OENO One
VINE AND WINE OPEN ACCESS JOURNAL
By IVES International Viticulture & Enology Society

DOI: https://doi.org/10.20870/oeno-one.2022.56.3.5374

OPEN ACCESS

ORIGINAL RESEARCH ARTICLE

Data fusion using Multiple Factor Analysis coupled with non-linear pattern recognition (fuzzy k-means): application to Chenin blanc

Mpho Mafata[1,2], Jeanne Brand[2] and Astrid Buica[1,2*]

[1] School for Data Science and Computational Thinking, Stellenbosch University, South Africa
[2] South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Stellenbosch University, South Africa

*correspondence:
astrid.buica@gmail.com

Associate editor:
Tristan Richard

Received:
18 January 2022

Accepted:
22 August 2022

Published:
27 September 2022

This article is published under the **Creative Commons licence** (CC BY 4.0).

Use of all or part of the content of this article must mention the authors, the year of publication, the title, the name of the journal, the volume, the pages and the DOI in compliance with the information given above.

ABSTRACT

Patterns in data obtained from wine chemical and sensory evaluations are difficult to decipher using classical statistics. Coupling data fusion with machine learning techniques could assist in solving these issues and lead to new hypotheses. The current study investigated the applicability of classical and machine learning pattern recognition approaches for oenological applications. A sample set of 23 Chenin blanc wines made from young (< 35 years) and old (> 35 years) vines were analysed (recently bottled (Year 1) and after two years of storage (Year 2)). Sensory (sorting) and chemical (NMR: nuclear magnetic resonance and HRMS: high-resolution mass spectrometry) data were collected. Multiple factor analysis (MFA) was used for the data fusion. Cluster analysis was performed by agglomerative hierarchical clustering (AHC) and fuzzy k-means. Optimal cluster conditions were found for both methods and the cophenetic coefficient was used to assess the confidence of fit. Given the large number of variables, the models were complex. Inconsistent clustering patterns were observed when varying clustering conditions, indicating high similarity between samples. Overall, fuzzy k-means resolved clustering patterns better than AHC and, coupled with data fusion, improved the interpretation of the complex oenological data.

KEYWORDS: sensometrics, chemometrics, Chenin blanc, data fusion, pattern recognition, multiple factor analysis, fuzzy k-means, cluster analysis, high-resolution mass spectrometry, nuclear magnetic resonance

INTRODUCTION

Pattern recognition in complex natural systems, such as wine, requires gathering not of just a large amount of data (and corresponding data variation) but diverse and information-rich data (and corresponding variability of measured parameters). Oenological studies can include variation by measuring several categories of chemical (e.g., volatile and non-volatile composition) and sensory (verbal and non-verbal) data (Stevenson, 2005; Valentin *et al*., 2012). Variability can be included through the use of both targeted and untargeted chemistry data and/or the combination of sensory with chemistry results. Combining these different sets requires appropriate data integration using statistical methods of data fusion and intelligent strategic approaches (Biancolillo *et al*., 2019; Borràs *et al*., 2015; Cocchi, 2019). Methods of data fusion for Oenology are usually supervised, seeking to find optimal discrimination of samples based on the cultivar, origin, or age, among others (Biancolillo *et al*., 2019; Borràs *et al*., 2015; De Carvalho Rocha *et al*., 2020). Not only the methods (chemistry, sensory, and statistical) but the strategy (data handling process) itself is important in such a scenario.

When it comes to discrimination problems in Oenology, some limitations are related to the type of information used and incorporated into the models. Chemical fingerprinting techniques are highly effective for discriminating samples in Oenology for problems such as cultivar (Figueiredo-González *et al*., 2012), origin (Versari *et al*., 2014), authentication (Garrido-Delgado *et al*., 2011), and ageing (Pereira *et al*., 2010). Techniques such as infrared (IR), nuclear magnetic resonance (NMR) (Alañón *et al*., 2015; Amargianitaki and Spyros, 2017; Ghasemi *et al*., 2013; Godelmann *et al*., 2013; Silvestri *et al*., 2014), liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Alañón *et al*., 2015), and gas chromatography-tandem mass spectrometry (GC-MS/MS) (Alañón *et al*., 2015; Seisonen *et al*., 2016) have previously been used. These techniques give results that are complex to model and require pre-modelling treatments (Engel *et al*., 2013; Rinnan *et al*., 2009).

In addition to these chemical fingerprinting techniques, sensory data can be used to model the behaviour of the wine from the sensorial perspective. Although profiling sensory methods such as descriptive analysis (DA) can suffice in discriminating samples (Cayuela *et al*., 2017; Vannier *et al*., 1999), this is not always the case and, thus, sensory data may need to be combined with chemistry data. In these cases, the problem then becomes how to combine the different data sets appropriately - specifically, chemistry with sensory data. Solving this problem requires methods of data fusion which combine and integrate data sets to create representative, information-rich models (Borràs *et al*., 2015; Cocchi, 2019; Seisonen *et al*., 2016).

In the fused models, assessing patterns of behaviour can be done in different ways, i.e., using classical (parametric) and non-classical (non-parametric) methods (Figueiredo-González *et al*., 2012; Myhre *et al*., 2018; Radovanovic *et al*., 2016). Classical methods of pattern recognition are based on linear regression or multiple linear regressions in the case of multivariate techniques (Granato *et al*., 2014; McKillup, 2005a; Salkind and Kristin, 2007). The assumption with classical methods is that there is a normal distribution, but in complex cases of high variation, a normal distribution is not always observed (Granato *et al*., 2014; McKillup, 2005a).

Parametric techniques used for pattern recognition include variants of cluster analysis (agglomerative hierarchical clustering/AHC, hierarchical cluster analysis/HCA), linear regression analysis (linear discriminant analysis/LDA, multiple linear regression/MLR), and discriminant/classification analysis (partial least squares/PLS, discriminant analysis/DA) (McKillup, 2005a). Non-parametric groups of methods for pattern recognition, such as k-nearest neighbours (kNN), artificial neural networks (ANN), and support vector machines (SVM), consider non-binary relationships (De Carvalho Rocha *et al*., 2020; Härdle and Simar, 2015; Myhre *et al*., 2018; Radovanovic *et al*., 2016).

These non-parametric pattern recognition methods are generally used in a supervised manner. Previous studies used kNN for the classification of wine age and found better results compared to the use of parametric linear discriminant analysis (LDA) (Pereira *et al*., 2010). An unsupervised ensemble approach has previously been described to generate robust and reliable clustering results (Myhre *et al*., 2018). The method looked at the effects of the pre-processing step (fitting/scaling the data), the clustering method itself (e.g., meanshift vs kernel density), and adjusting the parameters (e.g., number of clusters and bandwidth) on the reliability of the kNN clustering. Before kNN analysis, data were first scaled and modelled according to the appropriate technique; then, the clustering was applied to the model or its features (Myhre *et al*., 2018).

Fuzzy k-means, originally referred to as fuzzy c-means (Bezdek, 1981), is a variant of kNN that uses fuzzy algorithms for clustering observations (samples). The partitioning (grouping) of the clusters can be based on an average centroid, class membership (group designation), or random assignment of centroids (Bezdek, 1981). This is different from hard clustering techniques such as AHC and classification techniques such as PLS, which impose strict partitioning (Härdle and Simar, 2015; McKillup, 2005a). Of the available machine learning algorithms and based on the case studies by Myher *et al*. (Myhre *et al*., 2018) and studies on fuzzy c-means (Khang *et al*., 2020), fuzzy k-means was considered the most appropriate technique for the case study.

