Food Chemistry 339 (2021) 127852

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

^{1}H NMR chemometric models for classification of Czech wine type and variety

Anna Mascellani[a], Gokce Hoca[a], Marek Babisz[b], Pavel Krska[b], Pavel Kloucek[a], Jaroslav Havlik[a,*]

[a] Department of Food Science, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamycka 129, 165 00 Prague 6 – Suchdol, Czech Republic

[b] The National Wine Centre, Zamek 1, 691 42 Valtice, Czech Republic

ARTICLE INFO

Keywords:
^{1}H NMR
Chemometrics
Wine classification
randomForest
Wine analysis

Chemical compounds studied in this article:
Isobutanol (PubChem CID: 6560)
Isopentanol (Compound CID: 31260)
Isoleucine (Compound CID: 6306)
Leucine (Compound CID: 6106)
2,3-Butanediol (Compound CID: 262)
Ethanol (Compound CID: 702)
Ethyl Acetate (Compound CID: 8857)
Lactate (Compound CID: 612)
Alanine (Compound CID: 5950)
Proline (Compound CID: 145742)
Acetate (Compound CID: 175)
Methionine (Compound CID: 876)
Acetoin (Compound CID: 179)
Acetoacetate (Compound CID: 6971017)
Pyruvate (Compound CID: 107735)
Glutamate (Compound CID: 33032)
Succinate (Compound CID: 160419)
Choline (Compound CID: 305)
Myo-Inositol (Compound CID: 892)
Methanol (Compound CID: 887)
Glycerol (Compound CID: 753)
Fructose (Compound CID: 2723872)
Tartrate (Compound CID: 3806114)
Glucose (Compound CID: 5793)
Turanose (Compound CID: 5460935)
Uridine (Compound CID: 6029)
Catechin (Compound CID: 9064)
Epicatechin (Compound CID: 72276)
p-Hydroxyphenylacetic acid (Compound CID: 127)
Tyrosine (Compound CID: 6057)
Gallate (Compound CID: 370)
Phenethyl alcohol (Compound CID: 57361413)
Phenylalanine (Compound CID: 6140)
Chlorogenate (Compound CID: 1794427)

ABSTRACT

A set of 917 wines of Czech origin were analysed using nuclear magnetic resonance spectroscopy (NMR) with the aim of building and evaluating multivariate statistical models and machine learning methods for the classification of 6 types based on colour and residual sugar content, 13 wine grape varieties and 4 locations based on ^{1}H NMR spectra. The predictive models afforded greater than 93% correctness for classifying dry and medium dry, medium, and sweet white wines and dry red wines. The trained Random Forest (RF) model classified Pinot noir with 96% correctness, Blaufränkisch 96%, Riesling 92%, Cabernet Sauvignon 77%, Chardonnay 76%, Gewürztraminer 60%, Hibernal 60%, Grüner Veltliner 52%, Pinot gris 48%, Sauvignon Blanc 45%, and Pálava 40%. Pinot blanc and Chardonnay, varieties that are often mistakenly interchanged, were discriminated with 71% correctness. The findings support chemometrics as a tool for predicting important features in wine, particularly for quality assessment and fraud detection.

* Corresponding author.
E-mail address: havlik@af.czu.cz (J. Havlik).

https://doi.org/10.1016/j.foodchem.2020.127852
Received 18 May 2020; Received in revised form 13 August 2020; Accepted 14 August 2020
Available online 17 August 2020

Formate (Compound CID: 283)
Histidin (Compound CID: 6274)
Trigonellin (Compound CID: 5570)

1. Introduction

The Czech Republic is becoming an important European wine producer (Borák & Vacek, 2018). Marketed Czech wines primarily comprise varietal wines in which at least 85% of the grapes used for production represent a specific wine grape variety that must be defined on the label. European regulations specify labelling rules to inform customers and guarantee producers recognition of the quality of their products. In addition to the compulsory labelling particulars (EU No 2019/33), Protected Designation of Origin (PDO) wines of the Czech Republic must state additional particulars on the label, including the vintage year, the names of one or more wine grape varieties, the residual sugar content, and the community symbol, which indicates the PDO and geographical unit. The label data are verified by the Czech Food Inspection Authority in the process of classification that precedes the introduction to the market of PDO wines (International Organisation of Vine and Wine, 2013, 2015; Národní vinařské centrum, 2020; EU No 2019/33).

Wine is a complex system, and its chemical profile is the result of the environment, climate, wine grape variety, metabolism of yeast during fermentation, and various human interventions (such as temperature, barrel and/or bottle ageing, and addition of acids, enzymes and fining agents). High standards require analytical tools that are capable of assessing residual sugar content, wine grape variety, wine colour, origin, and the addition of sugar, sweeteners or flavourings. Nuclear magnetic resonance (NMR) and mass spectroscopy are the core methods of metabolomics (Hong, 2011; Johanningsmeier, Harris, & Klevorn, 2016), which is the study of global metabolite profiles in a system under a set of conditions (Rochfort, Ezernieks, Bastian, & Downey, 2010). When applied to the onventional methods allowing the high-clasfood production/consumption chain, metabolomics is called foodomics (Trimigno, Marincola, Dellarosa, Picone, & Laghi, 2015). NMR, a versatile high-resolution technique that requires minimal sample pre-treatment, is particularly suitable for wine analysis due to its broad metabolite coverage, high throughput and reproducibility. Recent applications of ^{1}H NMR to wine metabolite analysis have included investigations of differences between wine grape varieties, geographic origins and vintages (Ali, Maltese, Toepfer, Choi, & Verpoorte, 2011; Brescia, Košir, Caldarola, Kidrič, & Sacco, 2003; Du, Bai, Zhang, & Liu, 2007; Godelmann et al., 2013; Gougeon, da Costa, Guyon, & Richard, 2019; Larsen, Berg, & Engelsen, 2006; Mazzei, Francesca, Moschetti, & Piccolo, 2010; Papotti et al., 2013; Rochfort, Ezernieks, Bastian, & Downey, 2010; Son et al., 2008; Viggiani & Morelli, 2008), oenological practices (Amargianitaki & Spyros, 2017) or winemaking techniques (Mazzei, Spaccini, Francesca, Moschetti, & Piccolo, 2013). In addition, ^{1}H NMR is used for quality screening by control agencies (Minoja & Napoli, 2014).

To unravel the rich and complex compositional information present in NMR datasets, multivariate statistical analysis is used. The most widely used chemometrics methods for data mining with discrimination power are principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA). Alternative data treatment methods, such as random forest (RF) (Breiman, 2001), show great potential and advantages compared to these conventional methods allowing high-classification performances, including minimising the risk of over-fitting and class imbalance problems and eliminating irrelevant features. A shortcoming of these alternative methods is that the interpretation of the results is complex since the classification is not displayed as a graphical tree (Jiménez-Carvelo, González-Casado, Bagur-González, & Cuadros-Rodríguez, 2019; Menze et al., 2009); consequently, their application in food quality and authenticity remains scarce nevertheless, it has shown to be a state-of-the art method (Scott et al., 2013). RF has been previously applied in wine metabolomics based on gas chromatography-mass spectrometry with the aim of evaluating the discriminatory power of different compound classes for classification of four wine grape varieties (Gómez-Meire, Campos, Falqué, Díaz, & Fdez-Riverola, 2014). The RF algorithm is an unbiased supervised classification method based on a collection of decision trees formed using bootstrap samples from the learning dataset. The final classification is determined by computing the frequencies of predictions for each group over the whole forest (Breiman, 2001). The variables contributing the most to the separation could be evaluated by two different measures: “mean decrease accuracy” derived from statistical permutation tests, or “Gini importance” from the training of the random forest classifier (Breiman, 2001; Menze et al., 2009).

