Journal Pre-proof

A critical analysis of Adaptive Box-Cox transformation for skewed distributed data management: metabolomics of Spanish and Argentinian truffles as a case study

Leonardo Sibono, Massimiliano Grosso, Eva Tejedor-Calvo, Mattia Casula, Pedro Marco Montori, Sergi Garcia-Barreda, Cristina Manis, Pierluigi Caboni

PII: S0003-2670(25)00098-4

DOI: https://doi.org/10.1016/j.aca.2025.343704

Reference: ACA 343704

To appear in: *Analytica Chimica Acta*

Received Date: 4 April 2024

Revised Date: 20 January 2025

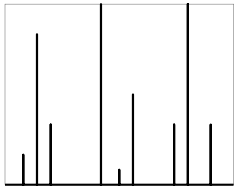
Accepted Date: 21 January 2025

Please cite this article as: L. Sibono, M. Grosso, E. Tejedor-Calvo, M. Casula, P.M. Montori, S. Garcia-Barreda, C. Manis, P. Caboni, A critical analysis of Adaptive Box-Cox transformation for skewed distributed data management: metabolomics of Spanish and Argentinian truffles as a case study, *Analytica Chimica Acta*, https://doi.org/10.1016/j.aca.2025.343704.

Argentina

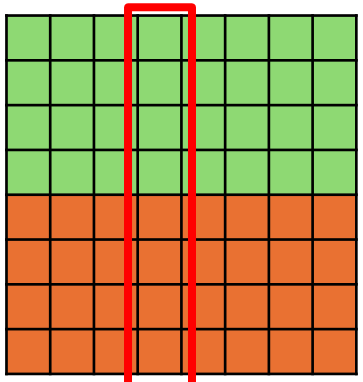
Spain

GC-MS analysis

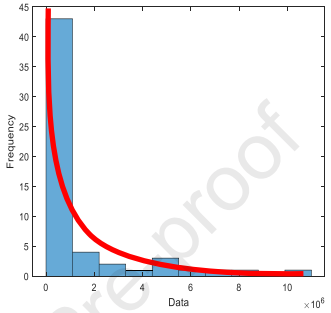


Metabolites

Samples

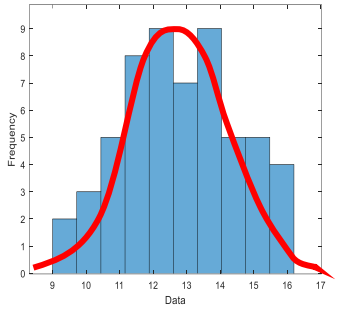


Skewed Distribution

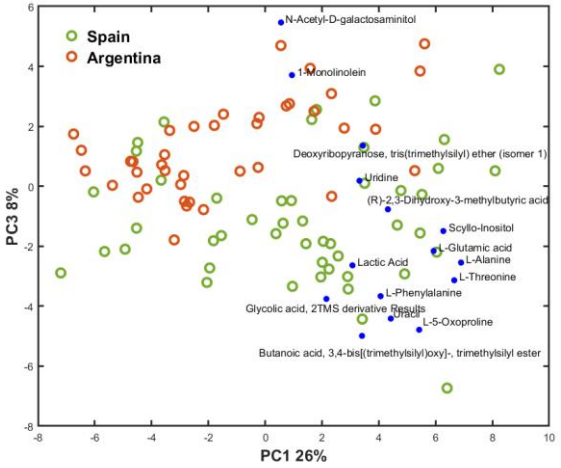


ABC TRANSFORMATION

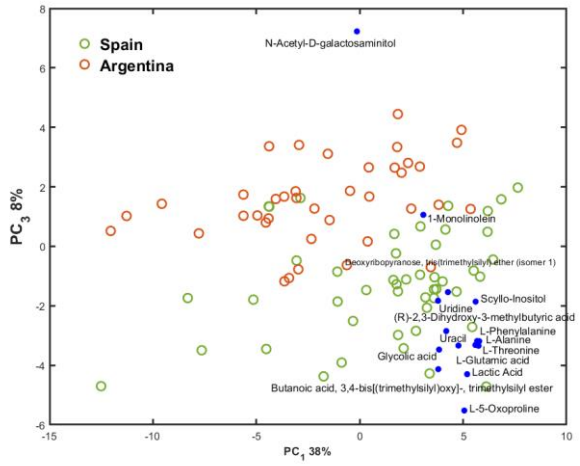
Gaussian Distribution



Acceptable separation



Enhanced separation



A critical analysis of Adaptive Box-Cox transformation for skewed distributed data management: metabolomics of Spanish and Argentinian truffles as a case study

Leonardo Sibono[1], Massimiliano Grosso[1*], Eva Tejedor-Calvo[2], Mattia Casula[3], Pedro Marco Montori[2], Sergi Garcia-Barreda[2], Cristina Manis[3], Pierluigi Caboni[3]

[1] Department of Mechanical, Chemical and Materials Engineering, Via Marengo 2, Università degli Studi di Cagliari, Cagliari, Italy

[2] Department of Plant Science, Agrifood Research and Technology Centre of Aragon (CITA), Agrifood Institute of Aragon – IA2 (CITA-Zaragoza University), Av. Montãnana, 930, 50059 Zaragoza, Spain

[3] Department of Life and Environmental Sciences, Cittadella Universitaria di Monserrato, Blocco A, Room 13, 09042, Monserrato, Italy

* email address: massimiliano.grosso@unica.it

Abstract

Background: Metabolic variations retrieved in metabolomic data are considered a benchmark for detecting biomatrix variability. Therefore, identifying target metabolites is crucial to keep track of any substrate modification and preserve it from any undesired alteration. Unfortunately, such a task can be negatively affected by detecting false positives, often triggered by complicated data distributions. In this work, we undertook an investigation of the metabolic profile of Spanish and Argentine truffles using a robust methodology. **The issue of skewed data distributions has been effectively addressed through a normalisation preprocessing, enhancing biomarker identification and samples classification.**

Results: A data normality-improved parametric test (ANOVA) was employed to define the target metabolites, which significantly vary between two regions of origin: Spain and Argentina. Specifically, Adaptive Box-Cox transformation was employed to improve the ANOVA test's performance so that data distributions were fitted to a Gaussian variable. Using the Bonferroni-Holm method for false discovery rate correction, we demonstrated the effectiveness of this transformation for the case under investigation. Results were compared with two non-parametric tests (Kruskall-Wallis and Permutation test), selected as a reference methodology, to provide a better understanding of non-normal distributions often encountered in metabolomic data analysis. 17 metabolites out of the 57 investigated metabolites exhibited notable variability across the two geographical regions. The validity of this methodology was supported through the discrimination of samples belonging to different groups. In this regard, both univariate and multivariate statistical models were tested through Monte Carlo simulations and yielded consistent results.

Significance: data analysis outcomes are sensitive to variables distributions. The present study shows an effective tool to increase data normality, thereby enhancing the statistical power for biomarker discovery and improving models' classification performances. These results find justification from the current

knowledge within the field of food sciences, enabling their application in advancing research in the truffle analysis domain.

Keywords: Metabolomics; Food; Mass spectrometry; Data preprocessing; Geographical origin; Biomarker discovery

1. Introduction

Metabolomics has gained broad importance owing to its potential to provide powerful tools for biomarker discovery and compound quantification [1], thus giving a valuable understanding of biological systems [2]. In recent years, the application of metabolomics to food technologies (foodomics) has acquired an increasing interest in food safety assessment, nutraceutical development and shelf life analysis [3,4]. Notably, understanding food metabolic profile can efficiently guide manufacturers in making decisions related to product quality, nutritional content, and flavour development.

A metabolomic study combines analytical techniques for metabolite content determination with statistical methods to extract useful information and provide data interpretation [5]. Mass Spectrometry (MS)-based analytical platform is the most popular due to its high sensitivity, selectivity and analyte coverage [6]. Indeed, it is widely recognized that using high-throughput analytical methods such as MS is imperative for any untargeted study [7]. The use of these techniques generates a vast amount of data, thereby complicating a metabolomic analysis [8]. Univariate statistics, including t-test and Analysis of Variance (ANOVA), are frequently used as multiple comparison tests to identify metabolic features that exhibit significant variations across various biological matrix stimulations. Although these methodologies offer straightforward interpretations of results, they are implemented under the assumption that the data to be analyzed follows a Gaussian distribution, a condition rarely met in real-world scenarios [9]. Accordingly, metabolomic data require pretreatment to satisfy such assumption, improve statistical power and thinning confidence intervals [10]. One approach to improve data normality is through Box-Cox Transformation (BCT) since it performs a non-linear operation on observations to reshape a skewed distribution into a more symmetrical form [11]. Recent literature has

introduced the Adaptive Box-Cox Transformation (ABC) to improve data normality using the BCT concept [12].

