Received: 3 October 2019 | Revised: 25 November 2019 | Accepted: 20 December 2019

DOI: 10.1002/cem.3216

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Comparison of supervised learning statistical methods for classifying commercial beers and identifying patterns

Dániel Koren[1] | Laura Lőrincz[2] | Sándor Kovács[3] | Gabriella Kun-Farkas[1] | Beáta Vecseriné Hegyes[1] | László Sipos[4]

[1]Department of Brewing and Distilling, Faculty of Food Science, Szent István University, Budapest, Hungary

[2]Egis Pharmaceuticals PLC, Budapest, Hungary

[3]Department of Economic and Financial Mathematics, Institute of Statistics and Methodology, Faculty of Economics and Business, University of Debrecen, Debrecen, Hungary

[4]Department of Postharvest Sciences and Sensory Evaluation, Faculty of Food Science, Szent István University, Budapest, Hungary

Correspondence

Sándor Kovács, Department of Economic and Financial Mathematics, Institute of Statistics and Methodology, Faculty of Economics and Business, University of Debrecen, 138 Böszörményi út, H-4032 Debrecen, Hungary.

Email: kovacs.sandor@econ.unideb.hu

Funding information

A MAGYAR ZÖLDSÉG- ÉS GYÜMÖLCSTERMESZTÉS AGRO-ÖKOLÓGIAI ÉS BIOLÓGIAI ERŐFORRÁSAINAK OPTIMÁLIS KIAKNÁZÁSA TECHNOLÓGIAI, EREDETVÉDELMI ÉS PIACFEJLESZTÉSI ESZKÖZÖKKEL, Grant/Award Number: VEKOP-2.3.3-15-2017-00022; Tudományos utánpótlás erősítése a hallgatók tudományos műhelyeinek és programjainak támogatásával, a mentorálás folyamatának kidolgozásával, Grant/Award Number: EFOP-3.6.3-VEKOP-16-2017-00005; János Bolyai Research Scholarship

Abstract

In this study, 13 properties (alcohol-, real extract-, flavonoid-, anthocyanin, glucose, fructose, maltose, sucrose content, EBC [European Brewery Convention] and L*a*b* color, bitterness) of 21 beers (alcohol-free pale lagers, alcohol-free beer-based mixed drinks, beer-based mixed drinks, international lagers, wheat beers, stouts, fruit beers) were determined. In the first step, multiple factor analysis (MFA) was performed for the whole data and five clusters (target classes) were determined; then, a bootstrapping was applied to establish a balanced data so as every cluster should contain 100 samples and the total sample size is 500. In the second step, 12 supervised learning algorithms (random trees [RND], Quinlan's C4.5 decision tree algorithm [C4.5], Iterative Dichotomiser 3 algorithm [ID3], cost-sensitive decision tree algorithm [CSMC4], cost-sensitive classification tree [CSCRT], *k*-nearest neighbors algorithm [KNN], radial basis function [RBF], multilayer perceptron neural network [MLP], prototype nearest neighbor [PNN], linear discriminant analysis [LDA], naïve Bayes with continuous variables [NBC], partial least squares discriminant analysis [PLS-DA]) were applied to classify each brand into the target classes. Furthermore, several error rates were calculated: re-substitution error rate (RER), cross-validated error rate (CV), bootsrap error (BOOT), leave-one-out (LOO), and train-test error rate (TRAIN). The MFA could discriminate five groups, which can be characterized by some analytical parameters, and the other multivariate methods performed similarly. The methods can be discriminated best based on the BOOT, CV, and LOO. The best estimation methods are the C4.5, CSMC4, and CSCRT; these performed best along the flavonoid content and EBC color. It identified that the methods most sensitive to the properties are the NBC. The classification ability fluctuated greatly in the case of three properties (glucose, maltose, sucrose). A remarkable fluctuation has been experienced in the case of L*a*b* color parameters, flavonoid content, EBC color, and bitterness by NBC method.

KEYWORDS
beer, error estimation, fruit beer, learning algorithms, multiple factor analysis (MFA)

1 | INTRODUCTION

Beer is one of the most consumed beverages worldwide. There are numerous studies, which have proven that low or moderate beer consumption can prevent cardiovascular, neurodegenerative diseases, and conformation of cancers. Furthermore, it is published well that it contains several components, which have beneficial effects on human health.[1] The most common parameters measured in breweries are the color, bitterness, alcohol, and real extract content. The EBC (European Brewery Convention) color of a sample is determined by a spectrophotometer at one wavelength (λ = 430 nm); thus, it is only able to differentiate between samples that have the same hue. This method was developed a long time ago to determine the color of traditional beers, which have a color from pale yellow through brown until dark brown but cannot differentiate between samples having different hues, for example, beers produced with various fruits. The bitterness of a sample means the isomerized alpha-acid content of a beer, which gives the bitter taste to a product. The alcohol content is usually determined by a beer analyzer, which is based on NIR (near infrared) spectroscopy. The real extract content of a beer means the dry matter content after fermentation, which are mainly unfermented carbohydrates such as dextrins.[2]

To meet consumer needs and to make beer more widely consumed, new beer types and styles brewed with fruits are produced and introduced to the public. Beers brewed with fruits can be divided into two main groups, one is real fruit beers and the other is beer-based mixed drinks, also known as radlers. The main differences between these two categories are that fruit beers are aged with whole fruits, fruit puree or juice, which are added in 5% to 10% to the product, usually after the main fermentation and aged together for months. Therefore, the alcohol content of these beers is higher and the taste of fruit is not dominating. In the case of radlers or beer-based mixed drinks, the final product, the beer is mixed with fruit juice in a ratio of about 50% to 50%, making the alcohol content lower, usually under 2 v/v %; thus, the fruit aroma is overwhelming the taste of the base beer. These beer-based mixed drinks are also produced in alcohol-free version.