The current study explored an unsupervised strategy consisting of data fusion coupled with pattern recognition. It included variation in the form of combinations of data sets (i.e., chemistry and sensory) and information-rich techniques (i.e., fingerprinting by NMR and HRMS). MFA was used for the fusion of sensory (verbal and nonverbal flexible sorting)

and chemistry (NMR and HRMS) data. Pattern recognition was done using AHC and fuzzy k-means. Fuzzy k-means was explored by varying coefficients of fuzziness and the number of clusters. The effectiveness of the different strategies of cluster analysis was evaluated using coefficients of variance in optimal classification (i.e., Wilks' Lambda and cophenetic correlation coefficient).

MATERIALS AND METHODS

1. Experimental design: winemaking and sensory analysis

The winemaking, wine treatments, as well as sensory evaluation have been previously published (Mafata *et al*., 2020). In brief, the experimental design consisted of 23 wines made from grapes harvested from vines aged 5 to 45 years old and made using the same vinification protocol. Wines are coded as YV (young vines, < 35 years) and OV (old vines, > 35 years), following the South African wine industry convention (Old Vine project, 2020). The wines were evaluated twice: three months in the bottle (Year 1) and two years in the bottle (Year 2). The sensory analysis consisted of two aspects, a flexible sorting task with a non-verbal (grouping) and a verbal (group description) step. The flexible (semi-directed) aspect of the sorting task meant that the wines had to be grouped according to YV (young vine) and OV (old vine), but there was an option to generate another group if necessary. A panel of 15 judges (industry professionals) evaluated the wines by sorting (in a single session for each year) and were asked to provide attributes they associated with each group. Full details are given in Mafata *et al*., 2020).

2. Chemical analysis

2.1. High-resolution mass spectrometry (HRMS)

Sample preparation: The 23 wine samples were clarified through centrifugation at 1000× *g* and sampled into a 2 mL HPLC vial.

Instrumental acquisition: This was done according to a published method (Garrido-Bañuelos *et al*., 2019; Panzeri *et al*., 2020) using a UPLC (Waters Corporation, Milford, MA, USA) equipped with a Synapt G2 quadrupole time-of-flight mass spectrometer (Q-TOF-MS, Waters Corporation, Milford, MA, USA). Separation was done on an Acquity UPLC HSS T3 column (1.8 µm internal diameter, 2.1 mm × 100 mm, Waters Corporation, Milford, MA, USA) at 60 °C using 0.1 % formic acid (mobile phase A) and acetonitrile (mobile phase B) and a scouting gradient. The mass spectrometric acquisition was done at 20 Hz using an electrospray ionisation probe in positive (150 to 600 m/z) and negative (40 to 600 m/z) mode; the cone voltage was set at 15 V, using N2 at 275 °C.

Data processing: Using the MarkerLynx software (Waters Corporation, Milford, MA, USA), the data were integrated and extracted as a retention time (RT) versus mass-to-charge ratio(m/z) matrix with values in terms of ion abundance. The peak detection was done using a standard marker intensity threshold set at 200 counts.

2.2. Nuclear magnetic resonance (NMR) spectroscopy

Sample preparation: The wine pH was adjusted to 3 using 10 M HCl and 10 M NaOH solutions. In a 5 mm Wilmad NMR tube (Rototec-Spintec GmbH, Bad Wildbad, Germany), 50 µL of a 1M tetramethylsilane standard (TMS in D2O) and 500 µL of the pH-adjusted wine were added. Nitrogen gas was passed over the headspace for preservation until analysis.

Instrumental acquisition: NMR was performed using a Varian Unity Inova 600MHz liquid state NMR Spectrometer and based on methods by López-Rituerto *et al*. (2012) and Godelmann *et al*. (2013). All spectra were acquired at 298.0 K. Manual tuning, matching, locking, and shimming were performed for each sample analysis. Two 1-H NMR analyses were performed, a pre-saturation for suppression of the water and ethanol signals followed by a 1-D NOESY experiment at 256 scans per sample (Supplementary Figure S1).

Data processing: was done offline using MestreNova software version 12.1.0 (Mestrelab Research S.L., Santiago de Compostela, Spain). The spectra were aligned using the TMS reference peak at 0.0 ppm. Cuts were done for the saturated peaks for ethanol alkyl groups (1.036 to 1.200 ppm and 3.610 to 3.670 ppm) and the water peak (4.735 to 4.880 ppm). The chemometric toolset was used to extract the spectra as untargeted data. The data were scaled according to the intensity of the TMS peak, then 0.04 ppm spectral bins were applied, and the data were captured based on the sum over the area. A data matrix was extracted with the bins as the variables and the wine samples as the observations and used for further statistical analysis.

3. Statistical analysis

The repeatability of the analyses was confirmed by using instrumental repeats (HRMS) and quality assessment using a known reference sample (NMR).

Exploratory data analysis consisted of Principal Component Analysis (PCA), Correspondence Analysis (CA), and Multidimensional Scaling (MDS). PCA was performed on the correlation matrices for HRMS (samples vs features) and NMR (samples vs bins) data sets based on Pearson correlation. CA was done on the verbal sorting results and MDS on the non-verbal sorting results of the sensory analysis.

Data fusion was performed by multiple factor analysis (MFA) on four data sets (HRMS, NMR, verbal and non-verbal sorting tasks). The exploratory data analyses were used as the first step in the MFA analysis. From the MFA, ordinal data of the group factor map (datasets), individual factor maps (samples), loadings factor maps (variables), and projected biplot maps (samples and variables) were extracted.

Pattern recognition was done on the ordinal matrix from the MFA by AHC and fuzzy k-means using XLSTAT software (version 2020, Addinsoft, New York, USA). AHC followed Ward's method of agglomeration based on the Euclidean distance; the threshold on dissimilarity was set according

to the cophenetic coefficient as a measure used to confine Euclidean distance-based clustering algorithms to a threshold of confidence. Fuzzy k-means clustering was done using Wilks' Lambda based on the Euclidean distance. Partitioning was done randomly with repetitions set at n = number of samples. Additional visualisations (graphical illustrations of the model output) were built using Statistica™ 13 (TIBCO, Dell Software, Inc., Texas, United States).

RESULTS

This section is presented following the manner of statistical execution, in which each step builds on the preceding step. The initial data analysis was done to explore the individual datasets for the purposes of data fusion; the data sets were scrutinised to elucidate any irregular patterns that may be associated with noise that might need to be corrected before fusion. The final models of the individual datasets were then fused by MFA and the features of the data fusion models were extracted for pattern recognition by AHC and fuzzy k-means clustering.

1. Exploratory data analysis

The sensory data sets (non-verbal) and verbal aspects of sorting (supplementary figure S2) were explored in detail in the article by Mafata *et al*. (2020). Here, the performance of the models of the chemistry data sets will be discussed, followed by the interpretation of the models in the context of the case study.

1.1. High-resolution mass spectrometry (HRMS)

The HRMS fingerprint for wine is commonly acquired in the positive mode since more compounds present in wine can be ionised in the positive mode, but metabolomic and untargeted studies use both ionisation modes (Alañón *et al*., 2015). Thus, in this study, the fingerprint included the acquisition in the negative mode since it also contains discriminating features, although fewer than in the positive mode. A block PCA analysis, which keeps the two modes separate, can elucidate the relationship between the two.

The configuration of the samples (Supplementary figure S3) for the two modes was used as an indication of the different information the acquisition modes contained. The configurations derived from the positive and negative modes had RV coefficients of 0.64 (Supplementary figure S3(c)) and 0.62 (Supplementary figure S3(d)) for Year 1 and Year 2, respectively, indicating a moderate similarity in configuration.