Although the use of BCT has been employed in several metabolomic works [12–16], it is noteworthy that such applications have investigated clinical contexts. To the authors' knowledge, the ABC method has not been investigated for metabolite identification in food-omic research. In addition, a choice criterion of normalization parameters is still a problem that is not widely addressed, resulting in a lack of understanding of the effectiveness of conventional data transformation strategies [12]. This work aims to expand the use of a feature-specific data transformation to improve data normality in target metabolite identification problems. In order to achieve this, a comparison is proposed between a non-parametric test and a parametric one. Furthermore, this study demonstrates how the transformation to normally distributed data can enhance the interpretability of results. The effectiveness of ABC in eliminating disturbances introduced by skewed data is assessed through the application of both univariate and multivariate classification models. Indeed, unlike its use in clinical studies, this work is the first to demonstrate the benefits of the ABC transformation in multivariate analysis of metabolomic data. This approach has strong potential also for industrial applications, where enhanced data handling and model accuracy are critical for product quality.

The analysis of truffles geographical origin is presented as a case study to apply the developed methodology, proposing the metabolite content of this food as a feature. Importantly, the present investigation focuses on GC-MS analysis, making it complementary to a previous work, in which the ABC transformation was demonstrated on LC-MS data with different data structures [12]. This approach presents a methodological protocol tailored to the inherent variability in food samples by identifying the appropriate data pre-processing combination and the most discriminant latent variables from multivariate models, thus discarding uninformative information.

2. Materials and methods

2.1 Truffle Sampling

Tuber melanosporum Vittad. ascocarps from Spain were harvested in various truffle orchards of Teruel, Zaragoza and Castellón provinces. Ascocarps from Argentina were collected in Espartillar (Saavedra, Buenos Aires, Argentina). Truffles were collected between 2020/2021 and 2021/2022 season. All truffles selected were characterized by the typical *T. melanosporum* aroma and the maturity stage of the truffles was assessed with a gleba sample reaching 5-10 mm under the peridium, taken with a scalpel. With this sample, a spore maturity index was calculated as the percentage of asci containing mature (i. e. dark brown and spiny) spores, following [17]. Fresh truffles were identified, selected, and processed [18]. Truffles were lyophilized (LyoBeta 15 lyophilizer, Telstar, Madrid, condenser temperature was -80 ºC), ground, mixed and sieved until a particle size lower than 0.5 mm was obtained. Powdered truffles were kept at −80 ºC until further use.

2.2 Sample preparation for GC-MS analysis

To extract 86 truffle samples the modified Folch method was adopted [19]. Briefly, 20 mg of each truffle were weighted inside 1.5 ml Eppendorf tube. 1 mL of a methanol-chloroform solution (2/1, v/v) containing 5 mg/L of 2,2,3,3-D4-succinic acid used ~~as~~ the internal standard and 90 µL of KCl 0.2M were added. Then, samples were first vortexed and then ultrasonicated 3 mins for 3 times. To break down the cell wall of the truffles effectively, the samples were stored at -20 °C overnight, and the next day they were sonicated again for 3 minutes and repeated 3 times. The obtained solution was centrifuged at 17700 rcf for 10 min, and 400 µL of aqueous layer were transferred into glass vial and dried under a nitrogen stream. The dried aqueous layer was derivatized using 70 µL of N-methyl-N-

(trimethylsilyl) trifluoroacetamide and heating the samples at 70°C for 1 hour. After, 930 µL of hexane were added and the samples were vortexed again before GC-MS analysis.

2.3 GC-MS Analysis

Qualitative analysis of aqueous truffles extracts was performed by an Agilent 8890 gas chromatograph equipped with an Agilent 7693 A autosampler, fitted with an Agilent 5977B single quadrupole mass detector (Agilent, Palo Alto, CA, USA). Chromatographic separations were performed on a HP-5MS capillary column: 30 m x 0.25 mm (i.d.) with a film thickness of 0.25 µm (Agilent J & W GC column, Palo Alto, CA, USA). The front inlet temperature was 200 °C and helium gas was used as the GC carrier gas. The temperature program was as follows: 1 min of isothermal heating at 70°C, which was then increased to 260 at 10°C/min and held for 2 min, to 280°C at 30°C/min and held for 15 min and finally to 330°C at 50°C/min and held for 5 min. The transfer line and the ion source temperatures were 280 and 180°C, respectively. Ions have been generated at 70 eV with electron ionization and a dwell time at 1.6 scans/s.

2.4 Data Overview and Tools

The metabolic dataset consists of 86 truffle samples, of which 47 were collected in Spain and 39 in Argentina. In all the samples, 57 metabolites were annotated after GC-MS. Data analysis was carried out on Matlab ® 2022b environment, and "Milano Chemometrics" toolbox (version 6.0) [20].

2.5 Statistical analysis

Truffle data from Spain and Argentina were primarily explored through descriptive statistics analysis specific to each variable (i.e. metabolite GC-MS peak). Initial examination of data frequencies revealed that their dispersion significantly deviated from the Gaussian assumption, resulting in highly skewed

distributions. Applying traditional hypothesis tests, such as the ANOVA test, may be deemed inappropriate in cases where the data are not Gaussian. Two distinct approaches were used in this study to address this issue. The first involved the application of two non-parametric tests:

1) The Kruskal-Wallis (KW) test to the raw data for each metabolite in order to identify the metabolites whose GC-MS peak median value significantly varied between the two distinct datasets corresponding to Argentina and Spain. This test is non-parametric, so it does not require any assumptions about data distribution, and it is robust to outliers [21,22].
2) The permutation test to evaluate whether the observed difference between the means from two groups (e.g., geographical origin) is statistically significant by randomly shuffling group labels and recalculating mean differences across 10^5 iterations. This generates a distribution of groups' average values differences under the null hypothesis (difference equal to zero), allowing comparison of the observed difference between the means from two classes to assess the p-value, which corresponds to the proportion of permuted samples (re-randomized mean differences) that show a mean difference equal to or greater than the observed difference between the groups [23].

In the second procedure, data is preprocessed using the ABC method [12,24]:

$$z_i(\lambda) = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases} \qquad i = 1, \dots, n \tag{1}$$

In Eq.1, λ is the power transformation parameter, x_i refers to the i-th metabolite raw data, z_i refers to the transformed one, and n is the sample size. When λ = 0, such operation degenerates to log-normal transformation. The purpose of this transformation was to normalize the data so that traditional parametric tests such as one-way ANOVA could be used to analyze it. The selection of the optimal λ value is crucial for proper data transformation. The procedure of the ABC transformation proposed by

[12] is thus employed. Briefly, to see how the BCT data were correlated to the Gaussian distribution, a normality Lilliefors hypothesis test [25] was conducted for each λ between -3 and 3 with an increment of 0.01. This procedure was performed individually for each metabolite. The null hypothesis of the Lilliefors test states that the data conforms to a Gaussian distribution, while the alternative hypothesis contemplates non-compliance with Gaussian behaviour. When λ changes, the Lilliefors test provides two different p-values for the two groups. The optimal λ for a given metabolite was selected as the one associated with the highest product of the p-values, representing the condition least prone to rejecting the null hypothesis and consequently closer to Gaussian transformation [12]:

$$\lambda_{opt,k} = \underset{\lambda_j \in [-3,3]}{\operatorname{argmax}}(p_{arg}(\lambda_j) \cdot p_{sp}(\lambda_j)) \quad (2)$$

In Eq. 2 $\lambda_{opt,k}$ is the optimal λ value for the k-th metabolite, $p_{arg}(\lambda_j)$ and $p_{sp}(\lambda_j)$ indicate the corresponding p-values derived from the Lilliefors test on Argentina's and Spain's samples, as λ_j varies. The selection criterion for the optimal λ followed the approach reported by Yu et al. [12]. However, it differs from the same study in that we employed the Lilliefors test for normality, which is more appropriate when the sample size exceeds 50 [26].

On transformed data, an ANOVA test was conducted in order to determine whether the metabolites showed significant differences based on the geographical origin of the samples. The significance level was set equal to 5%.

Both procedures were followed in order to correct false discovery rates. The Bonferroni-Holm test was used to set the false discovery rate at 1% after both procedures had been completed [27]. Metabolites that exhibited significant variation under both procedures were selected as target metabolites for geographical origin.

Figure 1 depicts the algorithm structure. In order to highlight the benefit of ABC transformation, a comparison between the results obtained from the second procedure and the ANOVA test conducted on raw data was performed. Target metabolites were individually tested through a Montecarlo simulation ANOVA test [28]. Spain and Argentina samples are denoted as class 1 and class 2, respectively. The methodology is reported in the following step procedure:

1. Randomly sample 50% of observations for both classes.
2. Perform the ANOVA test between the sampled and non-sampled class 1 observations. A false positive is detected if the average value is significantly different (p-value < 0.05). The same step is repeated for class 2 observations.
3. Perform the ANOVA test between the sampled class 1 and non-sampled class 2 observations. If p-value is > 0.05, a false negative is detected.
4. Repeat steps 1 to 3 for 500 times for each metabolite.
5. Calculate the False Positive Rate (FPR), the False Negative Rate (FNR) and the Accuracy.

The idea behind such a procedure is that if the target metabolites are highly discriminative between the classes, then the ABC transformation with ANOVA test can be considered effective. A comparison between parametric ABC-ANOVA and non-parametric tests' p-values might help at understanding the benefit of data transformation.