There is a lot of information available in the literature about the content of health-promoting components of beers, for example, phenolic composition, like flavonoids and anthocyanins, and antioxidant capacity, which was investigated by Moura-Nunes and co-workers who managed to model FRAP (Ferric ion Reducing Antioxidant Power) as a function of density, refractive index, bitterness, and ethanol content by using chemometrics.[3] There are studies that deal with the tristimulus color characteristics (L*a*b*), carbohydrate profile, and physicochemical properties of these products.[4,5] The L*a*b* color space is based on opponent color theory: L* shows the brightness, the position between light and dark, a* is red versus green, and b* is yellow versus blue[6].

Chemometric tools are used mainly in the case of results originating from NIR, FTNIR (Fourier transform NIR), e-nose, e-tongue, and sensory evaluation results.[7-9] Ghasemi-Varnamkhasti and Forina studied the use of NIR spectroscopy during aging of beers. For qualitative analysis principal component analysis (PCA), *k*-nearest neighbors (KNN), linear discriminant analysis (LDA), stepwise LDA, genetic algorithms (GA), and Gram-Schmidt supervised orthogonalization (SELECT) were employed.[10] Cetó used an electronic tongue for the discrimination of different commercial beer types. For the identification of initial patterns PCA and to achieve the correct recognition of sample varieties, LDA was used.[11] Varmuza and co-workers studied the concentration profiles of fresh and aged beer analyzed by univariate statistics (paired *t* test, correlation coefficients) and multivariate statistics. Their results showed that the PCA is able to find clusters of similar beer samples and detect outliers.[12] Multiple factor analysis (MFA) is a generalization of PCA in which grouped variables are equalized by a weight vector. The advantage of this analysis is that it can be applied to categorized and continuous data, and good visualization properties of MFA make the data interpretation easier.[13] Granato and co-workers determined the total polyphenol content (TPC), total flavonoid content (TFC), antioxidant activity (AA), and color. For further analysis, PCA, hierarchical cluster analysis (HCA), and supervised pattern recognition methods were used such as KNN and soft independent modeling of class analogy (SIMCA).[14]

There is less information about the multivariate evaluation of results of the bioactive components, carbohydrate content, color characteristics, and physicochemical properties of beer. Furthermore, there is no information about beer-based mixed drinks or fruit beers in scientific publications. Therefore, except for the basic parameters like alcohol content, real extract content, EBC color, and bitterness, in our study, the L*a*b* color, sugar composition, flavonoid, and anthocyanin content of the samples were determined and chemometrics was applied to be able to characterize the products in more detail and to see which parameter contributes to the classification of these products. Error estimation is critical to classification because the validity of the resulting classifier model is based on the accuracy of the error estimation procedure, and much greater effort needs to be focused on error estimation.[15,16]

Our aim was to fill this space by using multivariate statistics and chemometrics for the evaluation of our result. For this, the ranking of the classifying methods based on the error estimators has been carried out, the sensitivity of the methods depending on the variables involved has been tested, and the classification rank of the properties of beers has been determined.

2 | MATERIALS AND METHODS

2.1 | Beer samples

Twenty-one different beers were purchased, which are available in Hungarian retail. Apart from the most common ones, some special kinds of beers were included, for example, beers mixed with fruit juices or aged with fruits, Hungarian craft beers, as well as alcohol-free products. Beers were classified based on the Beer Style Guidelines of the Beer Judge Certification Program.[17] Three alcohol-free pale lagers, three alcohol-free beer-based mixed drinks, three beer-based mixed drinks, three international lagers, three wheat beers, three stouts, and three fruit beers were investigated. With this data selection, we aimed to involve many types of beer. Samples were homogenized and degassed by sonication in ultrasonic bath for 5 minutes prior to analysis (Table 1).

2.2 | Analytical methods

EBC color values were determined according to the standard EBC color measuring method.[18] Tristimulus values of the samples were measured by a Konica-Minolta CR-410 colorimeter in 10-mm path length quartz cuvettes. Alcohol and real extract content was determined by an Anton Paar DMA 4500 density meter and Alcolyzer Plus. The total flavonoid content was determined according to the description of Fernandes et al on the basis of aluminium complexation reaction.[19] The total monomeric anthocyanin pigment content determination was performed by pH differential method according to the description of Lee et al.[20] The carbohydrate content determination (glucose, fructose, sucrose, maltose) was carried out by Waters high-performance liquid chromatography using RI detector. Isocratic method was applied using water: acetonitrile (20:80%) as eluent and Waters Spherisorb S5 NH_2 column for separation.