In the present study, Czech wines with a PDO designation were investigated by an untargeted NMR spectroscopy approach coupled with multivariate statistical analysis. We evaluated the ability of NMR coupled with chemometrics to discriminate Czech wines according to (i) wine colour and residual sugar content (type), (ii) wine grape variety and (iii) geographic origin. In summary, we propose an innovative approach based on NMR spectrometry, combined with the RF algorithm, to provide an accurate prediction model for untargeted simultaneous discrimination of 13 wine grape varieties. Table 1

2. Material and methods

2.1. Wine samples

A set of 917 bottles of Czech wines were provided by The National Wine Centre's Wine Salon of the Czech Republic, an independent national wine competition, in 2019. The wines were from five sub-regions of Bohemia and Moravia (Supplement Material Figure S1): Mikulovská (250 samples), Slovácká (153 samples), Velkopavlovická (143 samples), Znojemská (111 samples) and Mělnická (8 samples); for two wines, the area was unclassified. The wines belonged to six types based on wine colour and residual sugar content (Supplementary Material Table S1): 494 wines were categorised as dry and medium dry white wines, 137 as medium white wines, 36 as sweet white wines, 44 as rosés and blancs de noir, 188 as dry red wines and 4 as other red wines. The wines were made from wine grape varieties including Riesling (88 samples), Chardonnay (70 samples), Pinot gris (70 samples), Sauvignon Blanc (63 samples), Welschriesling (62 samples), Pinot noir (58 samples), Grüner Veltliner (51 samples), Gewürztraminer (50 samples), Pinot blanc (47 samples), Blaufränkisch (37 samples), Pálava (36 samples), Cabernet Sauvignon (24 samples), Hibernal (24 samples), Zweigeltrebe (20 samples), Grüner Silvaner (19 samples), Saint Laurent (19 samples), Neuburger (14 samples), Merlot (13 samples), Muskat Moravsky (13 samples), Andre (12 samples), Müller-Thurgau (12 samples), Cabernet Moravia (11 samples), Blauer Portugieser (9 samples), Dornfelder (7 samples), Kerner (6 samples), Aurelius (5 samples) and Alibernet (3 samples); 16 samples were classified as other, and 45 samples were cuvée. The year of harvest ranged between 2007 and 2017. The sugars and the actual alcoholic strength of the wines were previously assessed by analytical methods approved by The International Organisation of Vine and Wine (OIV) and declared on the certificate of analysis, which is required for introduction to the market (International Organisation of Vine and Wine, 2019).

Table 1
^{1}H NMR chemical shifts and coupling constants (Hz) of wine compounds identified by references and using 1D and 2D NMR spectra (j-res, DQF-COSY, long-range COSY, HMBC).

Peak	Compound	δ_H[b]
1	Isobutanol	0.88 (d) 30.5 (m) 3.35 (d)
2	Isopentanol	0.88 (d), 1.43 (q), 1.65 (m), 3.63 (t)
3	Isoleucine	0.93 (t), 1.00 (d), 1.25 (m), 1.46 (m), 1.97 (m), 3.67 (d)
4	Leucine	0.95 (d), 1.70 (m), 3.74 (m)
5	2,3-Butanediol	1.13 (d), 3.71 (m)
6	Ethanol	1.17 (t), 3.65 (q)
7	Ethyl acetate	1.24 (t), 2.07 (s), 4.13 (q)
8	Lactate	1.40 (d), 4.16 (q)
9	Alanine	1.49 (d), 3.79 (q)
10	Proline	2.02 (m), 2.34 (m), 3.34 (m), 3.41 (m), 4.12 (dd)
11	Acetate	2.05 (s)
12	Methionine	2.10 (m), 2.64 (t), 3.86 (dd)
13	Acetoin	1.36 (d), 2.21 (s), 4.42 (q)
14	Acetoacetate	2.26 (s), 3.48 (s)
15	Pyruvate	2.36 (s)
16	Glutamate	2.14 (m), 2.49 (m), 3.79 (dd)
17	Succinate	2.65 (s)
18	Choline	3.19 (s), 3.51 (m), 4.06 (m)
19	Myo-Inositol*	3.27 (t), 3.53 (dd), 3.61 (t), 4.04 (t)
20	Methanol	3.35 (s)
21	Glycerol	3.55 (m), 3.65 (m), 3.76 (tt)
22	Fructose	3.55 (m), 3.69 (m), 3.80 (m), 3.88 (dd), 3.98 (m), 4.01 (d), 4.09 (d)
23	Tartrate	4.38 (s)
24	*α*-Glucose	5.21 (d)
25	*β*-Glucose	4.62 (d)
26	Turanose	5.30 (d), 5.20 (d)
27	Uridine	5.88 (d), 5.90 (d), 7.87 (d)
28	Catechin	2.55 (dd), 2.90 (m), 4.24 (m), 6.00 (d), 6.09 (d), 6.84 (dd), 6.93 (m)
29	Epicatechin	2.77 (dd), 2.93 (m), 4.33 (t), 6.09 (dd), 6.92 (m), 7.03 (s)
30	*p*-Hydroxyphenylacetic acid	3.44 (s), 6.84 (dt), 7.17 (dt)
31	Tyrosine	6.88 (d), 7.18 (d)
32	Gallate	7.15 (s)
33	Phenethyl alcohol	2.86 (t), 2.91 (m), 3.84 (t), 7.30 (m), 7.37 (m)
34	Phenylalanine	3.12 (dd), 3.28 (dd), 4.00 (dd), 7.32 (d), 7.40 (m)
35	Chlorogenate	2.04(m), 2.18 (m), 3.88 (dd), 4.2 (m), 5.32 (m), 6.43 (d), 7.03 (d), 7.13 (d), 7.21 (d), 7.68 (d)
36	Formate	8.30 (s)
37	Histidine	8.65(s), 7.39 (s), 4.04 (dd), 3.34 (m)
38	Trigonelline	4.43 (s), 8.08 (t), 8.83 (d), 8.84 (d), 9.13 (s)

[b] Peak multiplicities in parentheses: s, singlet; d, doublet; t, triplet; dd, doublet of doublets; dt, doublet of triplets; q, quartet; and m, multiplet. The chemical shifts were determined at pH 3.1 and expressed relative to DSS at 0 ppm. *, not of plant origin, probably phytate.

2.2. Sample preparation

All chemical reagents used for sample preparation for analysis were of analytical grade. NaN_3 (≥99.5%), H_3PO_4 (99%), 2,2-Dimethyl-2-silapentane-5-sulfonate sodium salt (97%, DSS), KH_2PO_4 (99%), D_2O (99.9%, D_2O) and HCl (36.5–38.0) were purchased from VWR (Radnor, PA, USA).

Phosphate buffer solution was prepared in D_2O by adding 1.5 M KH_2PO_4, 0.2% NaN_3 and 5 mM DSS. The pH was adjusted to 4 with H_3PO_4 and measured by an inoLab® pH 7110 pH meter (WTW, Germany). DSS was used as the internal standard to calibrate the chemical shift to 0 ppm. To each 540 µl of sample, 60 µl of phosphate buffer solution was added and mixed using an IKA® MS 3 basic vortex oscillator. The total 600-µl volume of prepared sample was transferred to a 5-mm NMR tube (NORELL Inc., Morganton, NC, USA) and subsequently analysed.