The complete HRMS fingerprint was a combination of both MS ionisation modes (positive/Pos. and negative/Neg.). The full fingerprint consisted of 187 and 257 unique features for Year 1 and Year 2, respectively (Supplementary Figure S4). These were then treated as the new variables and modelled by PCA (Figure 1). The new combined models were configurationally more similar to each mode than the modes were to each other. This was indicated by a high RV coefficient, 0.98 and 0.80 (combined vs Pos. and vs Neg.), and 0.99 and 0.73 (combined vs Pos. and vs Neg.) for Year 1 and Year 2, respectively.

The scree plot (Supplementary Figure S5) is an indication of the efficiency of the data model. A steep decline in the stress in Year 2 results implied greater efficiency, a consequence of the higher number of unique features compared to the Year 1 data set (257 vs 187). The inflection point (the point at which the stress begins to plateau) was around the fifth dimension for both years, for which 61 % and 69 % of the total variation can be explained for each year, respectively.

In the context of the original application, although both the separate modes and the combined HRMS data sets provide unique fingerprints, neither could distinguish the old vine samples from the young vine samples. Due to this, there were no markers linked to or which could be used to discriminate the samples according to vine age or class designation (young vine and old vine). As an exploratory analysis, the unsupervised PCA revealed no problematic issues with the

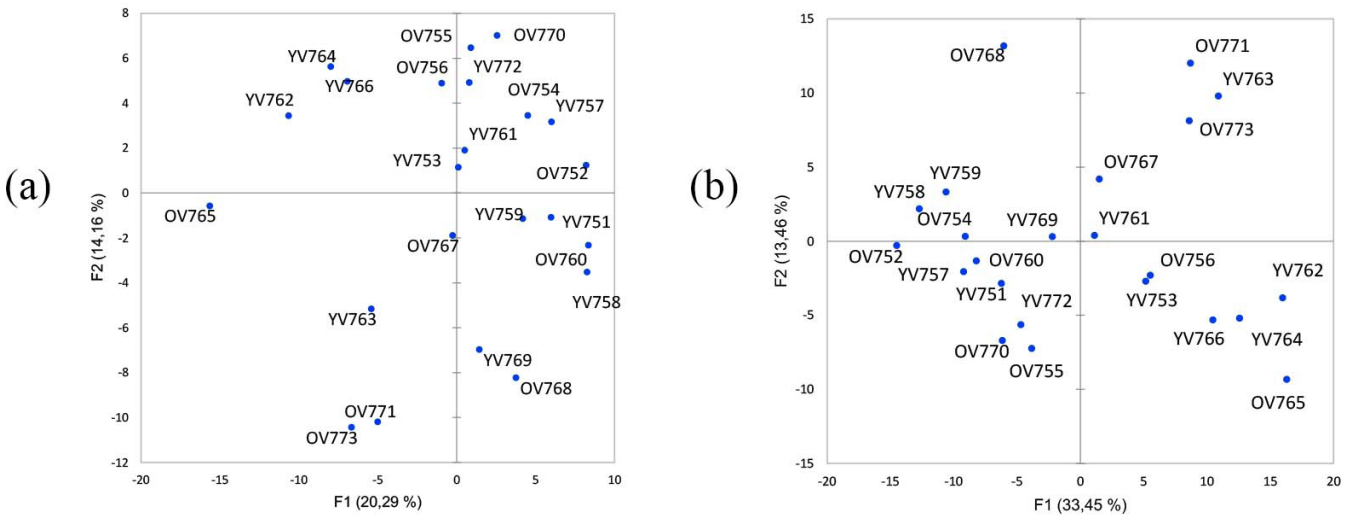


FIGURE 1. Block PCA scores of HRMS in combined (Pos. and Neg.) acquisition modes for Year 1 (a) and Year 2 (b).

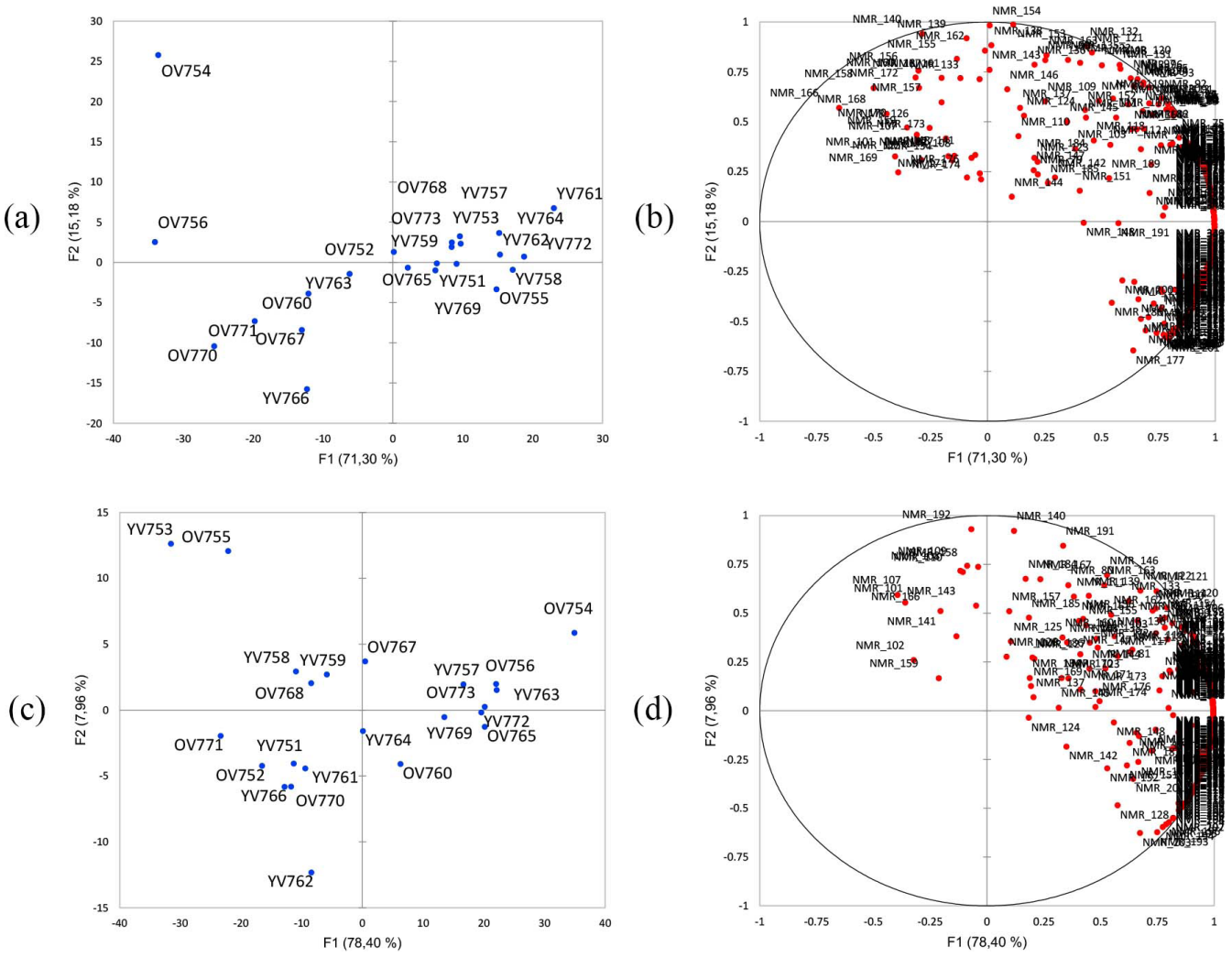


FIGURE 2. PCA models of NMR data for Year 1 (a) scores and (b) loadings, and Year 2 (c) scores and (d) loadings.