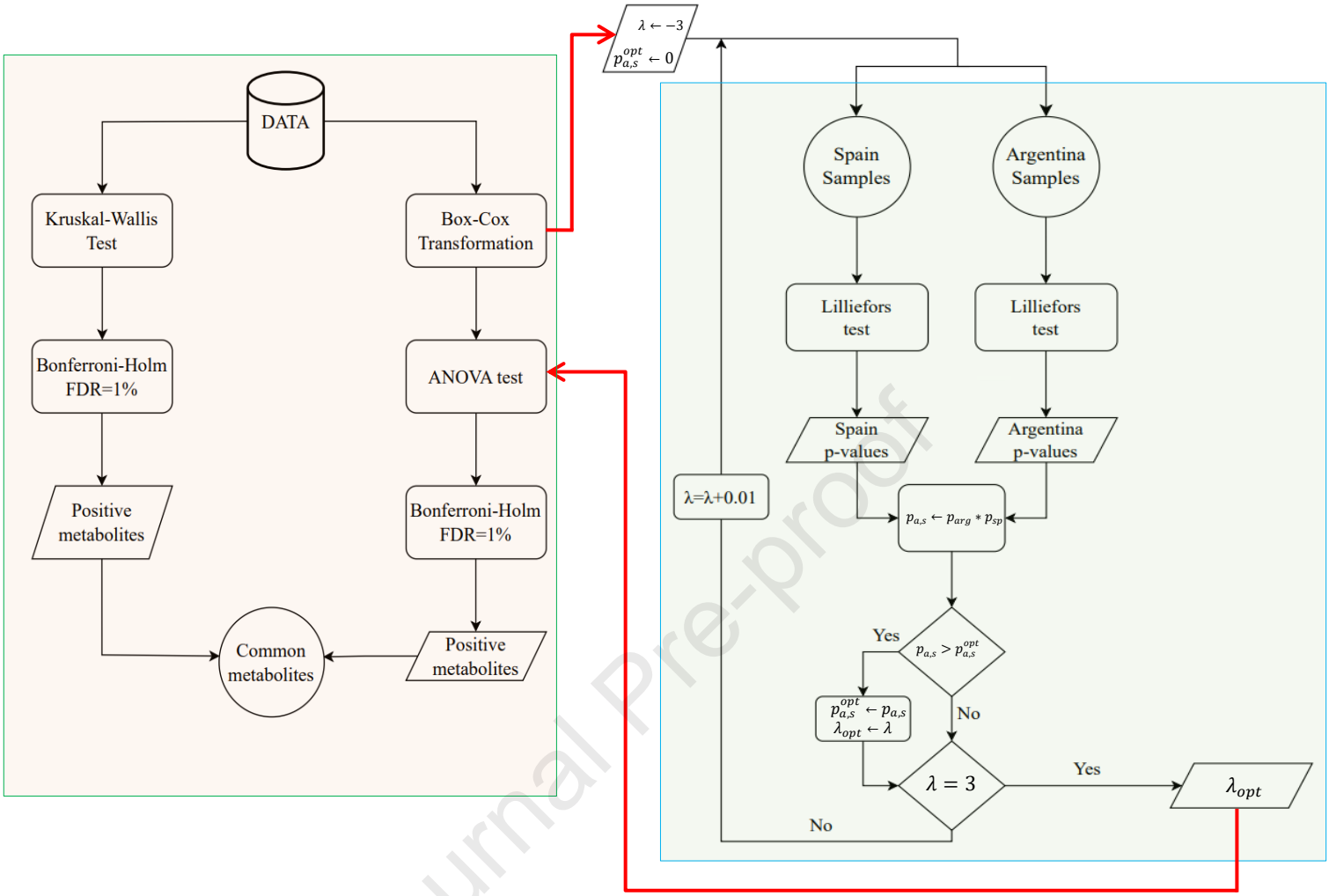


Figure 1. Algorithm structure for univariate analysis.

Subsequently, a PCA-LDA was employed on both the ABC-transformed and non-transformed samples to classify the samples originating from the two distinct geographical regions and assess the impact of the transformation on the quality of separation in the principal component space. In both cases, data were preprocessed through Unity Variance (UV) operation. The classification accuracy was assessed by evaluating the error rate in a 5-fold cross-validation. Subsequently, examining the loadings' values of the PCA model enabled the verification of the consistency between the target metabolites identified through the univariate procedure and those obtained through the multivariate approach. Finally, a PLS-

DA model was employed to further assess the benefits of the ABC transformation on a multivariate classifier. The same metrics as in the PCA-LDA model were adopted [20].

3. Results and discussion

Detection of significant differences between the two classes of metabolites has been accomplished by non-parametric tests (Kruskal-Wallis and Permutation test) on the raw data and parametric (ANOVA) tests on the ABC transformed data [9,12,13]. Moreover, the impact of data transformation on Principal Component Analysis is reported in this work to evaluate the effect of the ABC transformation from a multivariate perspective.

3.1 Data preprocessing: effect of the ABC transformation

Generally, the employment of a normalization step was highly recommended, as 80% of raw data metabolites were strongly asymmetric (abs(skewness) >0.5). Figure 2 depicts the effect of data transformation on their normality behaviour.

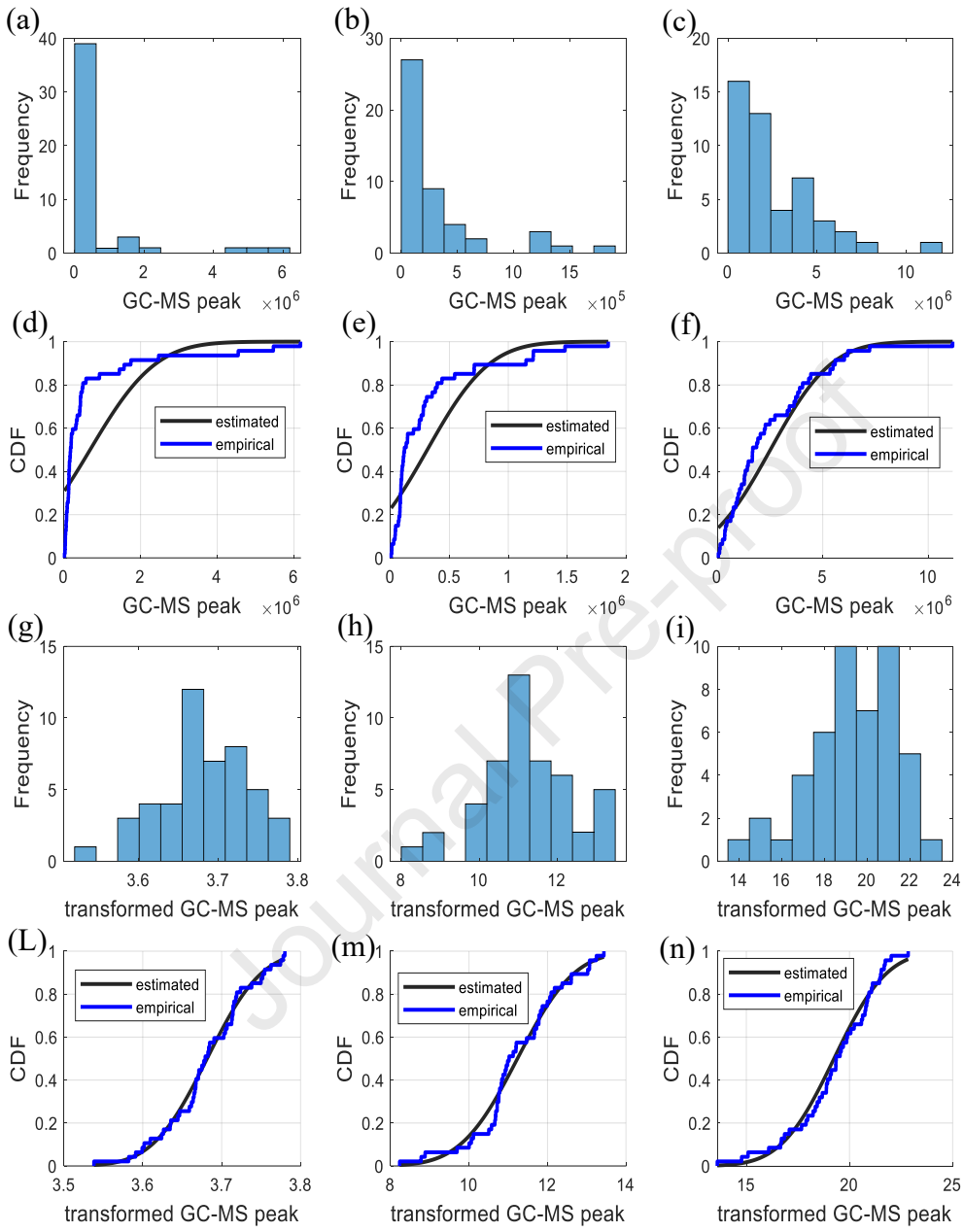


Figure 2. GC-MS frequency highlights the non-gaussian behaviour of metabolomic data. Frequency distribution of raw data from Spain for lactic acid, glycolic acid, and L-leucine (a-c). Comparative plots show empirical and estimated CDFs using mean and standard deviation for such three metabolites (d-f). Transformed data frequencies (g.i) and resulting Box-Cox CDFs (L-n) are presented for the same metabolites' GC-MS peaks.