2.3. NMR data acquisition

All spectra were recorded on a Bruker Avance III spectrometer equipped with a broad band fluorine observation (BBFO) SmartProbe™ with z-axis gradients (Bruker BioSpin GmbH, Rheinstetten, Germany) operating at a proton NMR frequency of 500.23 MHz. The temperature was set to 298 K (25 °C). ^{1}H NMR spectra were acquired and processed under the same conditions. The Bruker pulse sequence (noesypr1d) was applied to suppress the residual water signal at 4.704 ppm. For each sample, 128 scans and 4 dummy scans were collected as 64 K data points using a spectral width of 8 K Hz, receiver gain of approx. 18, relaxation delay of 1 s, acquisition time of 4.00 s, and mixing time of 0.1 s. The total acquisition time was 11 min. Tuning, lock, gain, 90° pulse calibration, and shimming were calibrated automatically for each sample by the standard module routine developed by Bruker (atma, lock, rga, pulsecal and topshim). The free induction decay (FID) was multiplied by 0.3 Hz line broadening prior to Fourier transformation. The raw NMR files are accessible from MetaboLights (study identifier: MTBLS1677, www.ebi.ac.uk/metabolights/MTBLS1677) (Haug et al., 2020).

2.4. Metabolite profiling

^{1}H NMR spectra of pooled wines were phased and baseline corrected using Chenomx NMR suite 8.5 software, professional edition (Chenomx Inc., Edmonton, AB, Canada). Signal assignment was performed using internal and in-house databases. The assignments of the metabolites were confirmed by ^{13}C NMR, J-resolved, ^{1}H–^{1}H COSY, long-range COSY and ^{1}H − ^{13}C HSQC (for settings, see Supplementary Material Table S2).

2.5. NMR data reduction and pre-processing

^{1}H NMR spectra were subjected to phase and baseline correction using the Whitakker smoother algorithm in MestReNova software version 14.1.0 (Mestrelab Reserach S.L., Santiago de Compostela, Spain). The spectra were calibrated to the internal standard DSS at 0.0 ppm and aligned by global alignment based on a reference spectrum. Specific scripts were implemented for data reduction. The ^{1}H NMR spectral data were reduced into 0.04-ppm spectral bins using the sum of the datapoints, and the region corresponding to water (4.74–4.98) was excluded. The binning width of 0.04 ppm represents a compromise between preserving sufficient data resolution and minimizing the loss of spectral information and the effect of peak drift. To avoid false grouping of samples in chemometrics models caused by the pH and paramagnetic properties of the wines, the bins from areas with large drift (1.02–1.34, 1.34–1.42, 2.02–2.10, 2.62–2.74, 3.58–3.66, 4.38–4.54, 8.22–8.42) were summed. A total of 226 bins were used for analysis. The spectra were then normalized to the average intensity of the internal standard and the total spectral area. The data were log-transformed, mean-centred and divided by the standard deviation of each variable (auto-scaled) obtained using the package 'MetaboAnalyst' version 2.0 (Chong, Yamamoto, & Xia, 2019) in R 3.5.0. The bins were used as variables for subsequent statistical analysis and modelling.

2.6. Exploratory data analysis and statistical modelling

The PCA was designed to reduce the dimensionality of the original data, identify sample composition trends, and exclude strong outliers. The heatmap was developed by applying hierarchical clustering (HC) analysis of the group averages and the features ranked using ANOVA. In this technique, the similarity among group averages was measured using the Pearson distance, while cluster aggregation was based on the average linkage method. Normalized bins were used for exploratory data analysis using the package 'MetaboAnalyst' version 2.0 (Chong et al., 2019) in R 3.5.0. The random forest (RF) algorithm was used for

classification by type and wine grape variety. This method generates a combination of decision trees using a bootstrap sample. The number of trees were set to 500; 70% of the samples were used as the training subset, and the remaining 30% of the samples were used as the testing subset. Calculations were performed using the 'randomForest' package 4.6–14 (Liaw & Wiener, 2002) available in R 3.5.0. A classification tree was fit to each bootstrapped sample, and each node within a tree was constructed by selecting a random subset of the environmental variables (for this parameter, mtry was set to 75). The RF algorithm was trained in two sequential steps. In the first stage, the RF model was trained to identify the type of wine. In the second step, the RF model was trained to discriminate different wine grape varieties. An additional RF model was built for the discrimination of Chardonnay from Pinot noir wines. Rosés and blancs de noir were excluded from the HC and RF analysis, as these are pinkish or white wines made from red grapes whose skins were removed before or after the beginning of fermentation. Another classification model was established by partial least squares-discriminant analysis (PLS-DA) obtained using the package 'MetaboAnalyst' version 2.0 (Chong et al., 2019) in R 3.5.0 with the aim of classifying the wines according to sub-region. The wine grape varieties with the largest numbers of samples (Chardonnay, Pinot noir and Riesling) were classified according to origin (Mikulovská, Slovácká, Velkopavlovická and Znojemská). The PLS-DA model was validated by full (leave-one-out) cross-validation in which each sample was predicted by the remaining samples and the procedure were repeated 100 times. Pearson's correlation analysis was used to investigate the links of residual sugar content and actual alcoholic strength with NMR bins in Excel (Microsoft Office Excel 2016).

3. Results and discussion

3.1. Metabolite profiling

A representative one-dimensional (1D) ^{1}H NMR spectrum acquired from a pooled sample of white and red wines is shown in Fig. 1. ^{1}H NMR spectroscopy is based on the nuclear magnetic resonance properties of the hydrogen nucleus. Each hydrogen nucleus within a molecule experiences a slightly different magnetic field because of its distinct chemical environment and absorbs energy at slightly different frequencies. Thirty-eight molecular structures were assigned and identified based on the analysis of 1D and 2D NMR spectra (Table 1) and comparisons with previous reports (Ali et al., 2011; Amargianitaki & Spyros, 2017; Consonni, Cagliani, Guantieri, & Simonato, 2011; Du et al., 2007; Larsen et al., 2006; Papotti et al., 2013). The compound classes comprised amino acids, carbohydrates, organic acids, alcohols and phenols.

A ^{1}H NMR spectrum is usually divided into three regions. The area between 0.8 and 4.0 ppm corresponds to amino acids as well as a few organic acids; peaks from alcohols (isobutanol, isopentanol, ethanol and methanol), aliphatic organic acids (lactic, pyruvic, acetic, and succinic), and amino acids (aspartate, methionine, alanine, γ-aminobutyric acid, glutamate and proline) were observable. The region between 4.0 and 6.0 ppm is considered the region for carbohydrate protons; spectral distinction is extremely difficult due to peak overlap, even in the 2D spectra. Fructose and glucose were identified. The aromatic spectral region between 6.0 and 8.5 ppm is one of the most interesting because it includes the aromatic compounds that characterize different wines. Gallic acid, catechin, phenylacetate, phenylalanine and *p*-hydroxyphenylacetic acid were observed. The profiling was designed to obtain a representative description of the spectra to characterize the signals in the bins.