2.3 | Statistical analysis

In our study, 13 properties of 21 beers were determined, and each analytical measurement was carried out in five parallels. First, the MFA was performed for the whole data and five clusters (target classes) were determined. We created the first MFA for 13 blocks, each block representing a single variable, and a block containing five measurements for the same variable. The goal is to create an average point on the MFA map, resulting in expert clustering.

Second, a bootstrapping was applied to establish a “balanced data” so as every cluster should contain 100 samples. Each group had five replicate measurements, so item numbers were as follows: in cluster 1 ($n = 7 \times 5$), cluster 2 ($n = 2 \times 5$), cluster 3 ($n = 3 \times 5$), 4 cluster ($n = 3 \times 5$), cluster 5 ($n = 6 \times 5$) (see Figure 1). These element numbers were incremented by bootstrapping so that each group contained 100 samples, which was large enough but also had the same number of elements. This resulted in a total sample size of 500 in the 5 groups. The next step was to apply 12 supervised learning algorithms—random trees (RND), Quinlan's C4.5 decision tree algorithm (C4.5), Iterative Dichotomiser 3 algorithm (ID3), cost-sensitive decision tree algorithm (CSMC4), cost-sensitive classification tree (CSCRT), *k*-nearest neighbor's algorithm (KNN), radial basis function (RBF), multilayer perceptron neural network (MLP), prototype nearest neighbor (PNN), linear discriminant analysis (LDA), naïve Bayes with continuous variables (NBC), and partial least squares discriminant analysis (PLS-DA)—to classify each brand into the target classes. These techniques can be grouped into logic-based (RND, C4.5, ID3, CSMC4, CSCRT), perceptron-based (MLP, PNN), statistical learning (LDA), or distance-based ones (KNN, NBC, PLS-DA). We would have liked to measure the classifying

TABLE 1 Investigated parameters of the beer samples

No.	EBC (n = 5)		L (n = 5)		a (n = 5)		b (n = 5)		Flavonoid Content [μM Catechin Equivalent] (n = 5)		Monomer Anthocyanin [mg/L Cyanidin-3-glucoside] (n = 5)		Fructose [g/L] (n = 5)		Glucose [g/L] (n = 5)		Sucrose [g/L] (n = 5)		Maltose [g/L] (n = 5)		Alcohol [v/v%] (n = 5)		Real extract [m/m%] (n = 5)		IBU (n = 5)	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	sd	M	sd	M	sd	M	sd
Alcohol-free pale lagers (beer type)																										
1	7.8	0.2	94.0	1.1	−2.1	0.0	37.0	0.2	36.5	1.3	0.0	0.0	0.6	0.0	0.4	0.0	0.0	0.0	11.1	0.6	0.3	0.1	3.7	0.1	14.1	0.1
2	8.4	0.8	89.6	0.1	−1.3	0.0	35.7	0.0	48.4	0.8	0.0	0.0	1.0	0.1	3.6	0.1	0.5	0.0	20.1	0.3	0.1	0.0	5.9	0.1	18.3	0.2
3	8.5	0.4	92.3	0.1	−3.1	0.0	39.3	0.1	46.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.3	0.3	0.4	0.1	4.7	0.0	17.7	0.1
Alcohol-free beer-based mixed drinks (beer type)																										
4	24.3	0.9	49.4	1.9	58.4	0.5	28.9	1.1	86.7	0.8	7.7	6.2	7.7	0.2	7.6	0.3	14.1	0.3	3.3	0.0	0.1	0.0	5.8	0.2	6.2	0.2
5	17.8	0.7	91.2	0.0	−2.3	0.0	26.5	0.0	9.4	0.7	0.0	0.0	13.2	0.4	3.2	0.1	0.0	0.0	8.5	0.0	0.1	0.0	4.6	0.0	8.4	0.1
6	20.7	0.6	84.6	0.5	7.3	0.0	35.5	0.1	0.0	0.0	0.0	0.0	14.3	0.3	3.5	0.2	0.0	0.0	6.1	0.2	0.1	0.0	6.1	0.1	5.6	0.1
Beer-based mixed drinks (beer type)																										
7	25.1	0.7	88.2	0.2	−2.8	0.3	24.3	0.1	1.8	1.4	0.0	0.0	6.8	0.2	1.2	0.0	0.0	0.0	0.0	0.0	1.4	0.1	3.1	0.1	6.7	0.1
8	19.8	0.8	84.6	0.2	8.5	0.0	32.2	0.0	0.0	0.0	0.0	0.0	15.4	0.4	5.2	0.2	0.0	0.0	0.0	0.0	1.3	0.2	4.1	0.0	5.5	0.3
9	26.0	0.1	50.5	1.0	56.0	0.8	29.3	0.7	47.6	2.0	357.6	47.2	10.0	0.1	9.9	0.0	2.4	0.0	0.0	0.0	1.4	0.1	5.2	0.2	7.4	0.1
International lagers (beer type)																										
10	7.9	0.4	92.8	0.1	−2.9	0.0	38.3	0.0	126.4	1.2	3.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	3.7	0.0	20.3	0.2
11	12.8	0.3	90.2	0.3	−1.8	0.0	51.7	0.1	141.6	2.1	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.4	0.1	5.0	0.1	27.4	0.2
12	12.6	0.8	89.6	0.0	−2.1	0.0	52.2	0.3	130.4	0.9	3.1	0.0	0.0	0.0	1.1	0.1	0.0	0.0	0.0	0.0	4.6	0.1	4.0	0.0	42.2	0.1
Wheat beers (beer type)																										
13	15.6	0.1	92.0	0.8	−1.4	0.0	45.1	1.0	126.8	0.6	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.5	0.0	4.4	0.2	12.4	0.2
14	28.3	0.6	82.2	1.0	4.7	0.1	56.2	0.2	135.5	4.0	10.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.2	4.2	0.1	12.6	0.1
15	33.5	0.5	80.9	0.5	1.8	0.1	41.1	0.2	165.6	1.4	13.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	3.8	0.1	11.3	0.1
Stouts (beer type)																										
16	91.6	0.4	37.7	0.3	26.3	0.2	29.6	0.5	413.1	2.8	35.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.2	0.1	5.0	0.0	18.7	0.1
17	107.3	0.9	30.6	2.3	21.7	1.2	15.9	1.1	365.2	1.1	45.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.1	3.2	0.0	17.2	0.2
18	108.9	0.7	34.4	1.9	25.2	1.6	19.5	0.7	394.7	0.6	43.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.9	0.0	4.7	0.1	18.5	0.1
Fruit beers (beer type)																										
19	44.7	0.2	49.7	0.3	38.7	1.3	29.8	0.9	393.2	0.9	40.6	0.0	6.5	0.1	10.9	0.1	0.0	0.0	0.0	0.0	8.1	0.2	5.5	0.2	9.6	0.1
20	56.5	0.5	51.8	2.1	31.1	1.0	26.2	3.8	451.8	3.8	39.1	0.5	12.9	0.2	4.5	0.1	0.0	0.0	0.0	0.0	6.2	0.2	6.7	01.	7.4	0.3
21	51.1	0.3	46.6	1.3	40.6	1.8	25.0	1.9	426.5	0.6	67.6	0.1	5.5	0.1	6.2	0.1	0.0	0.0	1.7	0.2	7.7	0.1	7.6	0.1	12.3	0.1

