Analyzing Geographical Origin of Grapes and Wines of Russia

Lev Oganesyants[1,*], *Alexandr* Panasyuk[1], *Elena* Kuzmina[1], *Dmitriy* Sviridov[1], and *Alexandr* Ilyin[2]

[1]All-Russian Scientific Research Institute of Brewing, Beverage and Wine Industry – Branch of V.M. Gorbatov Federal Research Center for Food Systems of RAS, Moscow, Russia
[2]Autonomous non-profit educational organization of higher education "Skolkovo Institute of Science and Technology", Russia, Moscow

Abstract. In connection with the growing consumer's interest to Russian wines with controlled place of origin PGI and PDO, the most pressing issue is the method of their identification. One of the most effective ways to confirm the wine's place of origin in world practice is a comprehensive research of the elemental profile and isotopic characteristics of "light" elements using the methods of statistical analysis. We have selected 32 samples of fresh grapes from various wine regions of Russia (Krasnodar Territory, Republic of Crimea, Republic of Dagestan). The grape must obtained from them was fermented under laboratory conditions. In the prepared wines, the elemental profile was determined, which included 71 indicators, as well as indicators $\delta^{18}O$, δD of released ethanol and $\delta^{18}O$ of the wine water. The resulting data set was analyzed using statistical methods PCA, Permanova, the Mann-Whitney test, and machine learning was also performed. It is shown that the difference between the values of the mass concentration of the elements Al, Fe, Br, Re, U for samples from Krasnodar Territory and the Republic of Crimea are statistically significant. On the matrix of the obtained values, the Random Forest model was trained, which was able to distinguish the regions of wine origin with an accuracy of 90%. When analyzing the nonlinear dependence, the indicators of Si, Li, Co, Cu, Ba, Na, Ni, U, Al, S, Fe, Mn, B and $\delta^{18}O$ of the water were determined by the model as important.

1 Introduction

As you know, wines with a controlled regional identity are popular among consumers due to their high quality, to the geographical conditions of the grapes place of origin and strictly regulated production technology. High quality regulated wines are usually very expensive and generate high incomes for producers. However, at the same time, there is remaining danger of substitution of authentic products with wines made from other wine-making regions grapes. In this regard, the control of the authenticity of the wines geographical origin is an important and priority area, contributing to the protection of producers of high-quality products.

* Corresponding author: labvin@yandex.ru

Typically, methods for confirming the geographical origin of a wine include fingerprinting using instrumental and statistical methods, followed by the establishment of identification ranges. Today, methods of isotope mass spectrometry (IRMS), as well as the research of the elemental profile [1-3], are most widespread in confirming the wines geographical place of origin.

One of the most informative indicators in the wines research for their geographical place of origin is the $\delta^{18}O$ indicator of the water of wine. As it grows, the grape plant consumes natural meteoric and groundwater. Their isotopic characteristics are determined by a combination of various geoclimatic factors of a particular region. Thus, the oxygen isotopic composition of the water of wine will be formed based on the specificity of metabolic processes in the grape plant and the geoclimatic features of the region, including the average annual temperature, precipitation, altitude and other factors [1, 4-8].

In our earlier researches, the efficiency of using the $\delta^{18}O$ indicator of the water of wines as a marker was confirmed when confirming their geographical place of origin [9].

In addition to the research of isotopic characteristics, researches aimed at researching the elemental profile of wine have become widespread [10-14]. Most of the mineral elements found in wine, including Si, K, Ca, Fe, Mn, Rb, Sr, Ti, Ni, go through their absorption by the grape plant from the soil where the grape is grown. Their qualitative and quantitative content forms a unique mineralogical signature associated with the wine terroir. Other elements, such as As, B, Pb, Cd, Cu, Sn, S, can be either native and reflect the soil composition, or can be artificially introduced through the introduction of fertilizers or during the wine production and storage [15-17].