3.2. Principal component analysis

PCA reduces the dimensions of the original data matrix by linear combinations of a starting set of variables based on their maximum variance. Loading vectors are associated with each class, correlated with the original variables, and orientated toward the direction in which the maximum variance of variables is expressed in order to highlight possible differences or similarities among samples (Liland, 2011).

After excluding 14 strong outliers, PCA analysis was performed on the data matrix of 911 samples classified by types. The PCA score plot showed clear separation between the red and white wine clusters, with the first two principal components explaining 10.1% and 7% of the variation, respectively (Fig. 2A). PCA also showed potential for classifying types of dry and medium dry, medium, sweet white and dry red wines. In the PCA cluster (Fig. 2A) describing white wines, three main sub-clusters were identified for dry and medium dry, medium and sweet white wines. The plot of the rosé and blanc de noir samples overlapped with that of dry and medium white wines. By contrast, relative dispersion of the sweet wine samples was observed, reflecting the varied nature of the samples. The loading plot (Fig. 2B) revealed the bins contributing to the differentiation. The loading plot showed high levels of bins containing signals from uridine (bin at 5.86) and unknown signals (bins at 4.38, 7.78) distinguishing white wines; red wines were discriminated by phenylalanine and phenylacetate (bin at 7.34), *p*-hydroxyphenylacetic acid, tyrosine and catechin (bins at 6.82, 6.86),

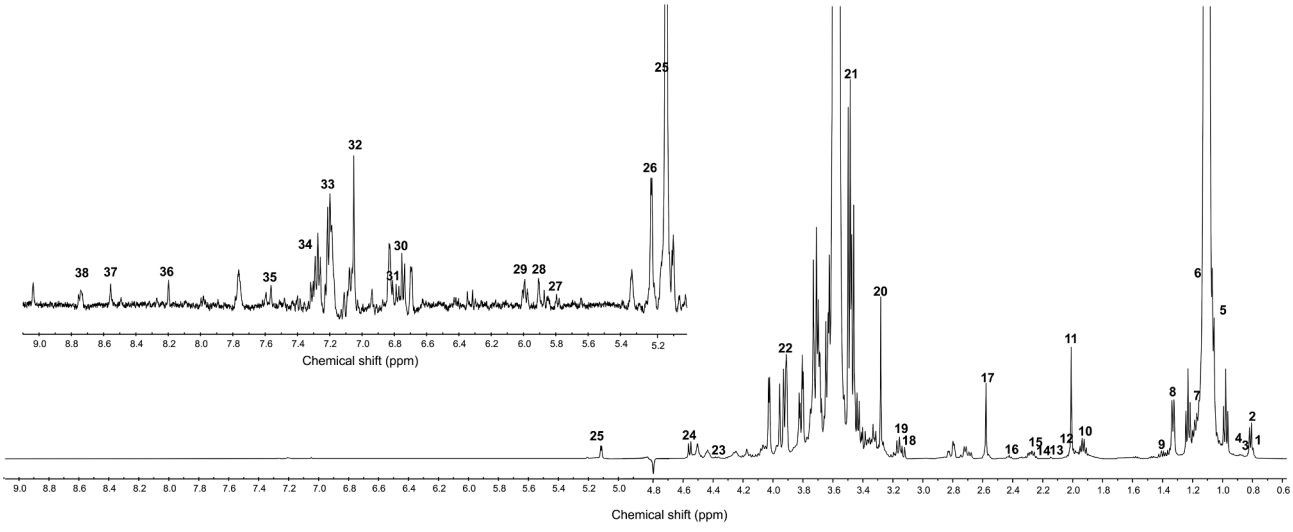

Fig. 1. ^{1}H NMR spectrum of a pooled wine sample (red, white and rosé) with metabolite assignment (Table 1).

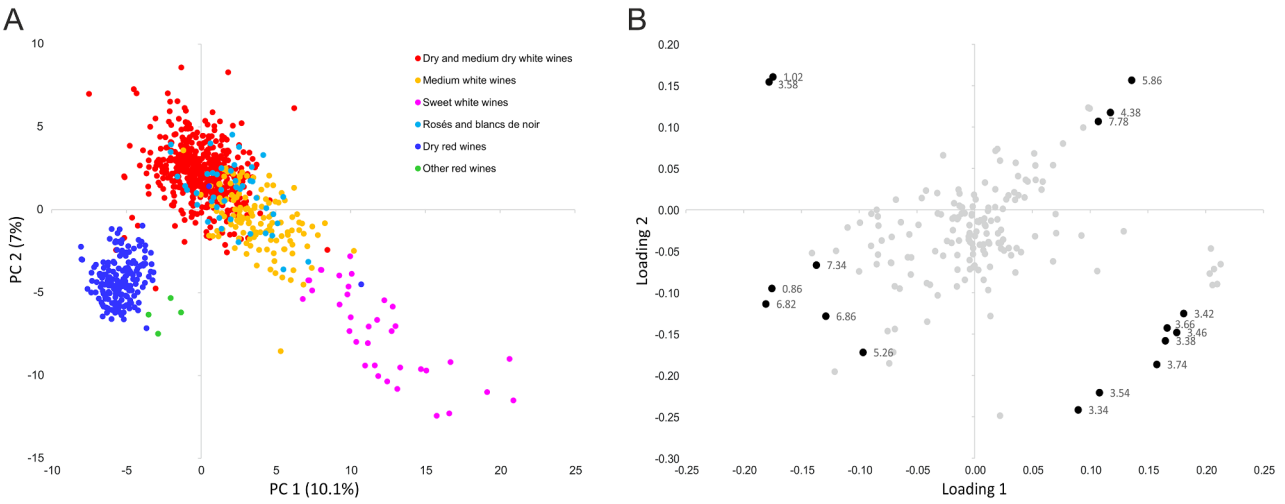

Fig. 2. PCA score plot (A) and loadings plot (B) derived from ^{1}H NMR spectra of Czech wines. Red dots: dry and medium dry white wines; yellow dots: medium white wines; pink dots: sweet wines; light blue dots: rosés and blancs de noir; blue dots: dry red wines; green dots: other red wines. The loadings in (B) represent uridine (bin 5.86), phenylalanine and phenylacetate (bin 7.34), *p*-hydroxyphenylacetic acid, tyrosine and catechin (bins 6.82, 6.86), turanose (bin 5.26), isobutanol, valerate and isopentanol (bin 0.86), glycerol and ethanol (bins 3.74, 3.56), sugars including glucose and fructose (bins 3.38–3.74) and unknown signals (bins 4.38, 7.78).

turanose (bin at 5.26) and isobutanol, valerate and isopentanol (bin at 0.86). White wines were separated into dry and medium dry, medium and sweet by bins by signals from glycerol and ethanol (bins at 3.74, 3.56) and sugars including glucose and fructose (bins 3.38–3.74).

3.3. Hierarchical clustering of wine grape varieties

A heatmap is a data visualization tool that uses colour for a set of parameters and an associated score on multiple objectives. It is a two-dimensional array in which the dimensions are groups and bins and groups are clustered based on similarity (Liland, 2011). In the heatmap, each row is a bin ranked by ANOVA, and each column is a category group average ordered by HC. HC is an unsupervised method that recognizes and distributes data groupings according to their affinity in clusters of progressive dissimilarity. These clusters are presented as a dendrogram in which it is assumed that closer objects in a space defined by variables have greater similarity in their properties. HC was conducted to verify the classification of types according to their mutual dissimilarities. The resulting heatmap with the dendrogram and main descriptive values is shown in Fig. 3A. Consistent with the results of PCA, grouping of the Czech wines based on types revealed two distinct clades. The first clade contained red wines, and the second included white wines, rosés and blancs de noir. Rosés and blancs de noir belonged to the same clade as dry and medium white wines. The features included in the heatmap were tentatively identified, and the colour coding revealed major intensities of red wines in bins with signals assigned to uridine (bins 5.86, 7.86), isopentanol (bin 0.86), tyrosine/phenols (bin 6.82), and acetate (bin 2.02), confirming the PCA results.