Note. Alcohol-free pale lagers: 1. Arany Ászok alkoholmentes, 2. Soproni Maxx alkoholmentes, 3. Dreher 24; Alcohol-free beer-based mixed drinks: 4. Pécsi Radler meggy mentes, 5. Borsodi citrom zero, 6. Borsodi grapefruit zero; Beer-based mixed drinks: 7. Borsodi citrom, 8. Borsodi grapefruit, 9. Pécsi Radler meggy; International lagers: 10. Heineken, 11. Pilsner Urquell, 12. Staropramen; Wheat beers: 13. Edelweiss, 14. Paulaner, 15. Valentins; Stouts: 16. Hercule stout, 17. Guinness, 18. Soproni Démon; Fruit beers: 19. Kasteel Rouge, 20. Liefmans Cuvee Brut, 21. Delirium Red.
Abbreviations: EBC, European Brewery Convention; IBU, International Bittering Units.

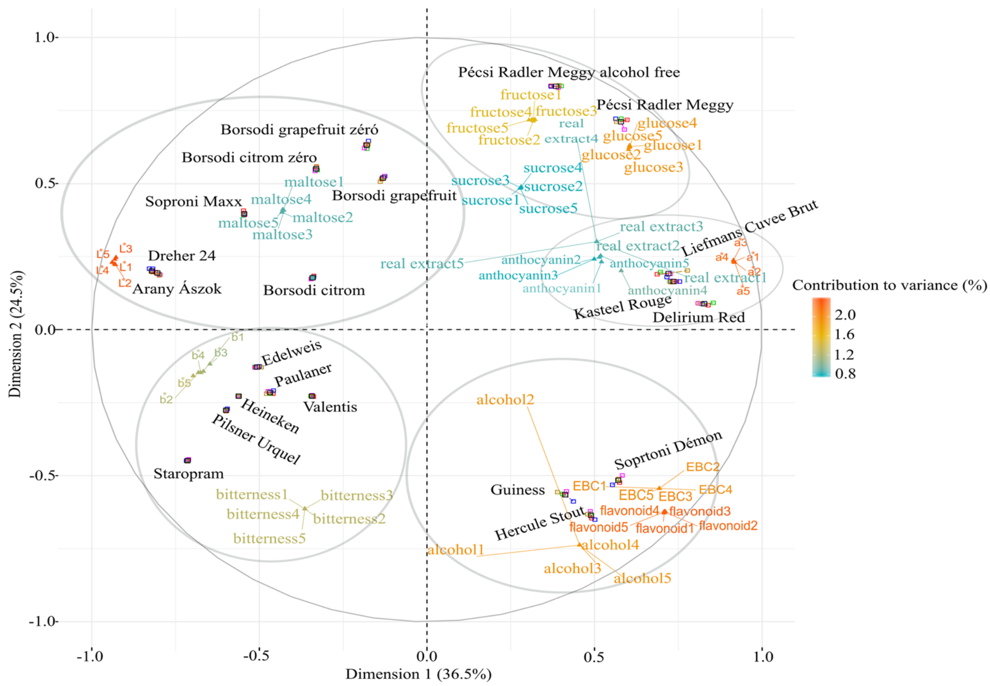

FIGURE 1 Multiple factor analysis of 13 properties and 21 beers samples (D1 = 36.5%; D2 = 24.5%). EBC, European Brewery Convention

ability of each supervised learning method for every determined property. Therefore, we used only one property for the classification at one time, and the following several error rates were calculated: re-substitution error rate (RER), cross-validated error rate (CV), bootsrap error (BOOT), leave-one-out (LOO) and train-test error rate (TRAIN). In this way, we got a data table of 12 methods for 13 properties and the columns were the 5 error rates.