For the research of the elemental wine profile, the most widely used method is inductively coupled plasma mass spectrometry (ICPMS) due to its high sensitivity and accuracy. Researchers also use methods such as atomic absorption spectrometry (AAS), inductively coupled plasma atomic emission spectrometry (ICP-OES), flame atomization absorption spectrophotometry (FAAS), neutron activation analysis (NAA), X-ray fluorescence analysis (XRF).

The identification criteria for the authenticity of the wines geographical origin is based on obtaining an array of data of various indicators and their analytical processing using various statistical methods. To obtain valid results, many one-dimensional and multivariate methods of statistical analysis, including the Student's test (t-test) and the Mann-Whitney test, require data to meet a certain set of criteria, which is satisfied by data preprocessing. Among such criteria may be the sample size, normality of data distribution, the number of indicators under research and compared regions, and others. Thus, the choice of the method is most often selected depending on the characteristics of the obtained data.

In recent years, to solve applied problems, researchers have increasingly begun to use machine learning methods aimed at creating models that can solve problems not according to a predetermined algorithm, but by learning from available data. There are supervised and unsupervised machine learning method types. In the first case, the training dataset on which the model learns contains the correct answers previously obtained in a different way. In the case of unsupervised learning, the model does not have access to the correct answers. After training, the models are able to solve the problem for which they were trained. Dimensionality reduction and machine learning methods are less demanding on the distribution of data, but, as a rule, in these cases, preliminary processing is necessary. The most common data analysis methods are analysis of variance (ANOVA), principal component analysis (PCA), linear discriminant analysis (LDA), formal independent modeling of class analogs (SIMCA), support vector machine (SVM) [5,18,19].

In many countries with a developed wine industry, researches aimed at verifying the authenticity of the wines geographical origin is widespread. For Russian winemaking, such researches are especially relevant, first of all, when controlling manufacturing of the

highest quality productcategories - wines of protected geographical indications (PGI) and wines of protected designation of origin (PDO).

2 Materials and Methods

32 samples of grapes from the main wine regions of the Russian Federation (Krasnodar Territory, Republic of Dagestan, Republic of Crimea) were selected. In laboratory conditions, the must of white grape varieties was fermented, red grape varieties were fermented on the pulp. Fermentation was carried out at a temperature of 22±2 °C, using dry yeast.

The isotopic ratios $\delta^{18}O$ and δD of ethanol isolated from wines were determined using a DELTA V Advantage isotope mass spectrometer (Thermo Fisher Scientific, USA - Germany) in the EA-MS configuration, which allows the mass spectrometer to be connected to an elemental analyzer Flash HT. Using an autosampler, the sample was introduced into the elemental analyzer. The analysis was carried out in a pyrolytic reactor at a temperature of 1400 °C. The pyrolysis products were fed through the ConFlow interface to the ion source of the isotope mass spectrometer for isotopic analysis. The EA-MS configuration of the mass spectrometer was calibrated according to the VSMOW2, SLAP2, USGS47 standards.

The $\delta^{18}O$ researches of the water of the obtained wines were carried out on a Delta V Advantage isotope mass spectrometer combined with a GasBench II module. The method is based on isotopic equilibration of oxygen molecules with water, the components of the analyzed sample and the oxygen of CO_2 molecules in a CO_2-He mixture. Isotopic equilibration took place in test tubes preliminarily purged with a CO_2-He mixture and sealed with lids. Isotope equilibration took place for 20 hours in a thermostat at a temperature of 24 °C. The gas mixture with a changed isotopic composition was analyzed in an isotope ratio mass spectrometer. The values of the $\delta^{18}O$ and δD indices of the released ethanol, as well as the $\delta^{18}O$ of the water in the samples were calculated using the ISODAT software package.