HRMS dataset and, thus, no corrective pre-processing was done. Therefore, the combined modes, without any feature selection, were further used for data fusion.

1.2. Nuclear magnetic resonance (NMR) spectroscopy

NMR data were processed to extract the spectra as untargeted data in the form of new variables ("bins"). This resulted in 390 and 392 variables for Year 1 and Year 2, respectively. These new variables were then modelled by PCA, and the efficiency of the model is illustrated by the scree plot in Supplementary Figure S6. The plot showed high efficiency in the models, with the inflection at F3 corresponding to 90 % of the cumulative variance for both years. The NMR configuration map of the samples (Figure 2(a) and 2(c)) showed a particular distribution for samples analysed during Year 1. With the exception of two samples (OV754 and OV756), there is an element of linearity diagonally across the first and second dimensions (at Y = X). The loadings indicate that the linearity may be due to intensity variations (McKillup, 2005a). However, this phenomenon was not observed in Year 2.

To try to find reasons for the linear variation in the Year 1 dataset and for further exploration of the NMR data, the spectra were submitted to block analysis. Three important spectral regions were used and processed separately: the alkyl region (≤ 3.61 ppm), carbohydrate region (3.67 to 4.74 ppm), and aromatic region (≥ 4.88 ppm). PCA on the carbohydrate region for Year 1 (Supplementary figure S7) is the only region that exhibited the linearity seen in the full NMR spectra. Correlated with the linear trajectory are 6 bins: (2.998 to 3.038 ppm, malic acid), (3.08 to 3.078 ppm, citric acid), (3.518 to 3.558 ppm, glycerol/fructose), (4.398 to 4.438 ppm, tartaric acid), (4.438 to 4.478 ppm, lactose), and (4.598 to 4.638 ppm, fucose) tentative assignment based on literature (López-Rituerto *et al.*, 2012; Mascellani *et al.*, 2021). The other regions (alkyl and aromatic) showed no discernible patterns in the spectra (observationally) or PCA (statistically). Year 2 results showed no discernible patterns, with a high similarity between all three regions (RV ranging from 0.70 to 0.91).

Looking at the high RV coefficient values for the full NMR spectra versus the carbohydrate region for Year 1, an additional exploratory approach was considered namely MFA. MFA is

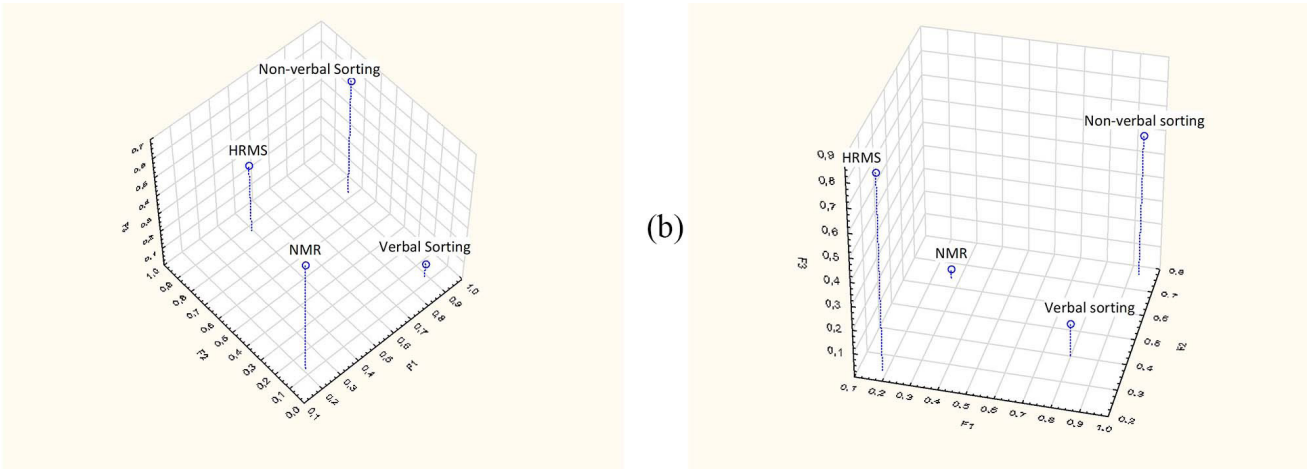


FIGURE 3. MFA group factor map for Year 1 (a) and Year 2 (b) depicted in the first three dimensions.

a multiblock approach that first builds individual PCAs for each block (i.e., NMR region) and then separately scales them according to their eigenvalues before building the final model. By doing this, the method avoids skewing by any one block (Burt, 1947). As the data fusion strategy in this study is also MFA, the NMR scaling issue was explored in two ways: (*i*) doing data fusion with the regions separate and (*ii*) first fusing the regions by MFA, then using the MFA as the representative NMR fingerprint. The latter resulted in increased RV coefficients (RV > 0.85 regions vs MFA), indicating that the issue may have been the differences in the scale between the regions (Supplementary Table S1).

For an unsupervised exploratory aim, the MFA approach would have been optimal; however, in order to advance the aims of the current study in devising a systematic approach to coupling data fusion with pattern recognition, the full spectra were used further without prior block analysis.

Contextually, similar to the HRMS results, neither the scores nor the loadings gave discriminating patterns based on vine age or class designation (young vs old vine wine) for both years, even when scaling by MFA was applied.

1. Data fusion by Multiple Factor Analysis (MFA)

The performance of the data fusion models is indicated by the decay in stress, visualised using the scree plots (Supplementary Figure S8). Year 1 saw a total of 649 observations modelled, while Year 2 had 743 observations due to more HRMS features and verbal sorting terms (49 in Year 1 and 71 in Year 2). Conversely, the Year 1 stress (2.54 eigenvalue) was relatively higher compared to Year 2 (2.18 eigenvalue). This meant that there were factors influencing the efficiency of the data fusion models other than the obvious number of observations. The scree plots showed that the inflection point for Year 1 (F8, corresponding to 53 % cumulative variance) and Year 2 (F9, corresponding to 59 % cumulative variance) were similar. Year 2 had better performance in terms of lower stress (eigenvalue) and higher efficiency (% cumulative variance) as 28 % was packed in the first three dimensions compared to 24 % for Year 1.

As a multiblock approach, the performance of the MFA data fusion models is determined by the blocks and their relationships to one another. The relationship of the four blocks in this case (HRMS, NMR, verbal sorting/VSorting,

TABLE 1. Pairwise RV coefficients for MFA scores configurations between the four individual data sets and the MFA data fusion models for both Year 1 and Year 2 of the study.