The final step was to analyze the results with MFA again. We arranged the results table in several ways so we performed two kinds of MFA. One is to present the strength and weakness of both methods and properties with respect to the error rate. The other one was performed to present the differences between the applied methods with respect to the properties. On the other hand, we were able to detect the most and the less influential properties for the classifying performance of the studied methods.

RER is the percentage of misclassified individuals in the sample. During the calculation, we used the same data set for training and testing according to Molinaro and coworkers[21]; therefore, we underestimate the *true error rate*. In order to take this problem, several error rates should also be calculated. Regarding the LOO error rates, we train on 499 instances and classify on 1 test instance. Regarding CV error, we subdivided our data set into 10 folds (consisted of 50 instances per fold) and repeated the process on 9 train folds and 1 test fold. Bootstrap error is somehow different compared with the other methods as it rather estimates the bias of the RER. Within this process, we calculate the RER several times by using 25 replications of the original data. In the training error calculation, 0.7 was used as a proportion of the train set (350 instances) and repeated the training 10 times by randomly selecting the new train set. The MFA-based expert classification was compared with the most widely applied hierarchical (agglomerative hierarchical clustering [AHC] Euclidean distance and Ward's method), centroid-based (*k*-means), and the novel spectral clustering methods.[22] Comparison was done by the following cluster validity indices: Dunn,[23] Silhouette,[24] and C-index.[25] The error calculations have been performed by using TANAGRA[26] and all other calculations in R system[27] (for hierarchical clustering: hclust package; for calculating cluster validity indices: clusterCrit package; for performing spectral clustering: SNFtool package).

3 | RESULTS AND DISCUSSION

The MFA was performed for the whole data and five clusters (target classes) were determined. More than 60% of the variance is explained, which is adequate. The L* color parameter, flavonoid content, and a* color parameter contributed to the variance to the highest extent. The L* color parameter describes the darkness; The higher the L* value, the lighter the color. In the case of traditional beer types produced without fruits, it is a good discriminating factor. In the case of beers produced with fruit, the a* value can discriminate between the different hues of the products on the axis of green and red: The more reddish the color, the higher the a* value. The flavonoid content depends on the raw materials used, such as malts, hops, and fruit, that is why it is a good discriminating factor (Figure 1.).

The discrimination of the five groups can be explained well from a technological point of view. In the first group, the alcohol-free and low-alcoholic pale products including traditional alcohol-free beers and beer-based mixed drinks produced with lemon and grapefruit juice can be found. These beers are the lightest in color (they have the lowest EBC color and the highest L* value) and have the lowest flavonoid content. In the second group, two beer-based mixed drinks produced with sour cherry juice are located. These have dark red color due to the sour cherry juice they were mixed with. They have low alcohol content, one of them is alcohol-free, while the other one contains less than 1.5 v/v%. In the third group, fruit beers aged with sour cherry can be found. These also have a dark red color, but their alcohol content is much higher (over 6 v/v%) than that of the previous group. In the fourth group, stouts are located, which have the lowest L* values and the highest EBC color values. In the fifth group, the international pale lagers and wheat beers are located; these have a pale color, an average alcohol content between 4.4 and 5 v/v%, and also an average flavonoid content compared with the other samples.

The MFA-based classification's match rate was 89% with *k*-means, 71% with AHC Ward's method, and 71% compared with spectral clustering. Each of the clustering indices performed much better for the MFA than for the other clustering. Based on the results, MFA had the highest Dunn and Silhouette index values and the lowest C-index values (Table 2). In summary, MFA clustering is very similar to all other clusters, but it has created a more valid and much better quality clusters.

TABLE 2 Comparison of clustering method's and cluster validity indices

	Multiple Factor Analysis (MFA)	*k*-means Clustering	Ward's Method (AHC)	Spectral Clustering
Dunn index	0.1576	0.0449	0.0635	0.1200
Silhouette index	0.3481	0.2428	0.2314	0.2402
C-index	0.0932	0.2579	0.1079	0.2282

Abbreviation: AHC, agglomerative hierarchical clustering.