In the resulting wines, the content of Li, Be, B, Na, Mg, Al, Si, P, S, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Y, Zr, Mo, Nb, Ru, Rh, Ag, Cd, In, Sn, Sb, Te, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu , Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Th, U was determined by ICP-MS (ICPMS) (X-7, Thermo Elemental, USA). The detector operates in a dual mode (pulse counting and analog); scanning mode Survey Scan and Peak Jumping; concentric sprayer - PolyCon; spray chamber - quartz, cooled (3 °C). The determination of elements in the samples was carried out by a quantitative method using standard solutions containing from 1 to 500 μg/L of the determined elements.

Principle Component Analysis (PCA) was used to visualize the data. It was carried out using python language [22] with sklearn [20] and pca [21] libraries. We used Hotelling's and dmodx's tests to remove outliers from the dataset found by PCA.

PERMANOVA is a multivariate test that allows you to compare samples not for each parameter separately, but for their totality. Skbio library [24] was used to perform PERMANOVA.

To compare 2 samples of wine samples, it was supposed to use the Student's test (t-test). However, the method has assumption that samples have a normal distribution of the same shape and the same variance. Appropriate tests were applied to each of the signs. The p-value was 0.05. The p-value threshold was selected as 0.05. For verification, we used the Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the Levene test. The pingouin library [25] was used for the Shapiro-Wilk and Levene tests, and the scipy library was used for the Kolmogorov-Smirnov test [26].

Since most of the parameters did not pass the tests, the Mann-Whitney test was taken instead the t-test as more unpretentious to data distribution. The Mann-Whitney test was performed using the pingouin library [25].

Random Forest is a supervised learning model capable of predicting observation class after training. For training, the data was split with stratification into training and test datasets. A set of models with different hyperparameters was trained, of which the best was selected using a grid search with cross-validation. The scikit-learn library was used to create Random Forest [20]

SHAP (SHapley Additive exPlanations) is a method that calculates the contribution of each of the features to the prediction of a class for observation by a machine learning model. It povides the evidence which features are important for classification. The shap [27] library in python3 was used to get the SHAP values

3 Results and Discussion

3.1 Analysis of results by PCA method

32 samples of wines were obtained from grapes selected in various wine-making regions of Russia during the grape harvesting and processing season. Of these, 12 samples from the Krasnodar Territory, 17 samples from the Republic of Crimea, 2 samples from the Republic of Dagestan. The elemental profile of samples was determined, which included 71 indicators of the mass concentration of elements, the δ18O and δD indicators of ethanol isolated from wines, as well as the δ18O indicator of the water of the resulting wines. The principal component analysis (PCA) method was used to visualize the resulting data set (Figure 1).

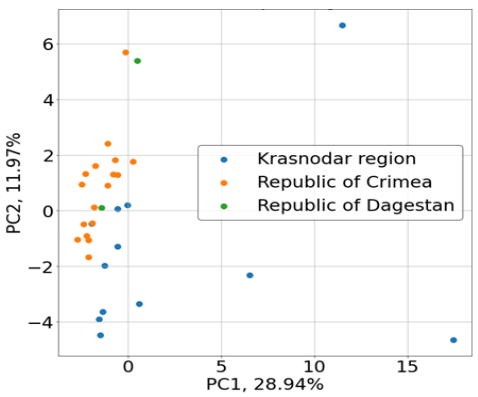

Fig.1. PCA of samples by area.

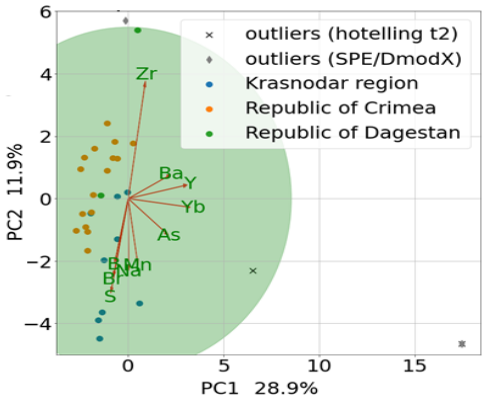

Fig.2. PCA-bioplot of samples by area.