By way of example, sample frequencies of GC-MS peaks from samples originating from Spain for three distinct metabolites (lactic acid, glycolic acid, and L-leucine) are displayed in (Fig. 2,a-c). Notably, the distributions for all three metabolites align closely with a log-normal distribution. Specifically, employing the ABC algorithm, optimal λ values close to 0 are obtained, indicating that a logarithmic transformation of the data produces variables exhibiting a Gaussian-like behaviour (Fig. 2, g-i). The effect of the nonlinear transformation becomes evident when comparing raw data Cumulative Distribution Functions (CDF) (Fig. 2, d-f) with transformed data CDF (Fig. 2, l-n), where the estimated CDF is that derived from a Gaussian function using sample means and variances. As can be seen, the estimation curve adequately fits the empirical cumulative frequency when involving transformed data. The optimal λ values were -0.26 for lactic acid, -0.01 for glycolic acid, and 0.04 for L-leucine. It appears that the distributions of the raw data for the second and third metabolites have a greater affinity for a log-normal distribution than those of the lactic acid. It is important to stress that the same λ values were applied to Argentina samples for the corresponding metabolite. A total of 19% of metabolites exhibited an optimal λ in absolute value smaller than 0.15, suggesting a reliable adherence to a log-normal distribution of the original data. The results of the elaboration of all power parameter values are given in Appendix A.

Figure 3 illustrates the distribution of p-values from Lilliefors normality test before and after the BCT for the different metabolites associated with a specific geographic region to provide an overall perspective of the effect of the BCT. Specifically, when raw data were evaluated, 38 and 47 metabolites out of 57 revealed a Lilliefors p-value<0.1 for Spain and Argentina, respectively. This result indicates a considerable deviation from Gaussian behaviour [12]. When data were transformed, only 11 metabolites were positive for Spain's normality test and 10 for Argentina.

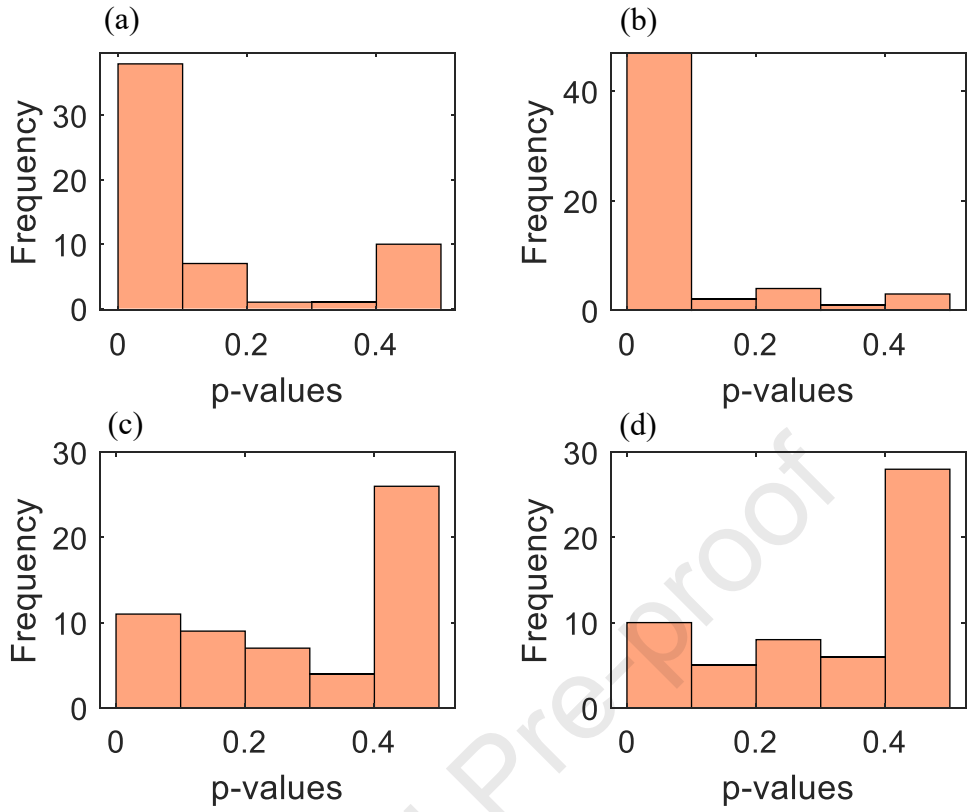


Figure 3. Results from Lilliefors normality test. p-values distribution obtained for Spain before (a) and after (c) transformation. p-values distribution obtained for Argentina before (b) and after (d) transformation

3.2 Identification of target metabolites

When only raw data were used, 31 metabolites resulted in significant variation between the two regions from the Kruskal-Wallis, Permutation and ANOVA tests. When Bonferroni-Holm with 1% FDR correction was applied on such 31 p-values, 17 tests were concluded as positives for Kruskal-Wallis p-values, 13 for Permutation test, while only 9 significant p-values were observed from ANOVA. All metabolites with significant p-values in the raw-data ANOVA and Permutation tests also showed significant p-values in the Kruskal-Wallis test. However, the employment of data transformation can potentially help identify additional metabolites, thereby enhancing the understanding of the biological matrix and the accuracy of a univariate model in discerning foods produced in distinct regions. Performing ANOVA on the raw data allows for a direct comparison of the statistical power before and

after applying the ABC transformation. This approach highlights the benefits of the ABC transformation, particularly in enhancing statistical power. Indeed, when ANOVA was performed on Box-Cox transformed data and Bonferroni-Holm correction was applied, 15 metabolites were reported as significant, all detected by the non-parametric method. However, when Kruskal-Wallis and Permutation tests were applied, lower p-values were generally obtained, meaning that the use of parametric tests when Gaussian distribution condition is met after data normalization improved the statistical difference between the geographical groups. Results are reported in Table 1. For the sake of simplicity, the larger set of metabolites identified as the target is reported for all three techniques. As can be seen, the normality transformation resulted in a clear decrease in p-value for all the presented metabolites except for the L-5-Oxoproline metabolite. It was indeed observed that the data distributions for this metabolite were remarkably different for the two regions. In detail, a uniform distribution for samples from Spain approximately describes this feature (the skewness value was equal to -8 e-04), whereas samples from Argentina show a skewness value equal to 1.29. As a result, the algorithm struggles to find a reasonable balance between the two datasets, leading to a loss of statistical power.

The decision to employ a stringent FDR correction method highlighted the behaviour of ANOVA p-values when data were transformed. Although some tests were still positive when raw data were tested, the application of the ABC transformation markedly enlarged the differences between the position indices associated with the two distinct classes. Consequently, a notable reduction in the resulting p-values was observed when Box-Cox transformed data were tested through ANOVA, thus making the statistical test outcomes more evident. For example, when the ANOVA test was performed on the transformed data, the p-values for lactic and glycolic acid decreased by four and five orders of magnitude when compared to the corresponding raw data. The benefit of such an operation is here underlined through validation techniques.

Table 1. P-values computed for each of the 17 metabolites classified as significant and identified through non-parametric analysis. Additionally, p-values obtained from ANOVA tests on both raw and ABC-transformed data are provided for these metabolites.

		Permutation test	Kruskall-Wallis	ANOVA on raw data	ANOVA on transformed data
Metabolite	Regulation	p-value	p-value	p-value	p-value
Lactic Acid, 2TMS	S	4.00 e-04*	6.83 e-07*	0.0061	5.35 e-07*
Glycolic acid, 2TMS	S	1.00 e-05*	4.51 e-08*	1.83 e-04*	3.17 e-09*
L-Alanine, 2TMS	S	3.00 e-05*	3.85 e-05*	1.74 e-05*	1.57 e-05*
Uracil, 2TMS	S	1.00 e-05*	1.96 e-06*	9.23 e-06*	2.24 e-06*
L-Threonine, 3TMS	S	2.00 e-05*	2.17 e-05*	7.19 e-06*	6.86 e-06*
2,3-Dihydroxy-3-methylbutyric acid, 3TMS	S	3.20 e-04*	4.96 e-04*	4.75 e-04*	2.38 e-04*
Butanoic acid, 3,4-dihydroxy	S	1.00 e-05*	7.79 e-09*	9.43 e-07*	2.31 e-10*
L-5-Oxoproline, 2TMS	S	2.00 e-05*	6.87 e-06*	1.42 e-06*	1.68 e-06*
L-Glutamic acid, 3TMS	S	1.00 e-05*	1.72 e-06*	7.72 e-06*	2.05 e-07*
L-Phenylalanine, 2TMS	S	5.00 e-05*	1.03 e-05*	0.0024	4.84 e-06*

Deoxyribopyranose, 3TMS	S	1.00 e-05*	2.50 e-08*	1.28 e-05*	4.29 e-09*
Scyllo-Inositol, 6TMS	S	6.50 e-04	4.50 e-04*	6.68 e-04	2.88 e-04*
D-Allofuranose, 5TMS	S	0.0014	4.96 e-04*	0.0012	5.87 e-04
N-Acetyl-D-galactosaminitol, 5TMS	A	0.0227	4.15 e-05*	0.0251	2.29 e-05*
9,12-Octadecadienoic acid, TMS	S	0.0027	1.47 e-05*	0.0029	6.30 e-05
Uridine, 3TMS	S	1.00 e-05*	1.03 e-07*	2.50 e-07*	2.21 e-08*
1-Monolinolein, 2TMS	S	2.60 e-04*	4.74 e-06*	0.0011	4.45 e-06*