The second heatmap showed clustering of wine grape varieties that were represented by more than 20 samples in our sample set (Fig. 3B). Czech wines were split into two distinct clades corresponding to wine grape varieties from red grapes and those from white grapes. Pinot noir and Cabernet Sauvignon belonged to the same clade, which was clustered with Blaufränkisch. This clustering was mainly based on the lower signal intensities of uridine (bin 5.86) and tartrate (bin 0.86) and the higher signal intensities of turanose (bin 5.26), catechin, *p*-

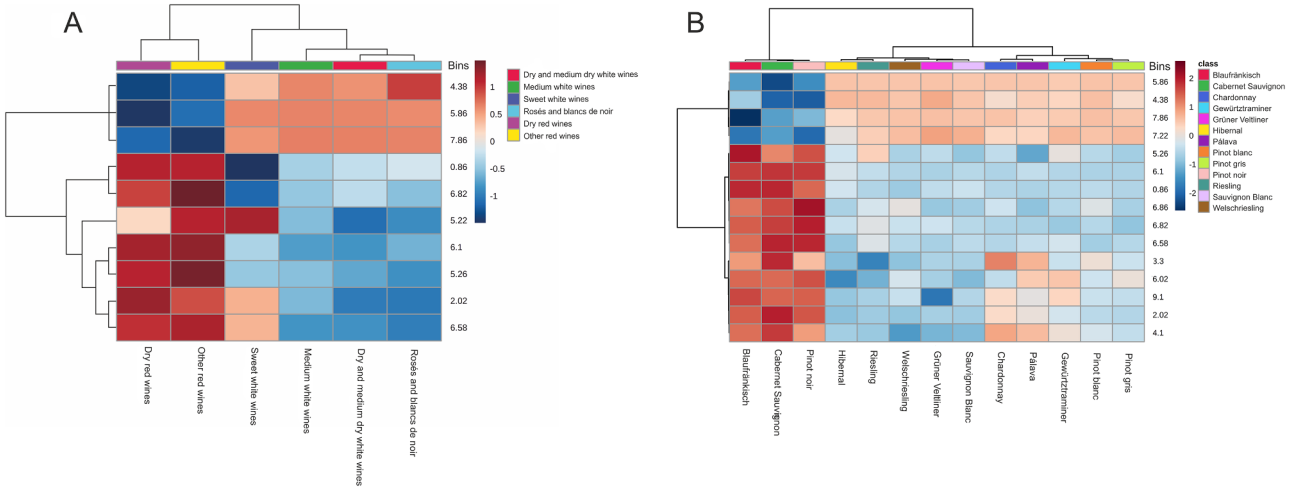

Fig. 3. Hierarchical clustering derived from ^{1}H NMR spectra of Czech wines by types (A) and wine grape varieties (B). The heatmap colour reflects the normalized intensity of the bins: blue: inferior; red: superior; grey: equal intensity. Clustering was performed on normalized and autoscaled binned-aligned ^{1}H NMR spectra using the Pearson distance and cluster aggregation based on the average linkage method. Groups were clustered using group means. The features were ranked by ANOVA. Uridine (bins 5.86, 7.86), isopentanol (bin 0.86), acetate (bin 2.02), turanose (bin 5.26), catechin, tyrosine and *p*-hydroxyphenylacetic acid (bins 6.08, 6.10, 6.82, 6.86), isobutanol and isopentanol (bin 0.86) and tartrate (bin 4.38).

hydroxyphenylacetic acid (bins 6.02, 6.10, 6.82, 6.86), isobutanol and isopentanol (bin 0.86). With respect to white wine grape varieties, the heatmap showed similarities between Pinot gris and Pinot blanc, which is a white mutation of Pinot gris; these wine grape varieties differed in bins 6.86, 6.02 and 3.3. Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir (Vezzulli et al., 2012), but here they showed no significant similarities with this predecessor. Hibernal and Riesling belonged to the same clade; interestingly, Hibernal is a genetic crossbreed of the Seibel and Riesling wine grape varieties, but this clade also includes Grüner Veltliner, which originated from a natural cross involving Sauvignon Blanc.

3.4. Random forest analyses for type and wine grape variety

The RF algorithm is an ensemble learning method used for classification that generates a combination of decision trees using a bootstrap sample (Breiman, 2001). One of the main advantages of RF is its ability to assess the relative importance of variables describing the target problem (Breiman, 2001; Gómez-Meire et al., 2014; Menze et al., 2009). First, the RF model was trained to identify wines by type; the training and test sets were composed of 635 and 268 randomly assigned ^{1}H NMR spectra, respectively. To reduce the risk of over-fitting, the test set was not used to construct the model. Internal validation of the RF model resulted in a rather satisfying OOB error (8.82%), where the classification error was always higher for rosés and blancs de noir. Rosés and blancs de noir were classified by the model as dry, medium dry or medium white wines, as shown in the confusion matrix (Supplementary Material Table S3) and the classification performance (Supplementary Material Table S6). The red wine types classified as "other" were represented by a relatively low number of samples; for a robust model, a higher number of samples is needed. The model classified white dry and medium dry, medium and sweet wines with greater than 93% correctness, with sensitivity values of 0.95, 0.89 and 1 and specificity values of 0.93, 0.95, and 0.98, respectively. Among dry red wines, 99.93% were correctly classified. The most important features for type prediction as ranked by Mean Decrease Gini (Supplementary Material Figure S2A) were bins containing signals of fructose (bins 3.78, 3.86, 4.06, 3.82, 3.7 ppm), uridine (bin 5.86 ppm), ethanol (bin 1.02 ppm), catechin (bins 5.98 and 6.06 ppm), tyrosine (bin 3.5 ppm) and glycerol (bin 3.5 ppm).

Given the high accuracy in classifying red and white wines, a second RF model was trained to discriminate 13 wine grape varieties (including at least 20 samples) using a total of 679 ^{1}H NMR spectra. The training and test sets were composed of 459 and 201 ^{1}H NMR spectra, respectively. To reduce the risk of over-fitting, the test dataset was not used to construct the model. An OOB error of 40.31% was obtained for this model. The trained RF model for wine grape varieties classified Pinot noir with 96% correctness, Blaufränkisch 96%, Riesling 92%, Cabernet Sauvignon 77%, Chardonnay 76%, Gewürztraminer 60%, Hibernal 60%, Grüner Veltliner 52%, Pinot gris 48%, Sauvignon Blanc 45%, Pálava 40%, Welschriesling 28% and Pinot blanc 13% (Supplementary Material Table S4 and Table S7). The most important features as ranked by Mean Decrease Gini (Supplementary Material Figure S2B) were the bins containing signals of proline (bins 1.98, 2.3, 3.3), phenylalanine (bin 7.46), methanol (bin 3.3), catechin (bins 6.02, 6.06, 6.9), tyrosine (bin 7.18) and epicatechin (bins 6.06, 7.02). Pálava was strongly misclassified by the model as Chardonnay; the heatmap (Fig. 3) clustered both wine types in the same clade as Gewürztraminer. Pinot blanc was mainly misclassified as Chardonnay and Pinot gris, and Pinot gris was mainly misclassified as Welschriesling.