Year 1						Year 2					
	HRMS	NMR	VSorting	NVSorting	MFA		HRMS	NMR	VSorting	NVsorting	MFA
HRMS	1	0.178	0.439	0.695	0.817	HRMS	1	0.165	0.264	0.508	0.649
NMR	0.178	1	0.109	0.266	0.419	NMR	0.165	1	0.166	0.277	0.455
VSorting	0.439	0.109	1	0.524	0.657	VSorting	0.264	0.166	1	0.604	0.714
NVSorting	0.695	0.266	0.524	1	0.943	NVSorting	0.508	0.277	0.604	1	0.94
MFA	0.817	0.419	0.657	0.943	1	MFA	0.649	0.455	0.714	0.94	1

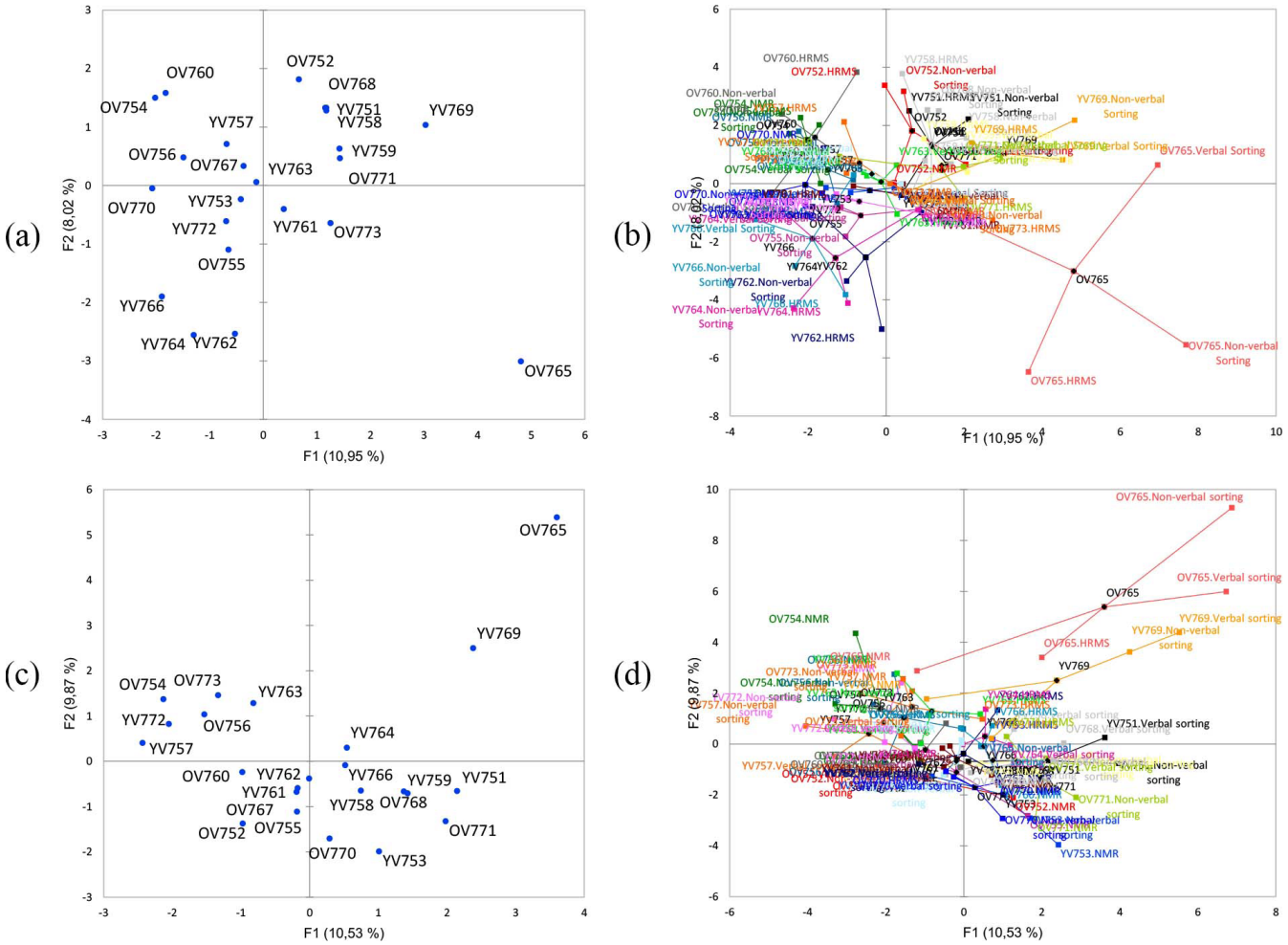


FIGURE 4. MFA results of Year 1 (a) scores and (b) coordinates for the projected points, and Year 2 (c) scores and (d) coordinates for the projected points.

and non-verbal sorting/NVSorting) could be seen in the three-dimensional representation of the group factor map (GFM) from the MFA (Figure 3). There was little overlap between the four data sets, indicating that they contained different (not redundant) information. Accompanying the GFM were pairwise RV coefficients (Table 1) calculated on the scores of each data set. Emphasising the point of lack of redundancy, the RV coefficients were low between the data sets, the highest being HRMS vs NVSorting (0.69) for Year 1 and NVSorting vs verbal sorting (0.60) for Year 2.

The low RV coefficients between the individual data sets indicated unique information contained in each data set and can be considered a positive result for unsupervised data fusion. In data fusion, it is preferable to have not only informational density but also informational variability. Informational variability guarantees low redundancies in the configurational maps of the samples and observations, lowering incidences of false correlations. The MFA data fusion models were representative, indicated by high RV coefficients between the MFA and the individual data sets for both years (Table 1). However, NMR exhibited low RV coefficients versus the MFA for both Year 1 (0.42) and Year 2 (0.45) due to the scaling issues previously discussed in Results section 1.2.

The advantage of data integration is that inferences on discriminating or unique elements can be made using the various compilation of data sets, facilitating comprehensive problem identification and description. This can be exemplified by the identification of the unique samples OV765 (Year 1 and Year 2) and YV769 (Year 1) (Figure 4). The partial axes from the MFA data fusion model (Figure 4(b) and (d)) showed the data sets from which the difference came. The graphs show that in both years, the sensory analysis (VSorting and NVSorting) consistently discriminated the unique samples from the rest. The same differences were noted in the analysis of the sensory results by Mafata *et al.*, 2020). The contextual meaning of these results was also discussed in the cited article.

3. Pattern recognition

3.1. Exploratory pattern recognition strategy by classical statistics and fuzzy k-means

A few patterns have already been highlighted in the previous sections; they were based on exceptions and not on discernible differences between samples according to

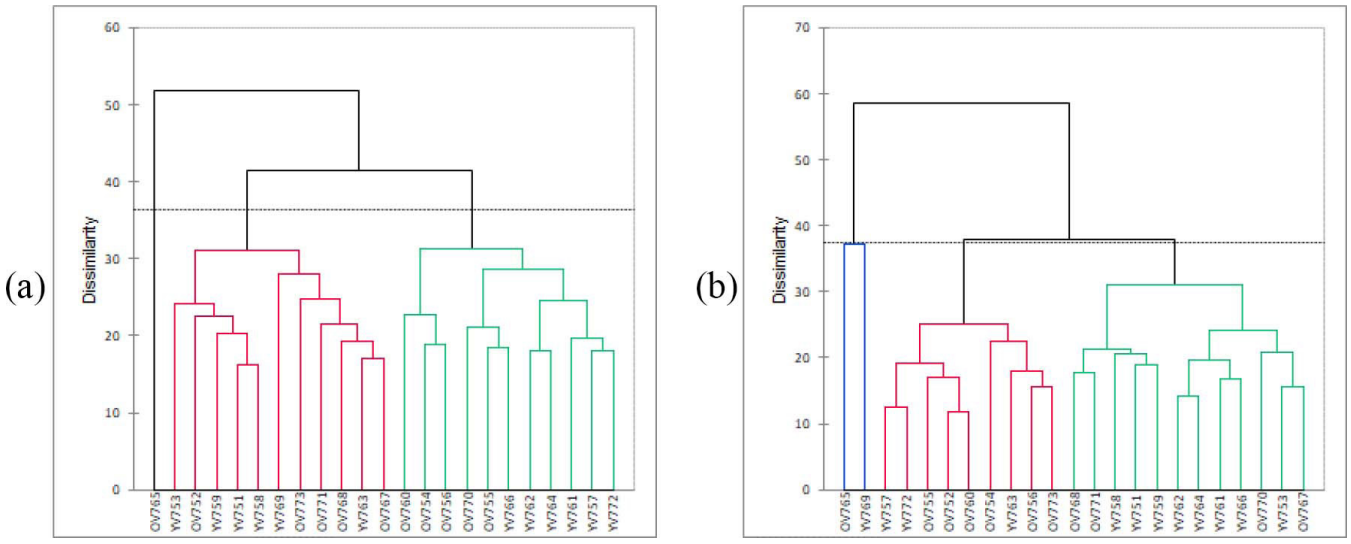


FIGURE 5. Agglomerative hierarchical clustering (AHC) dendrogram on the MFA scores of (a) Year 1 and (b) Year 2.