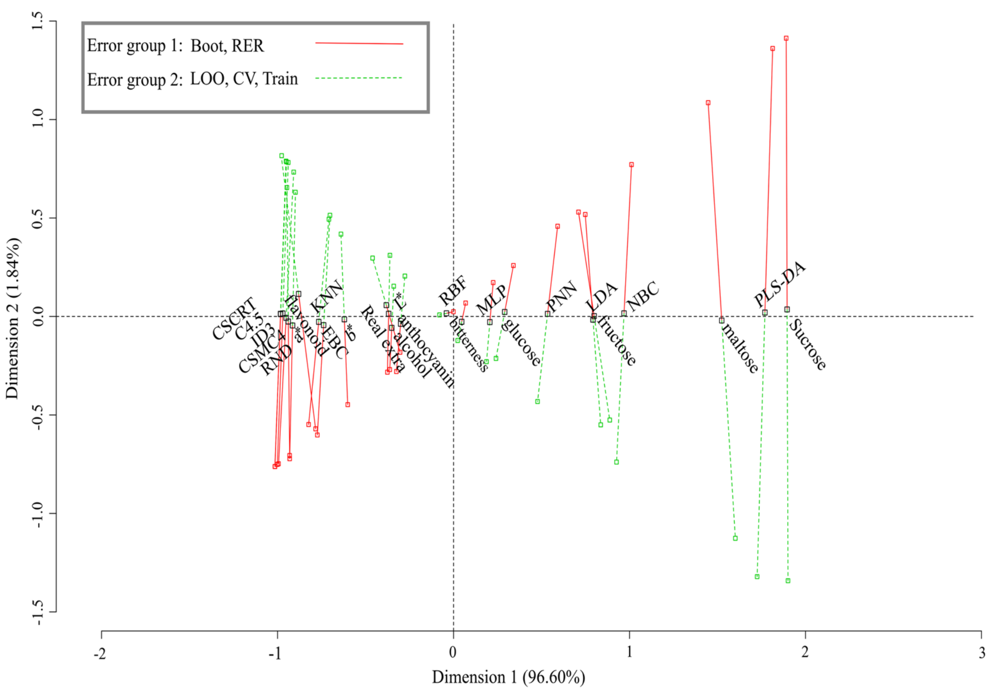

FIGURE 2 Comparison of methods based on two groups of errors (D1 = 96.6%; D2 = 1.84%). BOOT, bootstrap error; C4.5, Quinlan's C4.5 decision tree algorithm; CSCRT, cost-sensitive classification tree; CSMC4, cost-sensitive decision tree algorithm; CV, cross-validated error rate; EBC, European Brewery Convention; ID3, Iterative Dichotomiser 3 algorithm; KNN, *k*-nearest neighbors algorithm; LDA, linear discriminant analysis; LOO, leave-one-out; MLP, multilayer perceptron neural network; NBC, naïve Bayes with continuous variables; PLS-DA, partial least squares discriminant analysis; PNN, prototype nearest neighbor; RBF, radial basis function; RER, re-substitution error rate; RND, random trees; TRAIN, train-test error rate

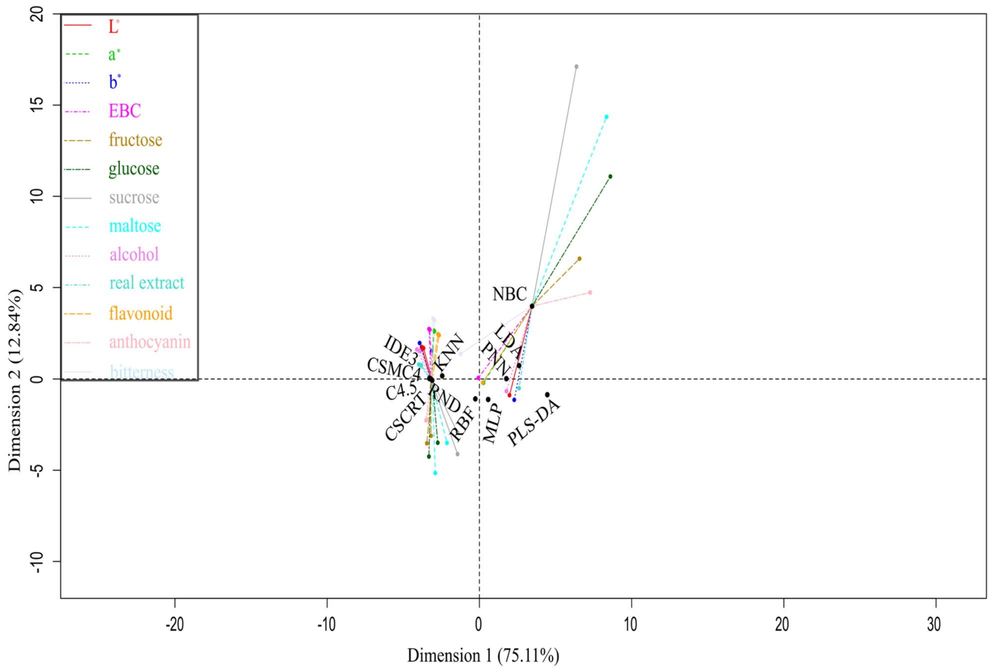

FIGURE 3 Classification ability of 12 methods with 13 analytical properties (D1 = 75.11%; D2 = 12.84%). C4.5, Quinlan's C4.5 decision tree algorithm; CSCRT, cost-sensitive classification tree; CSMC4, cost-sensitive decision tree algorithm; EBC, European Brewery Convention; KNN, *k*-nearest neighbors algorithm; LDA, linear discriminant analysis; MLP, multilayer perceptron neural network; NBC, naïve Bayes with continuous variables; PLS-DA, partial least squares discriminant analysis; PNN, prototype nearest neighbor; RBF, radial basis function; RND, random trees