^{1}H NMR has previously been reported to be a powerful method for assessing wine grape variety, although in much smaller studies than ours. Amargianitaki & Spyros (2017) used ^{1}H NMR spectroscopy combined with multivariate statistical analysis to classify different wine grape varieties cultivars grown in Germany. The wine grape varieties Pinot noir, Lemberger, Pinot blanc/Pinot gris, Müller-Thurgau, Riesling, and Gewurztraminer were successfully classified. The compounds responsible for differentiation in their study were shikimic acid, caftaric acid, and 2,3-butanediol (Godelmann et al., 2013). Wine grape varieties are known to carry specific molecular signatures. A previous characterization showed that Riesling wines have higher levels of catechin, caftarate, valine, proline, malate, and citrate (Ali et al., 2011). Clear separation of Cabernet Sauvignon and Shiraz wines based on their respective metabolite profiles has also been reported; Cabernet Sauvignon and Shiraz wines produced in Australia were separated by higher levels of proline in the Cabernet Sauvignon (Rochfort et al., 2010).

Chardonnay and Pinot blanc are two wine grape varieties of grape that represent a challenge for accurate classification. They are often mistaken, incorrectly classified by growers or incorrectly labelled. The two wine grape varieties are characterised by ampelographic similarities and historically have often been replaced for each other; in winemaking, they are often interchanged intentionally or by mistake. The RF model was used to classify Chardonnay (70 samples) and Pinot blanc (47 samples). The training and test sets were composed of 82 and 35 ^{1}H NMR spectra, respectively, and the test dataset was not used to construct the model in order to reduce the risk of over-fitting. The RF OOB error was 23.17%. The trained RF model classified Chardonnay with 81% correctness and Pinot blanc with 71% correctness (Supplementary Material Table S5 and Table S8). The most important features ranked by Mean Decrease Gini bins were bins at 1.98, 2.30 and 3.30 containing acetic acid and unknown signals.

3.5. Sub-region classification by PLS-DA

PLS-DA is a linear classification model based on the principal least squares regression algorithm. The model searches for latent variables (LVs) with maximum covariance across bins (Liland, 2011; Römisch et al., 2009). PLS-DA was used to classify the origins of the three most represented wine grape varieties: Chardonnay ($Q^2 < 0$, $R^2 = 0.5$), Riesling ($Q^2 < 0$, $R^2 = 0.6$) and Pinot noir ($Q^2 < 0$, $R^2 = 0.8$) (Fig. 4). For all wine grape varieties, an unclassifying prediction model was built. The winemaking sub-regions are located next to each other, and the samples are thus expected to show a gradient rather than

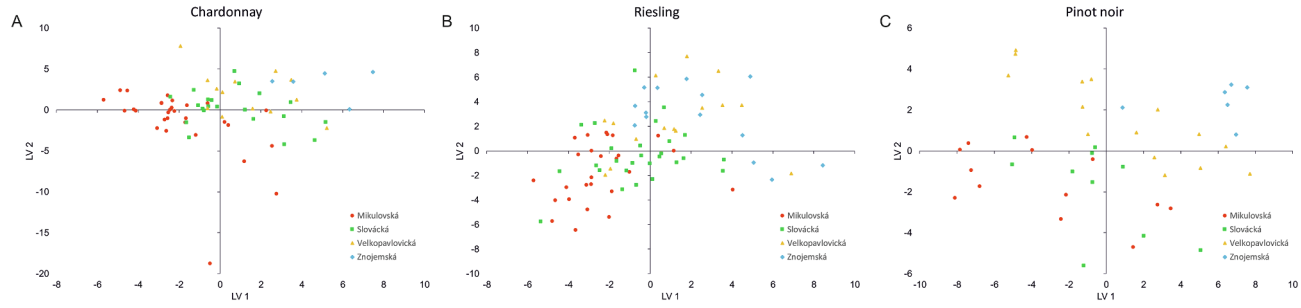

Fig. 4. PLS-DA score plots of LV1 and LV2 of Chardonnay (A), Riesling (B) and Pinot noir (C) from four sub-regions of Moravia.

distinct classes. However, on average, the sub-regions are characterized by different soil conditions, winemaking practices, grape culture and water management. In the Znojemská sub-region, grapes are grown on clay and humic clayey soil, while limestone soil prevails in the Mikulovská sub-region. Precipitation also differs slightly. NMR has been shown to be a suitable tool to investigate the influence of the geographic region where grapes are grown (Amargianitaki & Spyros, 2017). The main factors contributing to these classifications are environmental parameters such as soil geology and composition, climate, water availability, and light exposure. The origins of wines can be differentiated based on the content of succinic acid, sugars such as glucose and fructose and glycerol (Viggiani & Morelli, 2008; Mazzei et al., 2010; Ali et al., 2011; Godelmann et al., 2013). NMR spectroscopy and chemometrics were previously used to classify mature grape berries produced in four different regions in Bordeaux (France) (Pereira et al., 2005).

3.6. Correlations with actual alcoholic strength and residual sugar content

NMR is a highly quantitative analytic method with a wide linear signal range spanning more than 6 orders of magnitude (Mo & Raftery, 2008). Several quantitative NMR methods have been introduced and validated for the determination of natural compounds, biocides and alcohols (Maniara, Rajamoorthi, Rajan, & Stockton, 1998; Okaru et al., 2020; Un & Goren, 2017). In the process of introducing a wine to the market called classification, winemakers must provide a certificate of analysis for each batch that includes actual alcoholic strength and residual sugar content. These analyses are performed by commercial analytical services provided by certified laboratories. These metadata were correlated with NMR spectral bins using Pearson's correlation. As shown in Fig. 5A, there was only a moderate correlation with the declared actual alcoholic strength, r = 0.75, possibly due to either changes during maturation in the bottle.A tolerance of up to 0.5 0.8% v/v is permitted by the EU law (EU 2019/33). By contrast, the residual sugar content declared by the certified analysis was highly correlated with the sum of bins including signals from α-glucose and fructose, with r = 0.98 (Fig. 5B). The negative correlation between bins including signals from sugars and ethanol is the effect of alcoholic fermentation.

4. Conclusion

In this study, we investigated the capability of ^{1}H NMR spectroscopy for quality control of Czech wines. ^{1}H NMR spectroscopy coupled with advanced data analysis and chemometrics was effective in the classification of wine grape variety, type, and, in part, region.