the vine age (class designation). To try to identify these patterns, an extensive pattern recognition strategy was evaluated next. As previously mentioned in the Introduction, due to its sensitivity, fuzzy clustering is best recommended for attempting to discriminate samples for data with high similarity (Myhre *et al*., 2018; Radovanovic *et al*., 2016).

Pattern recognition in the form of parametric (AHC) and non-parametric (fuzzy k-means) cluster analyses were done on the correlation matrix of the MFA samples. Figure 4 shows two-dimensional representations of the MFA factor maps of the samples. The stress distribution of the MFA was gradual, having 50 % explained variance over eight factors for both years (Supplementary Figure S8). To get a fair representation of the relationship between samples, the AHC dendrogram was plotted for all factors (Figure 5). From the dendrograms, it was evident that sample OV765 for year 1 and samples OV765 and YV769 for year 2 were outliers. Due to the sensitivity of the fuzzy k-means method and the fact that the strategy used in this study is unsupervised (using random partitioning instead of supervised clustering, such as centroid or class membership), these outlying samples were removed before any further clustering was applied. The random partitioning was chosen because: (*i*) from the descriptive study, it was shown that there was no prototype that could be used as the best average example for classification (Mafata *et al*., 2020); (*ii*) the exploratory stages showed no distinguishable pattern between the samples (Results section 1); (*iii*) random partitioning was the most compatible approach to unsupervised clustering.

The outlying samples had a big effect on the clustering; removing the outliers resulted in the cophenetic correlation coefficient decreasing from 0.637 to 0.403 in Year 1 and 0.813 to 0.464 in Year 2. The percentage of the total variation within class was 91 % for Year 1 and 95 % for Year 2. Both the high variation and the low cophenetic correlation coefficient meant that the assignment of members of each cluster was unreliable (Bezdek, 1981).

Next, the fuzzy k-means was investigated by varying the fuzzy partition coefficient (Fk) as well as the number of clusters (k). As this study uses an unsupervised approach, reasonable Fk at the upper limit (1.05) and lower limit (1.001) were explored (Figure 6). Various degrees outside these parameters (i.e., Fk > 1.05) were also investigated, but the two final values were chosen because they illustrated the sensitivity of the fuzzy k-means technique (data not shown).

The objective function of the clustering, from which the goodness-of-fit criterion is based, is to maximise the discrimination between clusters based on Euclidean distance in this case (Bezdek, 1981). The goodness-of-fit criterion in this analysis is analogous to the R^2 used for goodness-of-fit used for PCAs, for instance (Härdle and Simar, 2015). Both Year 1 and Year 2 results showed a high goodness-of-fit criterion at $k < 9$ (Figure 6), but the Wilks' Lambda coefficient values were poor. This meant that although the data was well fitted, the class membership was unreliable. The Wilks' Lambda coefficient was lower (i.e., $\lambda < 1$) at $k \geq 12$ for both Year 1 and Year 2, and the percentage goodness-of-fit criterion increased (Figure 6). When Fk was set at 1.001, both the goodness-of-fit criterion and Wilks' Lambda coefficient improved ($\lambda = 0.085$ for both years) (Supplementary Table S2). Hence, the optimal clustering was with twelve clusters at a fuzzy coefficient of 1.001 for both years. Although the λ values reported in this study were relatively high compared to the parametric AHC, the fuzzy k-means clustering was more reliable than AHC at assigning cluster membership.

Correspondence analysis (CA) was performed on the membership probabilities at the optimal clustering by treating them as categorical/frequency-based data (Figure 7) (McKillup, 2005b), analogous to the use of CA in sensory sorting tasks (Cariou and Qannari, 2018; Valentin *et al*., 2012). The discrimination between samples was better at Fk set at 1.001 than at 1.05. Cluster analysis on the CA showed improved clustering compared to the parametric AHC without fuzzy k-means. Year 1 showed an increase in

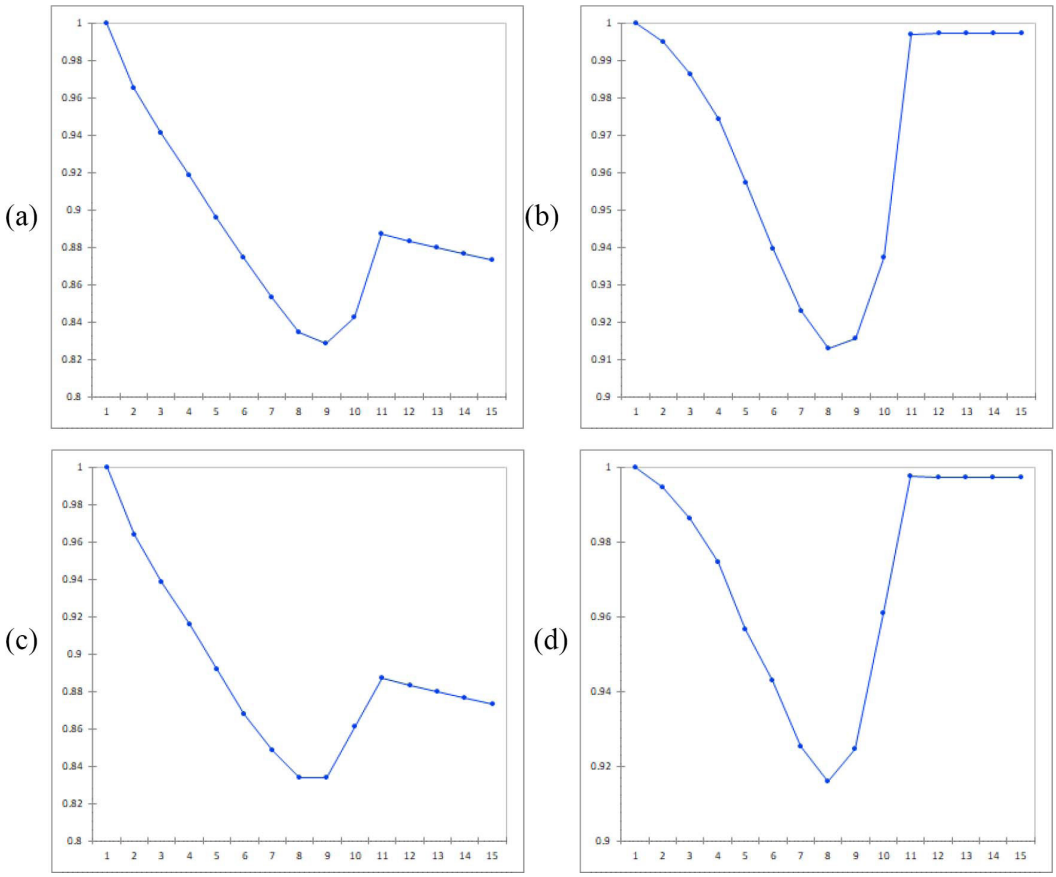


FIGURE 6. Evolution of the goodness-of-fit criterion in the fuzzy k-means analysis of Year 1 (a) with Fk set at the upper limit of 1.05 and (b) the lower limit of 1.001, and Year 2 (c) with Fk set at the upper limit of 1.05 and (d) the lower-limit of 1.001.

the cophenetic correlation coefficient, from 0.403 to 0.869 and the within-class variation decreased from 96 % to 16 % (Supplementary Table S3). For an unsupervised approach, these are markers of reliable clustering and discrimination between samples when using fuzzy k-means (Bezdek, 1981). Similarly, Year 2 showed an increase in the cophenetic correlation coefficient from 0.464 to 0.835 but a minimal decrease in the within-class variation (from 95 % to 87 %) (Supplementary Table S3).