The methods can be most distinguished, and their performance can be ranked based on the BOOT, CV, and LOO error estimators. The errors can be divided into two groups, which occur from their methodological specificity. In one group, the methods using the same data set were placed (BOOT, RER), and in the other group, the train-test split methods were placed (LOO, CV, TRAIN). We performed another MFA on the methods as well as on the properties where the blocks were the two groups of error rates. The methods with the best prediction ability in general were C4.5, CSMC4, and CSCRT. These gave good results along flavonoid content and EBC color properties. Then, the ID3 and RND methods followed. They gave good prediction along a* and b* color paremeters. Then, the third group of methods (KNN, RBF) performed well on real extract content and L* properties. The methods with average predictive ability can be identified, such as MLP, PNN, LDA, and properties (glucose, IBU [International Bittering Units], alcohol, antocyanin) with which average predictive ability can be achieved. The properties that, along with the methods, provided the worst prediction were maltose, sucrose, and fructose, and the worst predictive methods were PLS-DA and NBC. In the latter three methods, there were particularly large differences in the two types of errors (relatively higher RER, BOOT, lower LOO, TRAIN), while in the case of the best predictive methods, the inverse can be seen (lower RER, BOOT and relatively higher LOO, TRAIN, CV) (Figure 2.).

The third MFA was performed on all the error rates for each method, and the blocks were the different properties. The overall classification errors are the most volatile regarding NBC, CSMC4, C4.5, CSCRT, and RND. The classification ability of other methods was less volatile or less dependent on the result variable. It also appears that the most sensitive methods are NBC for properties. Particularly, in the case of three properties (glucose, maltose, sucrose), the classifying ability fluctuated greatly. Even with color components (L* a* b*), flavonoid content, EBC color, and bitterness, significant fluctuations in NBC were experienced (Figure 3.).

4 | CONCLUSIONS

Beers cannot be classified based on one or two properties, especially beers produced with fruits. For the appropriate classification, multivariate statistical methods are needed. In brewery practice, the four most common parameters of beer are tested: EBC, alcohol (v/v%), real extract (m/m%), and IBU. However, there are also other existing parameters, which are determined by international brewing factories of intermediate products during the brewing process or of the final products. In our study, besides the basic properties, we involved the total flavonoid content, monomeric anthocyanin pigment content, carbohydrate content (glucose, fructose, maltose, sucrose), and L*a*b* color of the samples. These properties are of particular interest to fruit beers because all of them are mainly influenced by the raw materials. These properties allow the samples to be distinguished more precisely and provide deeper information on different beers.

The MFA uses a weighting scheme for the variables, but this is no problem in this case, because we created the first MFA for blocks, each block representing a single variable, and a block containing measurements for the same variable. The goal is to create an average point on the MFA map, resulting in expert clustering. In a professional point of view, clusters can be perfectly explained by type of beer. This is why we also used them as a classification for supervised learning techniques. The MFA could discriminate five groups, and the other multivariate methods were performed similarly. The methods can be evaluated based on their performance features (error estimators). Based on these, a sequence of methods can be set up. Furthermore, there is a possibility to evaluate some properties based on the reliability of the estimation. The methods can be discriminated best based on the bootstrap, CV, and LOO.

The best estimation methods are the logic-based techniques (C4.5, CSMC4, CSCRT), which performed best along the flavonoid content and EBC color. It has been identified that the method most sensitive to the properties is the distance-based technique, NBC. The classification ability fluctuated greatly in the case of three properties (glucose, maltose, sucrose). A remarkable fluctuation has been experienced in the case of L*a*b* color parameters, flavonoid content, EBC color, and bitterness by NBC distance-based method.

ACKNOWLEDGEMENTS

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. This research project was supported by the Doctoral School of Food Sciences (Szent István University). The authors thank the support of VEKOP-2.3.3-15-2017-00022. The project is supported by the European Union and co-financed by the European Social Fund (grant agreement EFOP-3.6.3-VEKOP-16-2017-00005).

ORCID

Sándor Kovács https://orcid.org/0000-0002-1216-346X
László Sipos https://orcid.org/0000-0002-4584-6697