* *The metabolite content significantly changes between the two classes after FDR correction. S label denotes metabolites upregulated in Spanish truffles, A indicates their upregulation in the Argentina ones.*

As it can be seen, the compounds which resulted to significantly change between the two geographical origins constitute a heterogeneous pool of organic molecules encompassing fatty acids, amino acids, saccharides, alcohols and nucleotides. Several studies have previously focused the attention on the effect of different stimuli on truffle metabolic profile [4,29–32]. Despite the fact that the knowledge in this field is mainly confined to the analysis of volatile organic compounds, the results reported in previous works can corroborate the outcomes reported in this study. Indeed the upregulation of amino

acids like Alanine, Threonine, Glutamic acid and Phenylalanine or fatty acids like 9,12-Octadecadienoic acid in Spanish truffles might suggest different levels of lipolytic and proteolytic activity of microorganisms associated with those samples [4]. Moreover, a previous study demonstrated an up-regulation for Argentina truffles of several microbial volatile organic compounds such as 2-Methyl-1-butanol, 3-Methyl-1-butanol and 2-Butanone which are ones of the main contributors of the characteristic aromatic profile. On the other hand, these compounds were down-regulated in the Spanish counterpart, probably due to the soil and climate variation [29]. These differences may justify a lower content of heavier organic compounds (e.g. 2,3-Dihydroxy-3-methylbutyric acid and Butanoic acid, 3,4-dihydroxy) in Argentina samples investigated in this study, as a results of a greater extent of their degradation.

3.3 Classification of truffles through ANOVA test

Many metabolomic works employed ANOVA as a feature selection method [33–35], useful to both identify target metabolites and apply a filtering method aimed to simplify subsequent analysis (e.g. multivariate). In this study, the reliability of the ANOVA test was studied through Monte Carlo simulations to assess whether a metabolite identified as significant can be accurately used as a biomarker for the geographical origin of truffle samples. Table 2 shows the results of the 5-step procedure applied to raw, and Box-Cox transformed data. As expected, the FPR for all the metabolites is close to the significance level for the test (i.e., 5%), but the occurrence of false negatives was also considered. Several metabolites showed better performances when ANOVA was conducted on transformed data, while no or low differences were appreciated for others. In particular, lactic acid, N-Acetyl-D-galactosaminitol and 9,12-Octadecadienoic acid underwent a remarkable benefit in terms of accuracy when data were transformed. In view of this, it is important to emphasize that the ABC transformation can reduce the FNR when univariate methods are used to classify samples.

Table 2. Performance metrics for the ANOVA test applied on both raw and ABC transformed data for the metabolites identified as target.

	ANOVA on raw data			ANOVA on transformed data		
Metabolite	FPR %	FNR %	Accuracy %	FPR %	FNR %	Accuracy %
Lactic Acid 1	4.8	42.7	76.3	4.3	0.2	97.8
Glycolic acid 2	3.2	0.7	98.1	5.3	0	97.4
L-Alanine 3	5	5.8	94.6	5.3	5.8	94.5
Uracil 13	3.7	0.1	98.1	3.5	0.9	97.8
L-Threonine 16	5	1.5	96.8	6.0	1.9	96.1
(R)-2,3-Dihydroxy-3-methylbutyric acid 17	4.4	17.8	88.9	4.7	15.2	90.1
Butanoic acid, 3,4-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester 18	3.7	0	98.15	4.9	0	97.6
L-5-Oxoproline 20	4.6	0.7	97.4	3.9	1.6	97.3
L-Glutamic acid 23	4.8	0.9	97.15	4.5	0.4	97.6
L-Phenylalanine 24	3.2	19	88.9	5.2	1.9	96.5
Deoxyribopyranose, 5TMS 26	4.1	0.1	97.9	5.1	0	97.5
Scyllo-Inositol 37	4.7	25	85.2	3.6	18.9	88.8
D-Allofuranose, 5TMS 38	4.5	33.3	81.1	4.1	26.3	84.8

N-Acetyl-D-galactosaminitol 42	3.2	67.1	64.9	4.4	4.5	95.6
9,12-Octadecadienoic acid- 43	4.3	43.9	75.9	3.7	8.5	93.9
Uridine 47	4.6	0	97.7	5.8	0.2	97
1-Monolinolein 53	4.1	17.5	88.7	5.6	2.4	96

3.4 Multivariate analysis

The impact of data transformation on multivariate statistical analyses, such as PCA, was further assessed. It was observed that the application of the ABC transformation changed the overall data structure. Specifically, the separation between the two class samples became more pronounced when moving from raw data to transformed data, as evidenced in the biplot showing the first and the third PCs score values and the metabolites' loadings (Figure 4). Accordingly, 50.90% of data variability was captured by the first 3 PCs when raw data were used, whereas 60.45% of the total variance was explained by the same number of PCs when ABC transformed data were analysed. The score plot's display concerning PC1 and PC2 is not reported as it does not provide clear information regarding the separation between samples associated with different classes. Indeed, the most explanatory PCs may not coincide with largest PCs due to uncontrolled variability sources which cover the information of interest in correspondence of earlier PCs [36], as displayed in figure S1 reported in Supplementary Materials. As it can be seen, ABC transformation had a positive impact on PCA results, as the separation between PC1 and PC3 scores related to different geographical regions was improved. Similar conclusions were drawn by [12] in their metabolomic study of COVID-19 severity in different subjects [37], where BCT was employed as a preprocessing tool for image enhancement. Generally, it

is recognized that BCT is efficient at improving class separability. Indeed, this technique performs a non-linear normalization of the feature space, amplifying higher-value data to a greater extent than lower-value points [38]. As a result, it conveniently reshapes the principal component space for a classificatory model.

Interestingly, the discriminative information among samples from distinct geographical regions is mainly provided from the third principal component since this direction shows an evident class separation (Figure 4). As can be seen, most of the target metabolites identified by the procedure employed in this study exhibit upregulation in the samples from Spain, except N-Acetyl-D-galactosaminitol. Notably, such estimated overexpression of these metabolites aligns consistently with the results derived from examining the median values associated with individual metabolites, except for 1-Monolinein, for which an upregulation in Spain truffles was observed from univariate analysis. At the same time, no evident conclusions can be drawn from its loading analysis. It is worth noting that one can draw the same conclusion regarding the PC loadings analysis, regardless of the application ABC transformation with the exception of 1-Monolinein, for which an upregulation in Argentinian truffles was observed when raw data where used, while no clear geographical dependence can be inferred when ABC transformed where subjected to PCA.

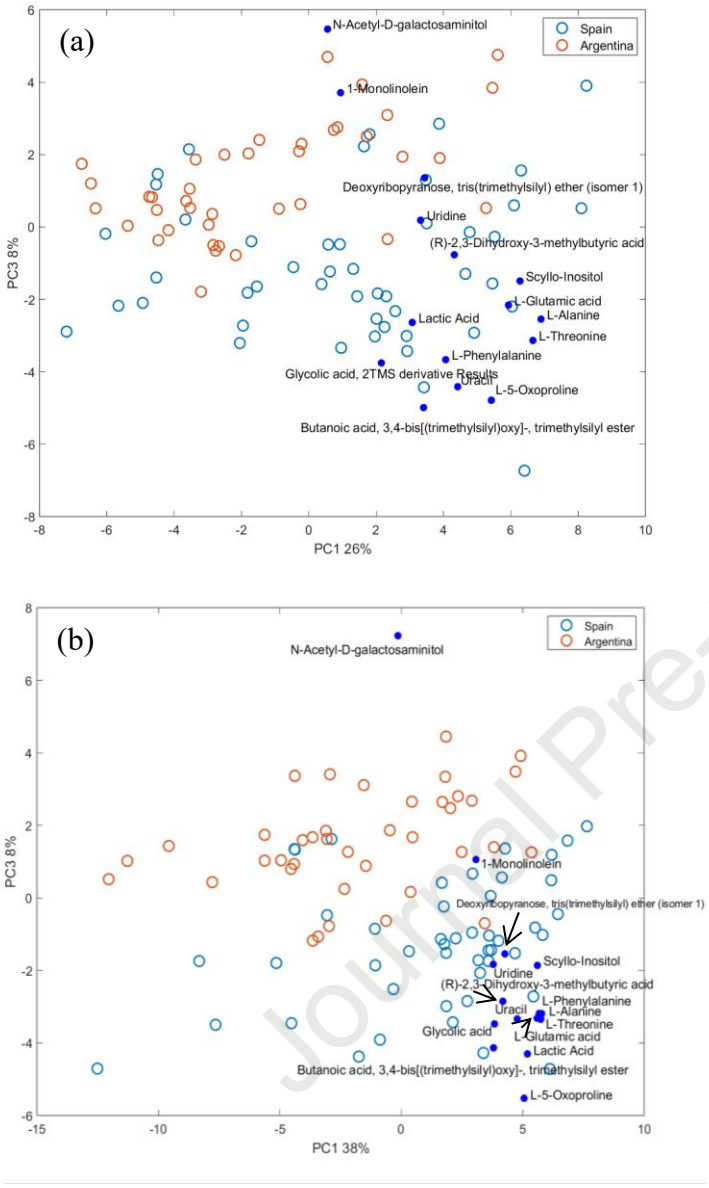


Figure 4. PCA biplot: scores related to truffles from Spain and Argentina assessed for raw data (a) and Box-Cox transformed data (b). The blue filled dot represents the loading value. For the sake of representation clarity, only the loadings corresponding to the metabolites targeted as significant are reported in the figure.