In order to identify outliers, a PCA-bioplot diagram was built (Figure 2). The vectors represent the contribution of the features to each of the principal components. Outliers were determined using the Hoteling test and the SPE/DmodX test. It should be noted that the samples taken in the Republic of Dagestan, in terms of the values of the researched indicators, are close to the samples from other wine-making regions. Due to the small sample, it is not possible to carry out reliable statistical researches of such samples. Thus, outlier samples and samples from the Republic of Dagestan were excluded from the data set. The result is presented as a PCA diagram in Figure 3.

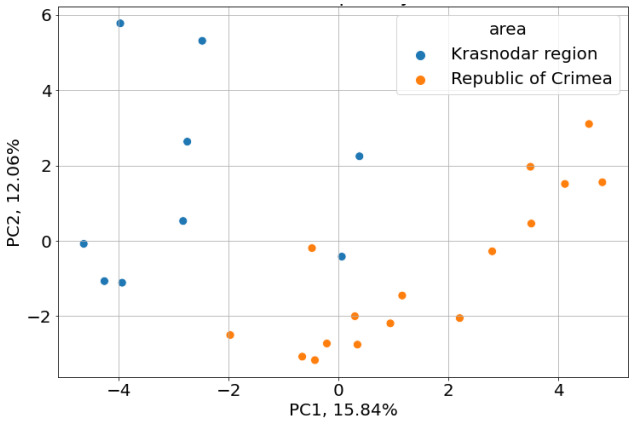

Fig.3. PCA of the remaining samples by area.

The figure shows that the samples were distributed by regions, which indicates a high probability of a statistically significant difference between the regions under research. The main division of samples from different regions into regions occurs along the abscissa axis, which indicates that the first main component (PC1) makes the greatest contribution to the statistical model. Elements B, S, V, Mn, Co, Ni, Br have the highest absolute value of the load, which indicates their importance in identifying the regional wines belonging.

Among the considered indicators, the indicators of isotopic ratios $\delta^{18}O$ and δD of ethanol and $\delta^{18}O$ of the water, reflecting the metabolic processes of the grape plant and the geoclimatic features of the region, deserve special attention. The data are presented in Figure 4.

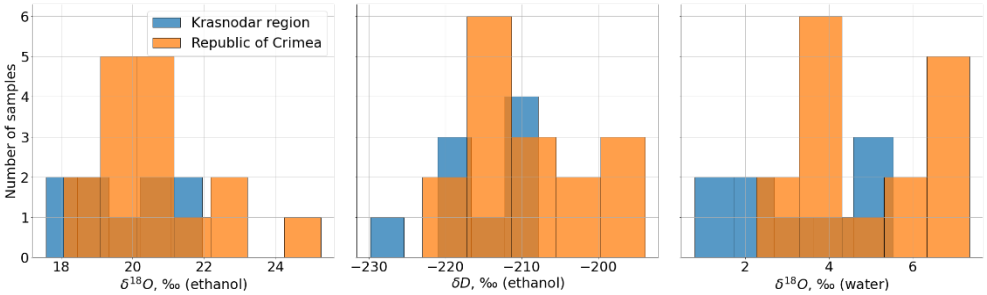

Fig.4. Distribution of isotopic ratios $\delta^{18}O$ and δD of ethanol and $\delta^{18}O$ of the water.

The figure shows that the ranges of values of the researched indicators for samples from the Krasnodar Territory and for samples from the Republic of Crimea partially coincide in all three cases. At the same time, the $\delta^{18}O$ values of the water of the samples from the Republic of Crimea differ from the $\delta^{18}O$ values of the water of the samples from the Krasnodar Territory towards a higher content of "heavy" isotopes. The average $\delta^{18}O$ values of the water for samples from Krasnodar Territory and the Republic of Crimea were 3.31‰ and 4.89‰, respectively, which can be significant with a deep statistical analysis.