The large dataset of 917 wine spectra allowed us to train a random forest model, capable of simultaneous classification of each wine type and wine grape variety with high correctness, particularly for Pinot noir, Blaufränkisch, Riesling, Cabernet Sauvignon, Chardonnay, Gewürztraminer, Grüner Veltliner, Pinot gris, Sauvignon Blanc, Pinot blanc and Welschriesling. The frequently interchanged wine grape varieties Chardonnay and Pinot blanc could be distinguished with

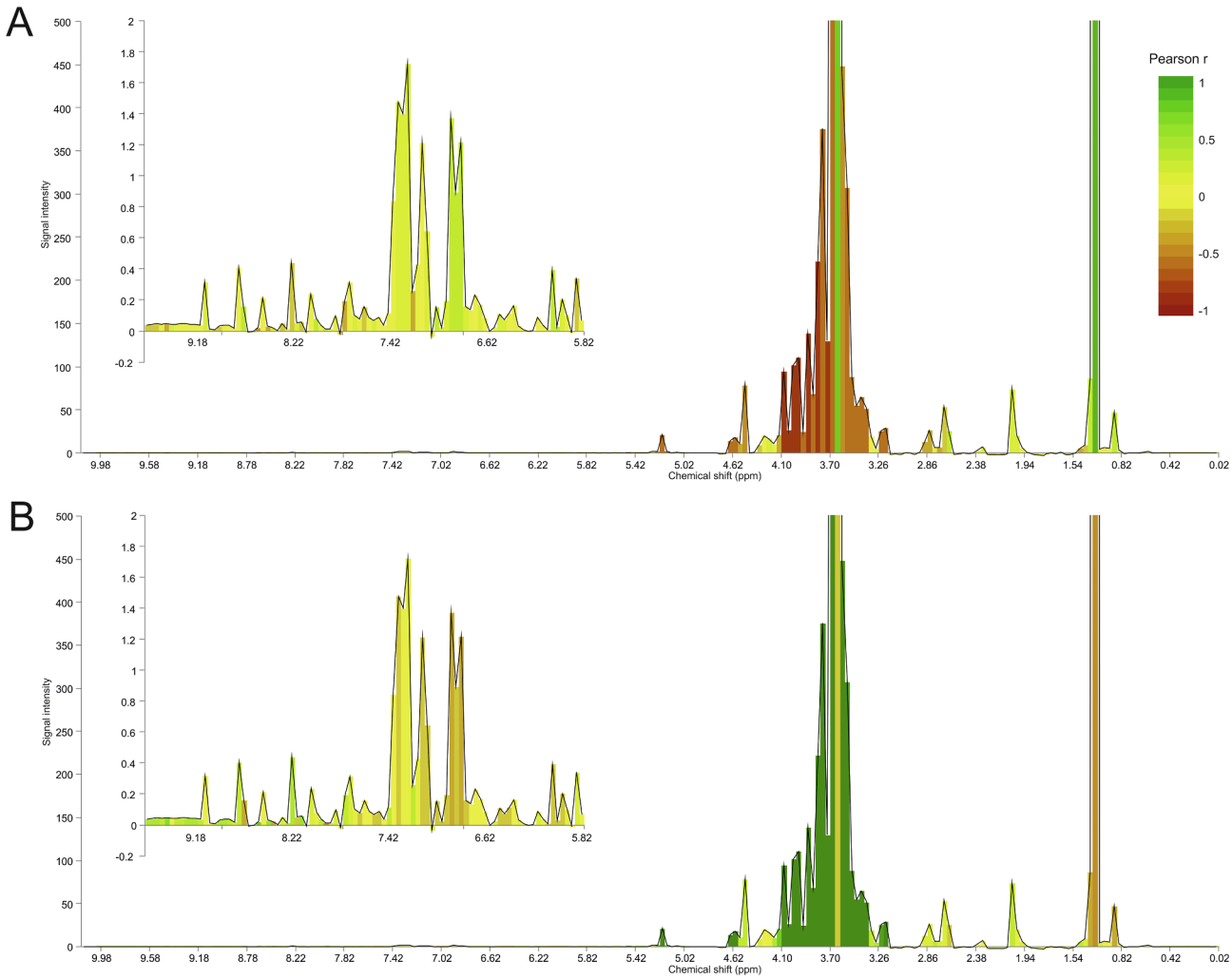

Fig. 5. Binned ^{1}H NMR spectrum colour-coded according to Pearson's correlation coefficient with residual sugar content and actual alcoholic strength metadata.

moderate correctness. Moreover, the model could partially distinguish between two sub-regions. Metabolomics not only provides a better understanding of wine character but, more importantly, is also a facile and rapid tool for assessing many aspects of wine quality. The application of the random forest algorithm for the purpose and scale of wine authentication of this study represents a novel, yet unexplored state-of-the-art approach with promising potential for Food Authorities.

CRediT authorship contribution statement

Anna Mascellani: Methodology, Validation, Investigation, Formal analysis, Data curation, Writing - original draft. **Gokce Hoca:** Investigation. **Marek Babisz:** Resources. **Pavel Krska:** Resources. **Pavel Kloucek:** Writing - review & editing, Project administration, Funding acquisition. **Jaroslav Havlik:** Methodology, Validation, Conceptualization, Formal analysis, Writing - review & editing, Supervision, Project administration.

Declaration of Competing Interest

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 paper.

Acknowledgments

This work was supported by METROFOOD-CZ research infrastructure project (MEYS Grant No: LM2018100) including access to its facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2020.127852.

References

Ali, K., Maltese, F., Toepfer, R., Choi, Y. H., & Verpoorte, R. (2011). Metabolic characterization of Palatinate German white wines according to sensory attributes, varieties, and vintages using NMR spectroscopy and multivariate data analyses. *Journal of Biomolecular NMR, 49*, 255–266. https://doi.org/10.1007/s10858-011-9487-3.

Amargianitaki, M., & Spyros, A. (2017). NMR-based metabolomics in wine quality control and authentication. *Chemical and Biological Technologies in Agriculture, 4*, 1–12. https://doi.org/10.1186/s40538-017-0092-x.

Borák, J., & Vacek, T. (2018). Czech foreign wine trade - Comparative advantages distribution in relation to the European Union. *Agris On-Line Papers in Economics and Informatics, 10*, 31–43. https://doi.org/10.7160/aol.2018.100303.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1201/9780429469275-8.

Brescia, M. A., Košir, I. J., Caldarola, V., Kidrič, J., & Sacco, A. (2003). Chemometric classification of Apulian and Slovenian wines using 1H NMR and ICP-OES together with HPICE data. *Journal of Agricultural and Food Chemistry, 51*, 21–26. https://doi.org/10.1021/jf0206015.

Chong, J., Yamamoto, M., & Xia, J. (2019). MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites, 9, 57-87. https://doi.org/10.3390/metabo9030057.

Commission Delegated Regulation (EU) No 2019/33 of 17 October 2018 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards applications for protection of designations of origin, geographical indications and traditional terms in the wine sector, the objection procedure, restrictions of use, amendments to product specifications, cancellation of protection, and labelling and presentation [2018] OJ L9.

Consonni, R., Cagliani, L. R., Guantieri, V., & Simonato, B. (2011). Identification of metabolic content of selected Amarone wine. *Food Chemistry, 129*, 693–699. https://doi.org/10.1016/j.foodchem.2011.05.008.

Du, Y. Y., Bai, G. Y., Zhang, X., & Liu, M. L. (2007). Classification of wines based on combination of 1H NMR spectroscopy and principal component analysis. *Chinese Journal of Chemistry, 25*, 930–936. https://doi.org/10.1002/cjoc.200790181.

Godelmann, R., Fang, F., Humpfer, E., Schütz, B., Bansbach, M., Schäfer, H., & Spraul, M. (2013). Targeted and nontargeted wine analysis by 1H NMR spectroscopy combined with multivariate statistical analysis. differentiation of important parameters: Grape variety, geographical origin, year of vintage. *Journal of Agricultural and Food Chemistry, 61*, 5610–5619. https://doi.org/10.1021/jf400800d.