Finally, PCA-LDA was employed on the samples to classify them according to their geographical origin. It is important to highlight the behaviour of the error rate in 5-fold cross-validation when a different number of PCs is introduced in the PCA model. It is observed that classification quality

improves when classifying data subjected to BCT followed by UV preprocessing, compared to those subjected only to UV preprocessing (Figure 5a). Indeed, the error rate curve associated with BCT-standardized data consistently stands below that of the raw standardized data, indicating a superior performance of PCA-LDA model on transformed data. In any case, the multivariate model achieves good classification performance: the elbow rule suggests an optimal number of principal components equal to six, corresponding to a Cross-Validation accuracy of 95%. Noteworthy, a sharp drop in error rate is observed when moving from 2 PCs to 3 PCs as a result of the scores' overall high separability along the third PC. It is important to highlight that Cross-Validation results are predominantly driven by the training data set size [39] since the error rate increases as the number of samples in the training data set decreases. Therefore, a robust validation analysis was conducted through Monte Carlo Cross-Validation. In this procedure 80% of the dataset is randomly partitioned for calibration, while the remaining observations constituted the validation set [40]. This process was iterated 1000 times, and the model performance was assessed through the average accuracy of classification on validation samples. Six principal components, previously determined as optimal, were selected for the analysis. In conclusion, 87% and 96% accuracy were achieved for raw data, and ABC transformed data using the same number of PCs (i.e. six), meaning that the classification performances are robust to different training set sizes. For a comprehensive multivariate analysis, PLS-DA was also employed to further confirm the transformation's benefits (Figure 5b). Similarly, a better performance was obtained for ABC-UV data, especially when smaller latent space dimensions were contemplated in the PLS-DA model. From the error rate analysis, it can be concluded that 2 LVs are sufficient to reach high classification performances from both UV and UV+ABC preprocessing. Remarkably, an accuracy of 100% was attained for four latent variables through all the 1000 Monte Carlo simulations, while 95% was obtained from raw data validation analysis. Again, the benefit of normalizing metabolites data distribution is emphasized. For completeness, additional combinations preprocessing methods were

applied prior to conducting the multivariate analysis. The resulting classification accuracies from non-ABC-transformed data were generally lower than those achieved using ABC-transformed data. The results are presented in Table S1 of supplementary materials. Such classification accuracy improves along with the ones obtained in other studies where similar gains were considered significant such as the classification trees proposed by Lubinsky [41], or the ensembled learning models proposed by Dhamayanthi and Lavanya [42], who obtained an improvement of 4.2%.

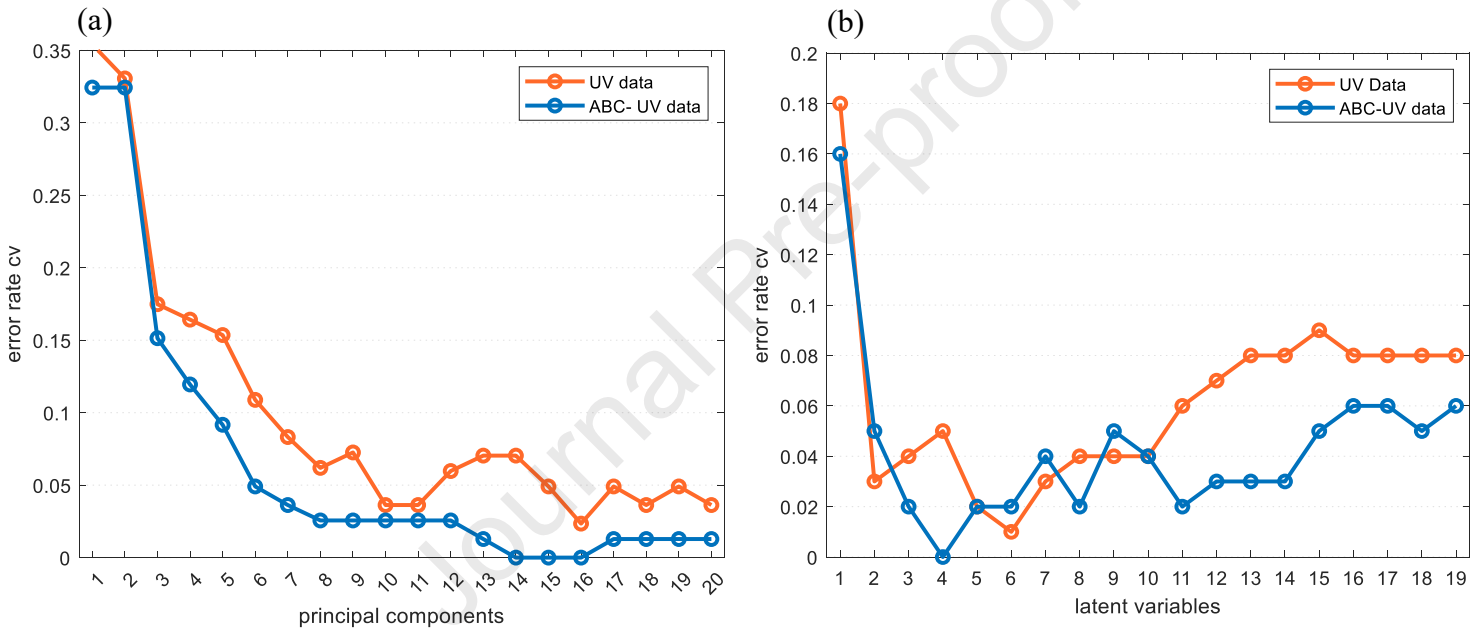


Figure 5. Error rate in a 5-fold Cross-Validation as a function of the number of the principal components included in the PCA-DA model (a) and the number of latent variables included in the PLS-DA model (b).

The VIP analysis was conducted to enable the assessment of each metabolite's contribution to the projection of observations within the latent space. This approach provides an additional validation of the identification process highlighted by the univariate analysis and allowed for a further assessment of the outcomes obtained from PLS-DA performed on both raw and ABC-transformed data. In the present analysis, only those metabolites having VIP

values greater than one were considered as discriminant between the classes. Two latent variables were selected for VIP values calculation, as an accuracy of at least 95% was achieved in both cases (Figure 5b). For each metabolite, Table 3 depicts the VIP values obtained from PLS-DA conducted on ABC transformed data. For the sake of simplicity, VIP values from raw data are reporte in table S2 from Supplementary materials. As it can be seen, the analysis of ABC-transformed data revealed that 27 metabolites had VIP>1, with 11 also identified as significant in the univariate analysis. Conversely, when raw data were used, 28 metabolites showed VIP values greater than 1, of which 9 overlapped with the univariate analysis results.

Table 3. Variable importance in projection values of metabolites from the 2-LVs PLS-DA for discrimination of samples from different geographical origin. The analysis was performed on ABC transformed data.

Compound	VIP	Compound	VIP	Compound	VIP
Lactic Acid, 2TMS*	1.03	L-5-Oxoproline, 2TMS	0.95	D-Gluconic acid, 6TMS	1.31
Glycolic acid, 2TMS derivative Results*	1.14	2,3,4-Trihydroxybutyric acid 4TMS	0.97	Palmitic Acid, TMS	0.59
L-Alanine, 2TMS	0.96	Pentanedioic acid, 2-3TMS	0.66	Myo-Inositol, 6TMS	1.05
Glycine, di-TMS	1.37	L-Glutamic acid, 3TMS*	1.04	N-Acetyl-D-galactosaminitol, 5TMS*	1.39
3-Hydroxybutyric acid, 2TMS derivative	0.80	L-Phenylalanine, 2TMS*	1.01	9,12-Octadecadienoic acid (Z,Z)-, TMS	1.26
2-Aminobutanoic acid, 2TMS	0.73	Arabinonic acid, 2,3,5-gamma.-lactone, 3TMS	1.14	11-Octadecenoic acid, (Z)-, TMS	0.72
L-Valine, 2TMS	0.99	Deoxyribopyranose, 3TMS*	1.14	Stearic acid, TMS	0.67
Urea, 2TMS	1.38	Xylitol, 5TMS	0.99	D-Myo-Inositol phosphate 7TMS	0.85
L-Leucine, 2TMS	0.87	Glycerol phosphate 4TMS	0.59	Uridine, 3TMS*	1.14
Glycerol, 3TMS	1.02	2-Keto-l-gluconic acid, 5TMS	1.56	8-Benzylquinoline	0.28
L-Isoleucine, 2TMS	1.02	(Z)-3-Hexenyl .beta.-glucopyranoside, 4TMS d	0.82	1-Monopalmitin, 2TMS	0.88
Glyceric acid, 3TMS	1.01	'D-(-)-Fructofuranose, pentakis(trimethylsilyl) ether (isomer 1) Results	1.48	Rosiridin, 5TMS derivative Results	1.28
Uracil, 2TMS*	1.00	Myristic acid, TMS	0.48	alpha.-D-glucopyranose, 1-O-(3-O-(2-methylbutanoyl)-.alpha.-D-glucopyranosyl), 7TMS	1.23
Serine, 3TMS	0.94	Talose, 5TMS	0.54	2-Monostearin, 2TMS	0.70
2,3-Dihydroxy-2-methylbutanoic acid, 3TMS	1.11	D-Glucitol, 6TMS	0.92	1-Monolinolein, 2TMS	0.97