3.2. Contextual interpretation of the cluster analysis

The contextual sensory perspective elements related to this application were investigated based on the typicality of old vine South African Chenin blanc from class designation and age of vines (Mafata *et al*., 2020). The discussion of the results in the previous study showed no indication of suitable prototypes (typical of old vine wine) or a distinguishable age-defined border between the classes (young vine wine versus old vine wine). Based on these findings, the fuzzy clustering approach could not use class membership or centroid partitioning. If a cluster analysis was to be performed for the four individual datasets used in the data fusion (NMR, HRMS, Non-verbal, and Verbal sorting), it would, much like the AHC, show the random distribution of samples demonstrating the high similarity between samples. Both statistical and contextual random effects of grouping were observed in this case. Statistically, by varying the different parameters in the fuzzy clustering (i.e., Fk and k), although the discrimination power was improved, there were no observable core centroids (average sample representative for each cluster). Hence, the algorithm could only exclude the most dissimilar sample at a time (Figure 8). The fuzzy k-means dendrograms show that the wines examined during Year 1 are more discriminable from one another than the aged wines of Year 2 (Figure 8).

The random associations inherent in this data were demonstrated at the individual data set explorations (Results section 1) and, most importantly, in the data fusion models (Table 1), where Year 2 had higher RV coefficients between the individual data sets and the MFA model. Although not shown here, when partitioning at Fk > optimal (i.e., approximately two members per cluster), there was a random assignment of memberships supporting the previously reported lack of a representative centroid sample. By varying the number of clusters, the analysis looks at how well the samples can be "pulled apart" and still be reliably closely associated. By varying the partition coefficient (Fk), the analysis looks at how wide the bandwidth around each centroid (using random assignment/partitioning) can be while retaining reliable clustering. The bandwidth would be comparable to borders

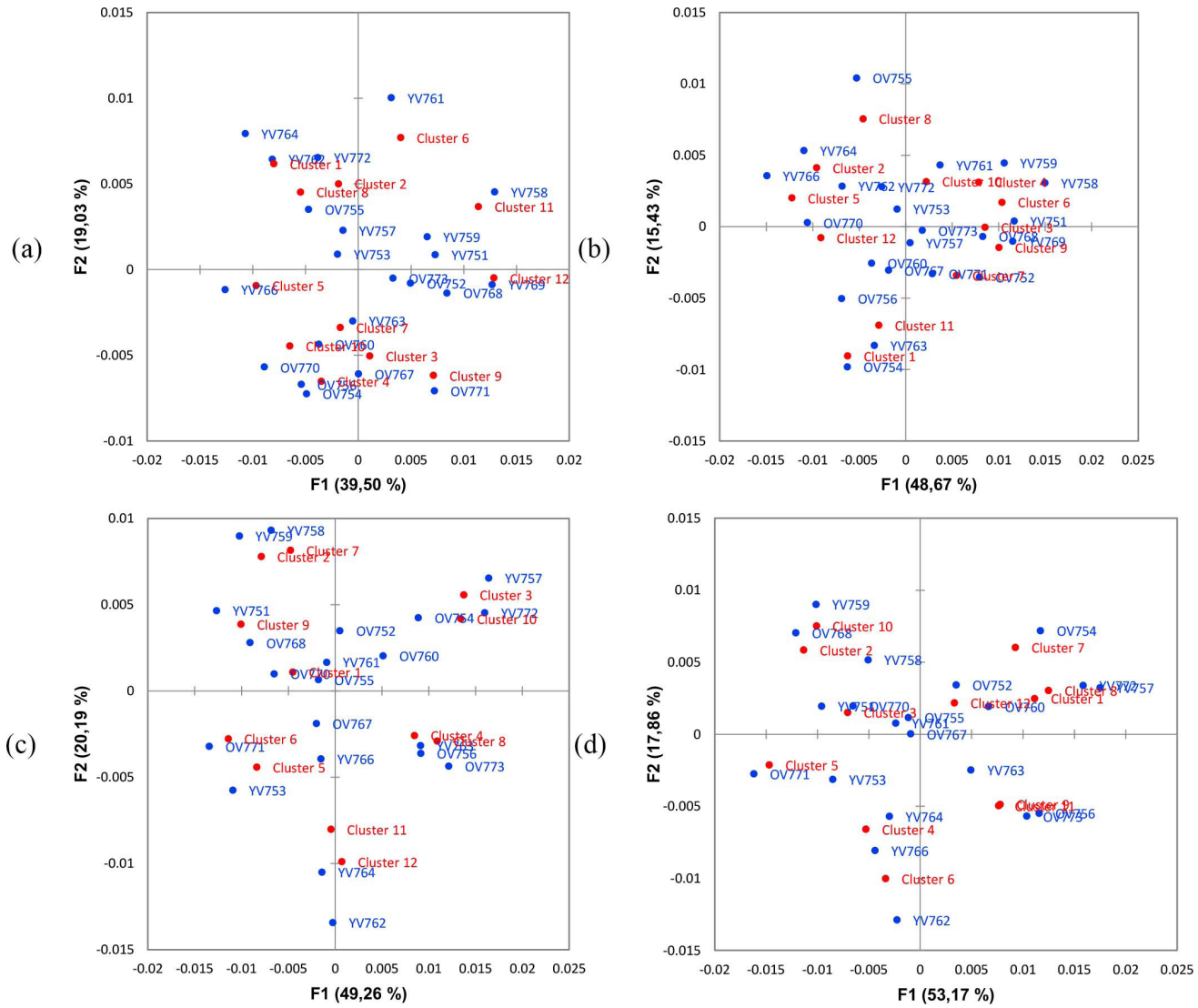


FIGURE 7. Correspondence analysis (CA) biplots on the membership probabilities at optimal clustering conditions for Year 1 (a) with Fk set at 1.001 and (b) 1.05, and Year 2 (c) with Fk set at 1.001 and (d) 1.05.

in typicality assignment in sensory typicality experiments (Ballester *et al.*, 2013). Although the discrimination between samples and hence the clustering was improved by using the fuzzy algorithms, it carried little contextual applicability in this case. This strategy could, however, be applied to cases such as the different styles of South African Chenin blanc to improve on the previous use of parametric statistical techniques (Buica *et al.*, 2017; Lawrence, 2012; Valente *et al.*, 2018).

