicc.htm*. Accessed 29 January 2016.
11. Pal M., Paul S.K. & Das A.K. (1998). Assessment of tea quality associating biochemical quality parameters and taster's score. *Two and a Bud* **45**(1): 19 – 22.
12. Shen N., Moizuddin S., Wilson L., Duvick S., White P. & Pollak L. (2001). Relationship of electronic nose analyses and sensory evaluation of vegetable oils during storage. *Journal of the American Oil Chemists' Society* **78**(9): 937 – 940. DOI: https://doi.org/10.1007/s11746-001-0367-z
13. Stone H. & Sidel J.L. (1993). *Sensory Evaluations Practices*. Academic Press, California, USA.
14. Vaamonde A., Sánchez P. & Vilariño F. (2000). Discrepancies and consistencies in the subjective ratings of wine-tasting committees. *Journal of Food Quality* **23**: 363 – 372.