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., & Fdez-Riverola, F. (2014). Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International, 60*, 230–240. https://doi.org/10.1016/j.foodres.2013.09.032.

Gougeon, L., da Costa, G., Guyon, F., & Richard, T. (2019). 1H NMR metabolomics applied to Bordeaux red wines. *Food Chemistry, 301*, Article 125257. https://doi.org/10.1016/j.foodchem.2019.125257.

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., & O'Donovan, C. (2020). MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Research, 48*, D440–D444. https://doi.org/10.1093/nar/gkz1019.

Hong, Y. S. (2011). NMR-based metabolomics in wine science. *Magnetic Resonance in Chemistry, 49*, S13–S21. https://doi.org/10.1002/mrc.2832.

International Organisation of Vine and Wine (2013). *International list of vine varieties and their synonyms.* Paris: International Organisation of Vine and Wine.

International Organisation of Vine and Wine (2015). *International Organisation of Vine and Wine International Standard for the Labelling of Wines.* Paris: International Organisation of Vine and Wine.

International Organisation of Vine and Wine (2019). *Compendium of International Methods of Wine and Must Analysis.* Paris: International Organisation of Vine and Wine.

Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International, 122*, 25–39. https://doi.org/10.1016/j.foodres.2019.03.063.

Johanningsmeier, S. D., Harris, G. K., & Klevorn, C. M. (2016). Metabolomic Technologies for Improving the Quality of Food: Practice and Promise. *Annual Review of Food Science and Technology, 7*, 413–438. https://doi.org/10.1146/annurev-food-022814-015721.

Larsen, F. H., van den Berg, F., & Engelsen, S. B. (2006). An exploratory chemometrics study of 1H NMR spectra of table wines. *Journal of Chemometrics, 20*, 198–208. https://doi.org/10.1002/cem.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*, 18–22.

Liland, K. H. (2011). Multivariate methods in metabolomics - from pre-processing to dimension reduction and statistical analysis. *TrAC - Trends in Analytical Chemistry, 30*, 827–841. https://doi.org/10.1016/j.trac.2011.02.007.

Maniara, G., Rajamoorthi, K., Rajan, S., & Stockton, G. W. (1998). Method performance and validation for quantitative analysis by 1H and 31P NMR spectroscopy. Applications to analytical standards and agricultural chemicals. *Analytical Chemistry, 70*, 4921–4928. https://doi.org/10.1021/ac980573i.

Mazzei, P., Francesca, N., Moschetti, G., & Piccolo, A. (2010). NMR spectroscopy evaluation of direct relationship between soils and molecular composition of red wines from Aglianico grapes. *Analytica Chimica Acta, 673*, 167–172. https://doi.org/10.1016/j.aca.2010.06.003.

Mazzei, P., Spaccini, R., Francesca, N., Moschetti, G., & Piccolo, A. (2013). Metabolomic by 1H NMR spectroscopy differentiates "fiano di Avellino" white wines obtained with different yeast strains. *Journal of Agricultural and Food Chemistry, 61*, 10816–10822. https://doi.org/10.1021/jf403567x.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics, 10*, 1–16. https://doi.org/10.1186/1471-2105-10-213.

Minoja, A. P., & Napoli, C. (2014). NMR screening in the quality control of food and nutraceuticals. *Food Research International, 63*, 126–131. https://doi.org/10.1016/j.foodres.2014.04.056.

Mo, H., & Raftery, D. (2008). Solvent Signal as an NMR Concentration Reference. *Analytical Chemistry, 80*, 9835–9839. https://doi.org/10.1021/ac801938j.

Národní vinařské centrum, O. p. . (2020). Classification of wines in Czech Republic. Retrieved from https://www.cmb-brno2020.cz/ website: https://www.cmb-brno2020.cz/en/viticulture-in-cr/classification-of-wines/.

Okaru, A. O., Scharinger, A., Rajcic de Rezende, T., Teipel, J., Kuballa, T., Walch, S. G., & Lachenmeier, D. W. (2020). Validation of a Quantitative Proton Nuclear Magnetic Resonance Spectroscopic Screening Method for Coffee Quality and Authenticity (NMR Coffee Screener). *Foods, 9*(1), 47. https://doi.org/10.3390/foods9010047.

Papotti, G., Bertelli, D., Graziosi, R., Silvestri, M., Bertacchini, L., Durante, C., & Plessi, M. (2013). Application of one- and two-dimensional NMR spectroscopy for the characterization of protected designation of Origin Lambrusco wines of modena. *Journal of Agricultural and Food Chemistry, 61*, 1741–1746. https://doi.org/10.1021/jf302728b.

Pereira, G. E., Gaudillere, J. P., Van Leeuwen, C., Hilbert, G., Lavialle, O., Maucourt, M., ... Rolin, D. (2005). 1H NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux, France. *Journal of Agricultural and Food Chemistry, 53*, 6382–6389. https://doi.org/10.1021/jf058058q.

Rochfort, S., Ezernieks, V., Bastian, S. E. P., & Downey, M. O. (2010). Sensory attributes of wine influenced by variety and berry shading discriminated by NMR metabolomics. *Food Chemistry, 121*, 1296–1304. https://doi.org/10.1016/j.foodchem.2010.01.067.

Römisch, U., Jäger, H., Capron, X., Lanteri, S., Forina, M., & Smeyers-Verbeke, J. (2009). Characterization and determination of the geographical origin of wines. part iii: Multivariate discrimination and classification methods. *European Food Research and Technology, 230*, 31–45. https://doi.org/10.1007/s00217-009-1141-x.

Scott, I. M., Lin, W., Liakata, M., Wood, J. E., Vermeer, C. P., Allaway, D., ... King, R. D. (2013). Merits of random forests emerge in evaluation of chemometric classifiers by external validation. *Analytica Chimica Acta, 801*, 22–33. https://doi.org/10.1016/j.aca.2013.09.027.

Son, H. S., Ki, M. K., Van Den Berg, F., Hwang, G. S., Park, W. M., Lee, C. H., & Hong, Y. S. (2008). 1H nuclear magnetic resonance-based metabolomic characterization of wines by grape varieties and production areas. *Journal of Agricultural and Food Chemistry,*

56, 8007–8016. https://doi.org/10.1021/jf801424u.
Trimigno, A., Marincola, F. C., Dellarosa, N., Picone, G., & Laghi, L. (2015). Definition of food quality by NMR-based foodomics. *Current Opinion in Food Science, 4*, 99–104. https://doi.org/10.1016/j.cofs.2015.06.008.
Un, İ., & Goren, A. C. (2017). Accurate determination of ethanol in water by qNMR: Validation and uncertainty assessment. *Journal of Chemical Metrology, 11*, 9–15. https://doi.org/10.25135/jcm.2.17.03.035.
Vezzulli, S., Leonardelli, L., Malossini, U., Stefanini, M., Velasco, R., Bitonti, B., & Moser, C. (2012). *Pinot blanc and Pinot gris arose induces as independent somatic In Posidonia oceanica cadmium changes in DNA mutations of and Pinot noir methylation chromatin patterning. 63*, 6359–6369. https://doi.org/10.1093/jxb/err313.
Viggiani, L., & Morelli, M. A. C. (2008). Characterization of wines by nuclear magnetic resonance: A work study on wines from the Basilicata region in Italy. *Journal of Agricultural and Food Chemistry, 56*, 8273–8279. https://doi.org/10.1021/jf801513u.