L-Threonine, 3TMS*	1.01	D-Allofuranose, 5TMS	0.98	Glycerol monostearate, 2TMS	1.15
2,3-Dihydroxy-3-methylbutyric acid, 3TMS*	0.74	D-Glucopyranose, 5TMS	0.70	D-Trehalose 8TMS	1.23
Butanoic acid, 3,4-dihydroxy, 2TMS*	1.22	Scyllo-Inositol, 6TMS Results*	1.01	Brassicasterol, 1TMS	0.68
Malic acid, 3TMS	0.73	D-Allofuranose, 5TMS	0.97	Ergosterol, TMS	0.59

* The metabolite was identified as significant Bonferroni-Holm correction was applied on p-values obtained from ANOVA test on ABC transformed data

To conclude the multivariate analysis, it is worth mentioning that while the ABC transformation significantly enhances data normalization and statistical power, it can potentially alter the original data scale, thus influencing the interpretability of metabolite concentrations. On the other hand, it results to be effective in enhancing model performance and reduce skewness, both crucial in medium to high-dimensional metabolomics data. In the context of this study, UV scaling applied for multivariate analysis aided in supporting data interpretability through the standardization of variable magnitudes, which did not undermine the biological meaning of the investigated variables. Indeed, the loading values interpretation obtained by multivariate models complies with findings from the univariate identification process.

4. Conclusions

The issue of significant feature selection is crucial for identifying biomarkers. The accurate identification of these metabolites may be compromised if they are not appropriately transformed. This study aimed to expand the recent advances in skewed data management by conceptualizing them within the significant feature selection framework in a food-omic contest for the first time. To this end, we first employed non-parametric significance tests, such as Kruskal-Wallis and Permutation tests, which are often used to analyse data that do not follow the normal distribution. However, parametric methods showed a higher statistical power than non-parametric ones when the data follow a Gaussian

distribution. Moreover, unlike the Kruskal-Wallis test, which relies on ranks, ANOVA makes use of more informative data characteristics, such as the mean and variance, which leads to more interpretable conclusions. Therefore, we conducted an ANOVA test, following a data transformation into a Gaussian-like distribution using the ABC transformation. The implementation of this procedure resulted in the identification of the same metabolites to those obtained through Kruskall-Wallis, except for only two of them. The statistical power of data was remarkably improved, thus demonstrating both the benefits achievable from the ABC transformation in enhancing the separability of samples from different groups and the robustness of the target metabolite selection criteria. Univariate and Multivariate statistics analysis were employed to corroborate the methodology from different perspectives. The proposed approach can improve the reliability and accuracy of biomarker identification, providing a more comprehensive understanding of metabolite responses to various stimuli. Although the ABC-transformed ANOVA and K-W test exhibited similar performance, the transformation significantly enhances PCA analysis. This methodology can be applied to more intricate studies involving multiple categories or continuous predictors, where non-parametric tests are not commonly utilized, or it can be extended to current challenges involving missing data imputation algorithms, where data are assumed to follow a Gaussian distribution. In conclusion, the proposed approach can facilitate the creation of new data management protocols tailored to specific cases.

Appendix A

Table A.1 shows the results obtained by executing the algorithm reported in Fig.1 in the Matlab environment for the optimal λ search.

Table A.1 Optimal λ identification. Results are reported for each metabolite.

Compound	λ	Compound	λ	Compound	λ
Lactic Acid, 2TMS	-0.26	L-5-Oxoproline, 2TMS	0.41	D-Gluconic acid, 6TMS	0.36
Glycolic acid, 2TMS derivative Results'	-0.01	2,3,4-Trihydroxybutyric acid 4TMS	0.54	Palmitic Acid, TMS	-0.45
L-Alanine, 2TMS	0.82	Pentanedioic acid, 2-3TMS	-0.06	Myo-Inositol, 6TMS	0.39
Glycine, di-TMS	0.35	L-Glutamic acid, 3TMS	0.29	N-Acetyl-D-galactosaminitol, 5TMS	-0.16
'3-Hydroxybutyric acid, 2TMS derivative	0.14	L-Phenylalanine, 2TMS	0.16	9,12-Octadecadienoic acid (Z,Z)-, TMS	3
2-Aminobutanoic acid, 2TMS	0.23	Arabinonic acid, 2,3,5-gamma.-lactone, 3TMS	0.16	11-Octadecenoic acid, (Z)-, TMS	1.3
L-Valine, 2TMS	0.61	Deoxyribopyranose, 3TMS	-0.18	Stearic acid, TMS	0.65
Urea, 2TMS	-0.09	Xylitol, 5TMS	0.46	D-Myo-Inositol phosphate 7TMS	0.06
L-Leucine, 2TMS	0.04	Glycerol phosphate 4TMS	0.66	Uridine, 3TMS	0.37
Glycerol, 3TMS	1.38	2-Keto-l-gluconic acid, 5TMS	0.44	8-Benzylquinoline	-1.66
L-Isoleucine, 2TMS	0.82	(Z)-3-Hexenyl .beta.-glucopyranoside, 4TMS d	-0.17	1-Monopalmitin, 2TMS	2.83
Glyceric acid, 3TMS	0.4	'D-(-)-Fructofuranose, pentakis(trimethylsilyl) ether (isomer 1) Results'	-0.09	'Rosiridin, 5TMS derivative Results'	0.43
'Uracil, 2TMS	-0.08	Myristic acid, TMS	-0.25	alpha.-D-glucopyranose, 1-O-(3-O-(2-methylbutanoyl)-.alpha.-D-glucopyranosyl), 7TMS	0.12
Serine, 3TMS	0.45	Talose, 5TMS	-0.23	2-Monostearin, 2TMS	-0.57
2,3-Dihydroxy-2-methylbutanoic acid, 3TMS	0.59	D-Glucitol, 6TMS	1.88	1-Monolinolein, 2TMS	-0.64
L-Threonine, 3TMS	0.5	D-Allofuranose, 5TMS	0.81	Glycerol monostearate, 2TMS	-2.05
2,3-Dihydroxy-3-methylbutyric acid, 3TMS	0.09	D-Glucopyranose, 5TMS	-0.17	D-Trehalose 8TMS	0.17
Butanoic acid, 3,4-dihydroxy, 2TMS	0.06	Scyllo-Inositol, 6TMS Results'	0.44	Brassicasterol, 1TMS	-0.18
Malic acid, 3TMS	-0.04	D-Allofuranose, 5TMS	0.21	Ergosterol, TMS	0.48

References

[1] A. Smolinska, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review, Anal. Chim. Acta 750 (2012) 82–97. https://doi.org/10.1016/j.aca.2012.05.049.

[2] M. San Nicolas, A. Villate, M. Olivares, N. Etxebarria, O. Zuloaga, O. Aizpurua-Olaizola, A. Usobiaga, Exploratory optimisation of a LC-HRMS based analytical method for untargeted metabolomic screening of Cannabis Sativa L. through Data Mining, Anal. Chim. Acta 1279 (2023). https://doi.org/10.1016/j.aca.2023.341848.

[3] J.M. Cevallos-Cevallos, J.I. Reyes-De-Corcuera, Metabolomics in Food Science, Adv. Food Nutr. Res. 67 (2012) 1–24. https://doi.org/10.1016/B978-0-12-394598-3.00001-0.

[4] P. Caboni, P. Scano, S. Sanchez, S. Garcia-Barreda, F. Corrias, P. Marco, Multi-platform metabolomic approach to discriminate ripening markers of black truffles (Tuber melanosporum), Food Chem. 319 (2020). https://doi.org/10.1016/j.foodchem.2020.126573.

[5] U. Roessner, J. Bowne, What is metabolomics all about?, Biotechniques 46 (2009) 363–365. https://doi.org/10.2144/000113133.

[6] U.W. Liebal, A.N.T. Phan, M. Sudhakar, K. Raman, L.M. Blank, Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (2020) 1–23. https://doi.org/10.3390/metabo10060243.

[7] G. Zhu, Y. Wang, W. Wang, F. Shang, B. Pei, Y. Zhao, D. Kong, Z. Fan, Untargeted GC-MS-Based Metabolomics for Early Detection of Colorectal Cancer, Front. Oncol. 11 (2021). https://doi.org/10.3389/fonc.2021.729512.