3.2 Data analysis using univariate and multivariate statistical methods

In order to identify the presence of a statistically significant difference in values between the two samples (Krasnodar Territory, Republic of Crimea), the multivariate test PERMANOVA was used. The test showed the following results:

test statistic name	pseudo-F
sample size	25
number of groups	2
test statistic	3.22
p-value	0.001
number of permutations	999

As a result of the PERMANOVA test, a statistically significant difference was found between the values of the researched parameters of the two samples (p-value <0.05).

Identification of specific indicators that determine statistically significant differences between the two groups of samples were performed using univariate tests. Before univariate statistical exploration of the data, assumptions were checked using Shapiro-Wilk, Kolmogorov-Smirnov and Levene tests. As a result of the analysis, it was revealed that 50 features have non-zero variance. 11 features of these have a normal distribution, 39 features have a similar distribution shape, and 44 features have the same (equivalent) variance between the two samples. Only 8 features satisfy all the assumptions of the above tests, which makes the use of the Student's test inappropriate. In this regard, a statistical research was carried out using the Mann-Whitney test. The results are presented in Table 1, where statistic is the U value of the Mann-Whitney test statistics, RBC is the rank biserial correlation, CLES is the common language effect size. The results are shown in Table 1.

Table 1. Data analysis using the Mann-Whitney test.

Elements	statistic	p-value	RBC	CLES
Al	119	0.0084	-0.6528	0.8264
Fe	111	0.0290	-0.4306	0.7708
Br	103	0.0445	-0.4306	0.7153
Re	39	0.0428	0.4583	0.2708
U	34.5	0.0284	0.5208	0.2396

According to the test, the samples differ significantly in terms of the indicator if its pvalue is less than the threshold value of 0.05. Five indicators were identified, the values of which differ statistically significantly. BoxPlot charts of the identified indicators are presented in Figure 5.

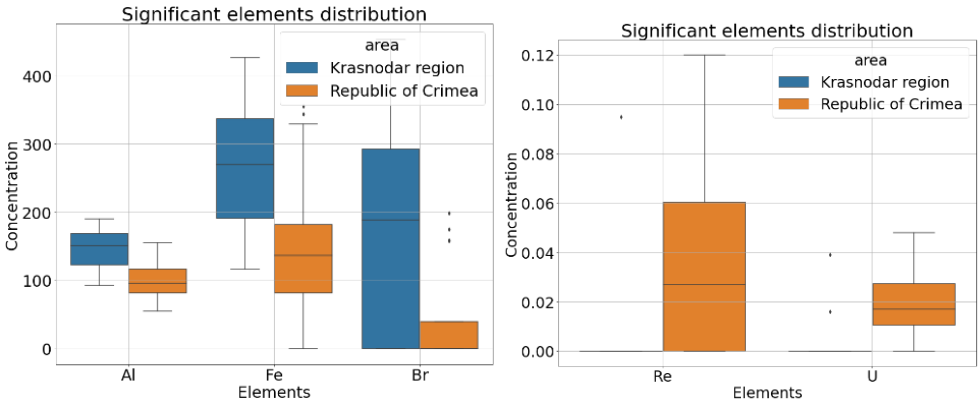

Fig.5. Mass concentration of statistically significant elements in wine samples.

The figure shows that the content of the elements Al, Fe, Br in the samples from the Krasnodar Territory is significantly higher than in the samples from the Republic of Crimea. The presence of the elements Re and U, at concentrations above the detection limit, is observed in samples from the Krasnodar Territory only in some cases. Re was found in one sample, U was found in 2 samples. In most of the samples from the Republic of Crimea, both of these elements were identified. This distribution of elements made it possible to recognize them as statistically significant.