REFERENCES

1. De Gaetano G, Costanzo S, Di Castelnuovo A, et al. Effects of moderate beer consumption on health and disease: a consensus document. *Nutrition, Metabolism and Cardiovascular Diseases.* 2016;26(6):443-467. https://doi.org/10.1016/j.numecd.2016.03.007
2. Mosher M, Trantham K. *Brewing Science: A Multidisciplinary Approach*. Cham, Switzerland: Springer; 2017:38-43.
3. Moura-Nunes N, Brito TC, da Fonseca ND, et al. Phenolic compounds of Brazilian beers from different types and styles and application of chemometrics for modeling antioxidant capacity. *Food chemistry.* 2016;199:105-113. https://doi.org/10.1016/j.foodchem.2015.11.133
4. Sohrabvandi S, Mousavi SM, Razavi SH, Mortazavian AM. Application of advanced instrumental techniques for analysis of physical and physicochemical properties of beer: a review. *International Journal of Food Properties.* 2010;13(4):744-759. https://doi.org/10.1080/10942910902818145
5. Arfelli G, Sartini E. Characterisation of brewpub beer carbohydrates using high performance anion exchange chromatography coupled with pulsed amperometric detection. *Food chemistry.* 2014;142:152-158. https://doi.org/10.1016/j.foodchem.2013.07.008
6. Bello-Cerezo R, Bianconi F, Fernández A, González E, Di Maria F. Experimental comparison of color spaces for material classification. *Journal of Electronic Imaging.* 2016;25(6):1-10 061406. https://doi.org/10.1117/1.JEI.25.6.061406
7. Biancolillo A, Bucci R, Magrì AL, Magrì AD, Marini F. Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. *Analytica chimica acta.* 2014;820:23-31. https://doi.org/10.1016/j.aca.2014.02.024
8. Duarte IF, Barros A, Almeida C, Spraul M, Gil AM. Multivariate analysis of NMR and FTIR data as a potential tool for the quality control of beer. *Journal of agricultural and food chemistry.* 2004;52(5):1031-1038. https://doi.org/10.1021/jf030659z
9. Vera L, Aceña L, Guasch J, Boqué R, Mestres M, Busto O. Characterization and classification of the aroma of beer samples by means of an MS e-nose and chemometric tools. *Analytical and bioanalytical chemistry.* 2011;399(6):2073-2081. https://doi.org/10.1007/s00216-010-4343-y
10. Ghasemi-Varnamkhasti M, Forina M. NIR spectroscopy coupled with multivariate computational tools for qualitative characterization of the aging of beer. *Computers and electronics in agriculture.* 2014;100:34-40. https://doi.org/10.1016/j.compag.2013.10.001
11. Cetó X, Gutiérrez-Capitán M, Calvo D, del Valle M. Beer classification by means of a potentiometric electronic tongue. *Food chemistry.* 2013;141(3):2533-2540. https://doi.org/10.1016/j.foodchem.2013.05.091
12. Varmuza K, Steiner I, Glinsner T, Klein H. Chemometric evaluation of concentration profiles from compounds relevant in beer ageing. *Eur Food Res Technol*. 2002;215(3):235-239. https://doi.org/10.1007/s00217-002-0539-5
13. Stanimirova I, Walczak B, Massart DL. Multiple factor analysis in environmental chemistry. *Anal. Chim. Acta*. 2005;545(1):1-12. https://doi.org/10.1016/j.aca.2005.04.054

14. Granato D, Branco GF, Faria JDAF, Cruz AG. Characterization of Brazilian lager and brown ale beers based on color, phenolic compounds, and antioxidant activity using chemometrics. *Journal of the Science of Food and Agriculture.* 2011;91(3):563-571. https://doi.org/10.1002/jsfa.4222
15. Dougherty ER, Braga-Neto U. Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. *Journal of Biological Systems.* 2006;14(01):65-90. https://doi.org/10.1142/S0218339006001726
16. Dougherty ER, Sima C, Hanczar B, Braga-Neto UM. Performance of error estimators for classification. *Current Bioinformatics.* 2010;5(1): 53-67. https://doi.org/10.2174/157489310790596385
17. Beer Judge Certification Program. Style Guidelines. 2015; https://www.bjcp.org/docs/2015_Guidelines_Beer.pdf visited: 23. 06. 2019.
18. European Brewery Convention. Analytica-ebc. Getränke-Fachverlag. Nürnberg. 1998
19. Fernandes AJD, Ferreira MRA, Randau KP, de Souza TP, Soares LAL. Total flavonoids content in the raw material and aqueous extractives from Bauhinia monandra Kurz (Caesalpiniaceae). *The Scientific World Journal.* 2012;1-7.
20. Lee J, Durst RW, Wrolstad RE. Determination of total monomeric anthocyanin pigment content of fruit juices, beverages, natural colorants, and wines by the pH differential method: collaborative study. *Journal of AOAC international.* 2005;88(5):1269-1278.
21. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21(15): 3301-3307.
22. Gere A, Kovács S, Sipos L. Spectral clustering in eye-movement researches. *J. Chemom*. 2018;32(4):1-10. e3003 https://doi.org/10.1002/cem.3003
23. MacKay D. Chemometrics, econometrics, psychometrics-how best to handle hedonics? *Food Qual Prefer.* 2006;17(7–8):529-535. https://doi.org/10.1016/j.foodqual.2005.09.003
24. Chen G, Jaradat SA, Banerjee N, Tanaka TS, Ko MSH, Zhang MQ. Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Stat Sin*. 2002;12:241-262. 14
25. Hubert L, Schultz J. Quadratic assignment as a general data analysis strategy. *Br J Math Stat Psychol.* 1976;29(2):190-241. https://doi.org/10.1111/j.2044-8317.1976.tb00714.x
26. Rakotomalala, R. TANAGRA: a free software for research and academic purposes, in Proceedings of EGC'2005, RNTI-E-3, 2005;2: 697-702.
27. R Core Team. *R: A Language and Environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; 2019.

How to cite this article: Koren D, Lőrincz L, Kovács S, Kun-Farkas G, Vecseriné Hegyes B, Sipos L. Comparison of supervised learning statistical methods for classifying commercial beers and identifying patterns. *Journal of Chemometrics*. 2020;e3216. https://doi.org/10.1002/cem.3216