[8] Q. Yang, B.H. Shi, G.L. Tian, Q.Q. Niu, J. Tang, D.D. Linghu, H.Q. He, B.Q. Wu, J.T. Yang, L. Xu, R.Q. Yu, GC–MS urinary metabolomics analysis of inherited metabolic diseases and stable metabolic biomarker screening by a comprehensive chemometric method, Microchem. J. 168 (2021). https://doi.org/10.1016/j.microc.2021.106350.

[9] J. Sun, Y. Xia, Pretreating and normalizing metabolomics data for statistical analysis, Genes Dis. (2023). https://doi.org/10.1016/j.gendis.2023.04.018.

[10] I. Karaman, Preprocessing and pretreatment of metabolomics data for statistical analysis, Adv. Exp. Med. Biol. 965 (2017) 145–161. https://doi.org/10.1007/978-3-319-47656-8_6.

[11] J.I. Vélez, J.C. Correa, F. Marmolejo-Ramos, A new approach to the Box–Cox transformation, Front. Appl. Math. Stat. 1 (2015). https://doi.org/10.3389/fams.2015.00012.

[12] H. Yu, P. Sang, T. Huan, Adaptive Box-Cox Transformation: A Highly Flexible Feature-Specific Data Transformation to Improve Metabolomic Data Normality for Better Statistical Analysis, Anal. Chem. 94 (2022) 8267–8276. https://doi.org/10.1021/acs.analchem.2c00503.

[13] I. Martínez-Arranz, R. Mayo, M. Pérez-Cormenzana, I. Mincholé, L. Salazar, C. Alonso, J.M. Mato, Enhancing metabolomics research through data mining, J. Proteomics 127 (2015) 275–288. https://doi.org/10.1016/j.jprot.2015.01.019.

[14] A.B. Frahm, P.R. Jensen, J.H. Ardenkjær-Larsen, D. Yigit, M.H. Lerche, Stable isotope resolved metabolomics classification of prostate cancer cells using hyperpolarized NMR data, J. Magn. Reson. 316 (2020). https://doi.org/10.1016/j.jmr.2020.106750.

[15] M.A. Mohamed, A.S. Fayed, M.A. Hegazy, N.N. Salama, E.E. Abbas, Fully optimized new sensitive electrochemical sensing platform for the selective determination of antiepileptic drug ezogabine, Microchem. J. 144 (2019) 130–138. https://doi.org/10.1016/j.microc.2018.08.062.

[16] S. Gouveia-Figueira, M. Karimpour, J.A. Bosson, A. Blomberg, J. Unosson, M. Sehlstedt, J. Pourazar, T. Sandström, A.F. Behndig, M.L. Nording, Mass spectrometry profiling reveals altered plasma levels of monohydroxy fatty acids and related lipids in healthy humans after controlled exposure to biodiesel exhaust, Anal. Chim. Acta 1018 (2018) 62–69. https://doi.org/10.1016/j.aca.2018.02.032.

[17] S. Zeppa, C. Guidi, A. Zambonelli, L. Potenza, L. Vallorani, R. Pierleoni, C. Sacconi, V. Stocchi, Identification of putative genes involved in the development of Tuber borchii fruit body by mRNA differential display in agarose gel, Curr. Genet. 42 (2002) 161–168. https://doi.org/10.1007/s00294-002-0343-6.

[18] C. Susana Rivera, M.E. Venturini, R. Oria, D. Blanco, Selection of a decontamination treatment for fresh Tuber aestivum and Tuber melanosporum truffles packaged in modified atmospheres, Food Control 22 (2011) 626–632. https://doi.org/10.1016/j.foodcont.2010.10.015.

[19] J. FOLCH, M. LEES, G.H. SLOANE STANLEY, A simple method for the isolation and purification of total lipides from animal tissues., J. Biol. Chem. 226 (1957) 497–509. https://doi.org/10.1016/s0021-9258(18)64849-5.

[20] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Anal. Methods 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

[21] B. Shaun, Missing values, outliers, robust statistics & non-parametric methods, Sci. Data Manag. 1 (1998) 32–38.

[22] C. Potvin, D.A. Roff, Distribution-free and robust statistical methods: viable alternatives to parametric statistics?, Ecology 74 (1993) 1617–1628. https://doi.org/10.2307/1939920.

[23] K.J. Berry, J.E. Johnston, P.W. Mielke, Permutation methods, Wiley Interdiscip. Rev. Comput.

Stat. 3 (2011) 527–542. https://doi.org/10.1002/wics.177.

[24] R.M. Sakia, The Box-Cox Transformation Technique: A Review, Stat. 41 (1992) 169. https://doi.org/10.2307/2348250.

[25] Y. Bee Wah, N. Mohd Razali, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, J. Stat. Model. Anal. 2 (2011) 21–33.

[26] R.R. de Souza, M. Toebe, A.C. Mello, K.C. Bittencourt, Sample size and Shapiro-Wilk test: An analysis for soybean grain yield, Eur. J. Agron. 142 (2023). https://doi.org/10.1016/j.eja.2022.126666.

[27] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.

[28] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemom. Intell. Lab. Syst. 56 (2001) 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2.

[29] E. Tejedor-Calvo, S. Garcia-Barreda, J. Sebastián Dambolena, D. Pelissero, S. Sánchez, P. Marco, E. Nouhra, Aromatic profile of black truffle grown in Argentina: Characterization of commercial categories and alterations associated to maturation, harvesting date and orchard management practices, Food Res. Int. 173 (2023). https://doi.org/10.1016/j.foodres.2023.113300.

[30] N. Šiškovič, L. Strojnik, T. Grebenc, R. Vidrih, N. Ogrinc, Differentiation between species and regional origin of fresh and freeze-dried truffles according to their volatile profiles, Food Control 123 (2021). https://doi.org/10.1016/j.foodcont.2020.107698.

[31] F. Wernig, F. Buegger, K. Pritsch, R. Splivallo, Composition and authentication of commercial and home-made white truffle-flavored oils, Food Control 87 (2018) 9–16. https://doi.org/10.1016/j.foodcont.2017.11.045.

[32] Y.J. Tang, G. Wang, Y.Y. Li, J.J. Zhong, Fermentation condition outweighed truffle species in affecting volatile organic compounds analyzed by chromatographic fingerprint system, Anal. Chim. Acta 647 (2009) 40–45. https://doi.org/10.1016/j.aca.2009.05.027.

[33] D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, E. Pujos-Guillot, Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data, Front. Mol. Biosci. 3 (2016). https://doi.org/10.3389/fmolb.2016.00030.

[34] Y. Fan, T.B. Murphy, J.C. Byrne, L. Brennan, J.M. Fitzpatrick, R.W.G. Watson, Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer, J. Proteome Res. 10 (2011) 1361–1373. https://doi.org/10.1021/pr1011069.

[35] M. Pérez-Cova, S. Platikanov, D.R. Stoll, R. Tauler, J. Jaumot, Comparison of Multivariate ANOVA-Based Approaches for the Determination of Relevant Variables in Experimentally Designed Metabolomic Studies, Molecules 27 (2022). https://doi.org/10.3390/molecules27103304.

[36] R.G. Brereton, Chemometrics for Pattern Recognition, 2009. https://doi.org/10.1002/9780470746462.

[37] A. Cheddad, On Box-Cox Transformation for Image Normality and Pattern Classification, IEEE Access 8 (2020) 154975–154983. https://doi.org/10.1109/ACCESS.2020.3018874.

[38] M. Bicego, S. Baldo, Properties of the Box–Cox transformation for pattern classification, Neurocomputing 218 (2016) 390–400. https://doi.org/10.1016/j.neucom.2016.08.081.

[39] C.M. Rubingh, S. Bijlsma, E.P.P.A. Derks, I. Bobeldijk, E.R. Verheij, S. Kochhar, A.K. Smilde, Assessing the performance of statistical validation tools for megavariate metabolomics data,

Metabolomics 2 (2006) 53–61. https://doi.org/10.1007/s11306-006-0022-6.

[40] L. Sibono, M. Grosso, S. Tronci, M. Errico, M. Addis, M. Vacca, C. Manis, P. Caboni, Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction, Metabolites 13 (2023). https://doi.org/10.3390/metabo13070877.

[41] D.J. Lubinsky, Increasing the performance and consistency of classification trees by using the accuracy criterion at the leaves, Proc. 12th Int. Conf. Mach. Learn. ICML 1995 (1995) 371–377. https://doi.org/10.1016/b978-1-55860-377-6.50053-0.

[42] N. Dhamayanthi, B. Lavanya, Improvement in Software Defect Prediction Outcome Using Principal Component Analysis and Ensemble Machine Learning Algorithms, Lect. Notes Data Eng. Commun. Technol. 26 (2019) 397–406. https://doi.org/10.1007/978-3-030-03146-6_44.

- Statistical power from normalised data is greater than that from skewed raw data.
- Adaptive Box-Cox (ABC) transformation enhances data interpretability.
- Data normalisation improves the performances of multivariate analysis tools.
- Spain and Argentina truffle metabolites are analysed and compared.
- Both ABC transformed ANOVA and non-parametric test correctly detect the biomarkers.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.