3.3 Data analysis using machine learning method

In order to identify the nonlinear dependence of the researched indicators in wines on the geographical place of grape growing, the supervised machine learning method of Random Forest was used, which is able to predict the geographical origin of observations (class) based on the values of the variables after training. Random Forest is a collection of Decision Tree models, each of which is trained on a random subset of the data and outputs an individual set of rules for predicting the observation class.

Within the framework of this work, Random Forest was trained on a matrix of values of the concentrations of the revealed elements and the values of isotopic characteristics in wine samples, and the region of grape growing was predicted. For training, the data was split with stratification into train and test datasets. A set of models with different hyperparameters was trained, of which the best was selected using a lattice search with cross-validation in accordance with its accuracy - a parameter showing the proportion of correct model responses. As a result of the work, a model was obtained, the accuracy of which was 90%.

In addition to the resulting classifier, which is able to predict the place of wine origin, a list of important features was also obtained that were used by the model to classify samples. The results are presented in Figure 6 in the form of a SHAP diagram.

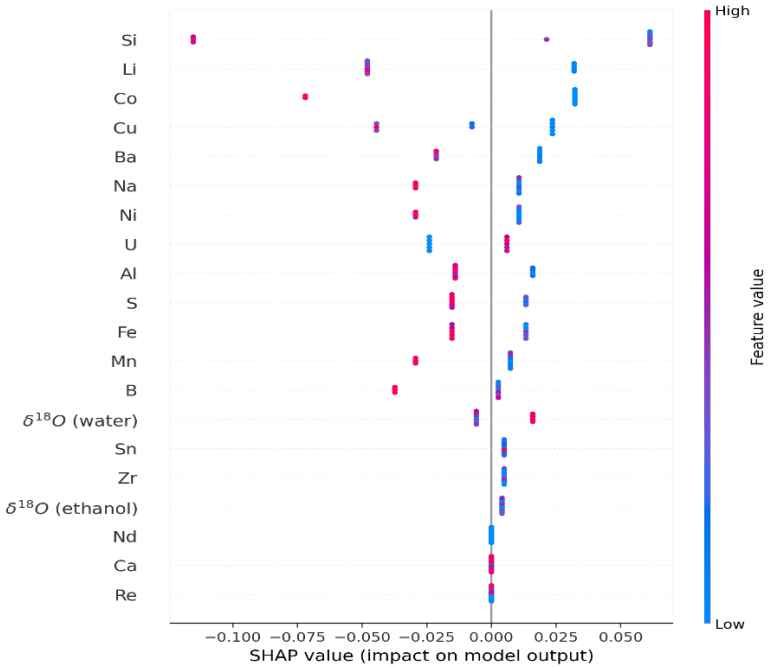

Fig.6. SHAP trait importance diagram.

On the SHAP feature diagram, the x coordinate indicates importance of the feature for model prediction, and the y coordinate lists features in order of increasing importance (less important at the bottom and more important at the top). The SHAP values of each of the samples are shown for each trait. The color reflects the value of the trait in the sample: blue - low value, red - high value. The distance of the points from zero reflects the importance of the feature for predicting the class of observations. If the feature points have a large absolute SHAP value (spaced from the center), the feature is considered important for prediction. As a result of the research, the indicators of Si, Li, Co, Cu, Ba, Na, Ni, U, Al, S, Fe, Mn, B and $\delta^{18}O$ of the water were determined as important for predicting the region of grape growing by the machine learning method.

4 Conclusion

As part of the research, new data were obtained on the values of some isotopic ratios of ethanol elements and the water, as well as the elemental profile of wines made from grapes selected in various wine-growing regions of Russia. The use of statistical analysis methods made it possible to establish the presence of significant differences between the obtained values, as well as to identify important indicators for identifying wine by the place of its geographical origin. Thus, using one-dimensional tests, the difference between the values of the mass concentration of the elements Al, Fe, Br, Re, U for samples from the Krasnodar Territory and the Republic of Crimea was found to be statistically significant. Elements B, S, V, Mn, Co, Ni, Br have been identified as important using the PCA method. On the matrix of the obtained values, the Random Forest model was trained, which was able to distinguish the regions of wine origin with an accuracy of 90%. When analyzing the nonlinear dependence, the indicators of Si, Li, Co, Cu, Ba, Na, Ni, U, Al, S, Fe, Mn, B and $\delta^{18}O$ of the water were determined by the model as important. This line of research is of particular importance in the identification of Russian wines. Annual sampling, increasing the sample and expanding its geographical coverage will allow for the confirmation of the geographical origin of the wine with high accuracy.

References

1. M. Niculaua, S. Cosofret, V.V. Cotea, Isotopes in Environmental and Health Studies **48**, 25-31, (2012)
2. L. Adami, S.V. Dutra, A.R. Marcon, Rapid Communications in Mass Spectrometry **24**, 2943-2948, (2010)
3. R. Ferrarini, G. Maria, C.F. Camin, Journal of Membrane Science **498**, 385-394, (2016)
4. C.F. Camin, N. Dordevic, R. Wehrens, Analytica Chimica Acta, **853**, 384-390, (2015)
5. L. Adami, S. V. Dutra, A. R. Marcon, Food Chemistry, **141**, 2148-2153 (2013)
6. N. Dordevic, R. Wehrens, G.J. Postma, Analytica Chimica Acta, **757**, 19-25 (2012)
7. C.F. Camin, L. Bontempo, M. Perini, Food Control **29**, 107-111 (2013)
8. D. Luo, H. Dong, H. Luo, Food Chemistry **174**, 197-201 (2015)
9. L.A. Oganesyanc, A.L., Panasyuk E.I. Kuz'mina, D.A., Sviridov, Food industry, **12**,78-80 (2020)
10. S. Orellana, A.M. Johansen, C. Gazis, Food Chemistry, (2019)

11. S.A. Drivelos, C.A. Georgiou, TrAC Trends in Analytical Chemistry **4**, 38-5,1 (2012)
12. F.D. Bora, A. Donici, T. Rusu, Notulae Botanicae Horti Agrobotanici Cluj-Napoca, 223-239, (2018)
13. E.I. Geana, C. Sandru, V. Stanciu, Food Analytical Methods, 63-73 (2017)
14. C. Durante, L. Bertacchini, L. Bontempo, Food Chemistry **210**, 648-659 (2016)
15. R.D. Paola-Naranjo, M.V. Baroni, N.S. Podio, Journal of Agricultural and Food Chemistry, 7854-7865 (2011)
16. H. Hopfer, J. Nelson, T.S. Collins, Food Chemistry, 486-496, (2015)
17. S. Pepi, C. Vaccaro, Environmental Geochemistry and Health, 833-847 (2018)
18. S. Fan, Q. Zhong, H. Gao, Journal of Food and Drug Analysis, **26**, 1033-1044 (2018)
19. S.M. Azcarate, L.D. Martinez, M. Savio, Food Control. 268-274 (2015)
20. R. Vallat, Journal of Open Source Software **3**, 1026, (2018)
21. E. Taskesen, pca, GitHub repository (2019) https://github.com/erdogant/pca.
22. F. Pedregosa, G. Varoquaux, A. Gramfort, Journal of Machine Learning Research **12**, 2825-2830 (2011)
23. S.M. Lundberg, S.I. Lee, Advances in Neural Information Processing Systems, **30** (2017)
24. The scikit-bio development team, scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers (2020) http://scikit-bio.org.
25. Python Software Foundation. Python Language Reference, version 3.8. Available at http://www.python.org.
26. M.L. Waskom, Journal of Open Source Software **60**, 3021 (2021) https://doi.org/10.21105/joss.03021.
27. P. Virtanen, R. Gommers, T.E. Oliphant, Nature Methods, **17**, 261-272 (2020)