CONCLUSION

This study aimed to use an unsupervised strategy that consisted of data fusion coupled with an exploration of pattern recognition (comparing parametric and non-parametric clustering). The study used MFA to fuse four data sets, namely, HRMS, NMR, verbal, and non-verbal sorting. The NMR fingerprint produced unique sample configurations indicated by low RV coefficients vs other data sets. The nature of the discrimination (noise or legitimate uniqueness) in the NMR was investigated by using alternative scaling by MFA or by blocking (i.e., NMR regions). Identifying the cause of the discriminant results of the NMR was not the objective of this study, thus, this was not explored further. As a result, all four data sets were fused without any pre-processing or alternative scaling methods. To try to elucidate any further patterns, cluster analysis was applied to the MFA sample configurations using AHC and fuzzy k-means clustering. Through a series of exploratory steps, which included outlier sample exclusion, varying the coefficient of fuzziness (Fk) and the number of clusters (k), the optimal clustering conditions were found. At the optimal clustering conditions (Fk = 1.001 and k = 12), fuzzy k-means clustering was more reliable than AHC, indicated by a higher cophenetic correlation coefficient and lower within-class variation. This meant that fuzzy k-means was sensitive to small variations between samples and discriminated samples between and within classes differently.

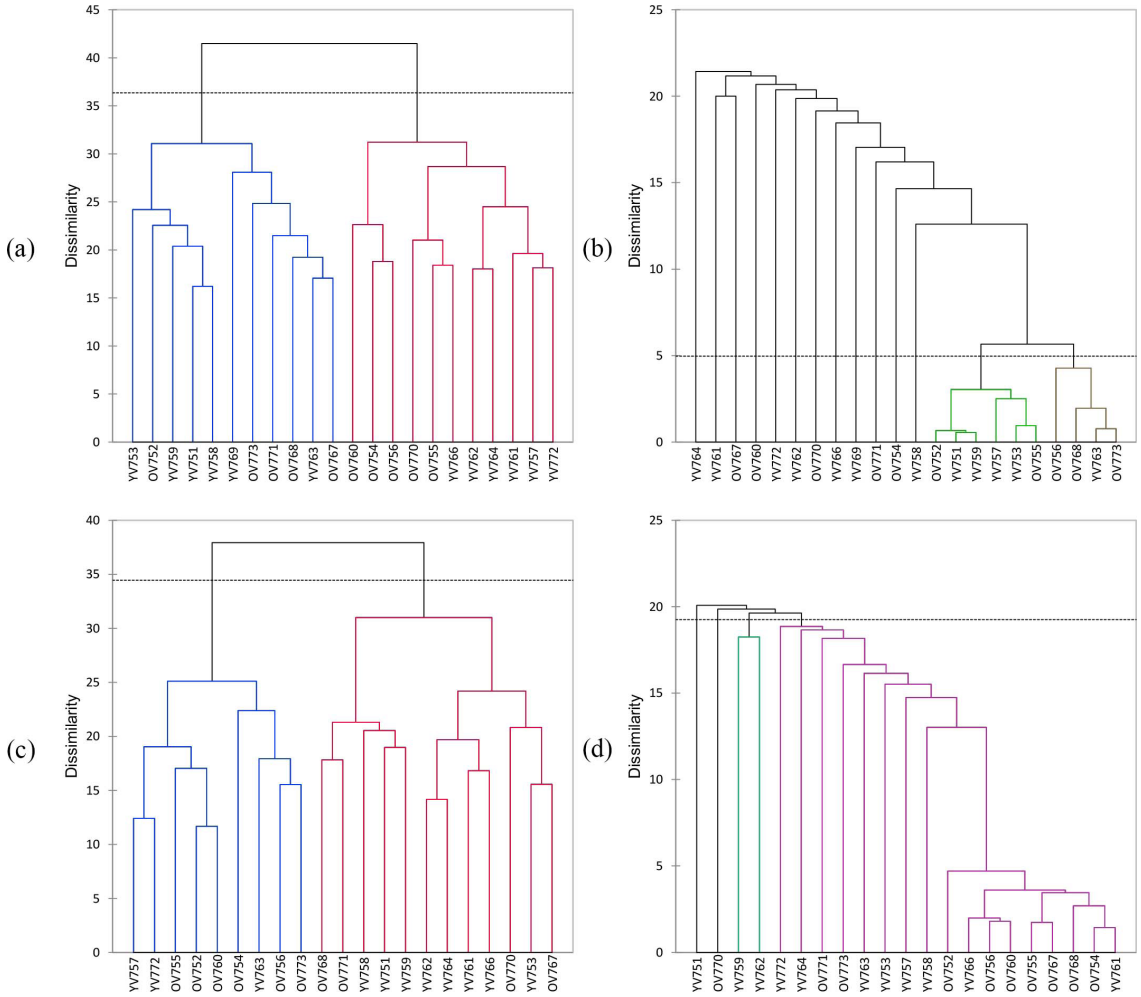


FIGURE 8. AHC dendrogram for Year 1 based on (a) parametric clustering and (b) fuzzy k-means clustering, and Year 2 based on (c) parametric clustering and (d) fuzzy k-means clustering. Optimal clustering conditions were set at 12 clusters and a fuzzy coefficient (Fk) of 1.001.

The data fusion did not elucidate any obvious patterns related to the applied context: are South African old vine Chenin blanc wines chemically and/or sensorially discriminable from young vine wines? The multi-layered approach demonstrated that the old vine Chenin blanc samples in this study were too similar to each other and to the young vine wines to obtain any class or age discrimination. However, the discrimination power of fuzzy k-means can be used for cases where AHC shows some discrimination, but the borders between classes are too fuzzy.

ACKNOWLEDGEMENTS

The authors would like to thank the following: Herman Redelinghuys from the Stellenbosch University Centre for Research on Evaluation, Science, and Technology (CREST) for the discussions on the HRMS data handling; Prof. Maritjie Stander from the Mass Spectrometry Unit of the Central Analytical Facility (Stellenbosch University) for her advice and help with the HRMS; Dr D.J. Brand and Elsa Malherbe from the Nuclear Magnetic Resonance Unit of the Central Analytical Facility (Stellenbosch University) for their advice and help with NMR analysis.

This research was funded by the National Research Foundation (NRF) of South Africa.

REFERENCES

Alañón, M., Pérez-Coello, M., & Marina, M. (2015). Wine science in the metabolomics era. *Trends in Analytical Chemistry*, *74*, 1–20. https://doi.org/10.1016/j.trac.2015.05.006

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

Ballester, J., Mihnea, M., Peyron, D., & Valentin, D. (2013). Exploring minerality of Burgundy Chardonnay wines: A sensory approach with wine experts and trained panellists. *Australian Journal of Grape and Wine Research*, *19*(2), 140–152. https://doi.org/10.1111/ajgw.12024

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. In *Pattern Recognition with Fuzzy Objective Function Algorithms* (First, Vol. 66). Springer US. https://doi.org/10.1007/978-1-4757-0450-1

Biancolillo, A., Boqué, R., Cocchi, M., & Marini, F. (2019). Data Fusion Strategies in Food Analysis. In *Data Handling in Science and Technology* (Vol. 31, pp. 271–310). Elsevier Ltd. https://doi.org/10.1016/B978-0-444-63984-4.00010-7

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta*, *891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Buica, A., Brand, J., Wilson, C., & Stander, M. (2017). Evaluating South African Chenin blanc wine styles using an LC-MS screening method. *Studia Universitatis Babes-Bolyai Chemia*, *62*(2Tom1), 113–123. https://doi.org/10.24193/subbchem.2017.2.09

Burt, C. (1947). MULTIPLE FACTOR ANALYSIS. *British Journal of Educational Psychology*, *17*(3), 163–169. https://doi.org/10.1111/j.2044-8279.1947.tb02224.x

Cariou, V., & Qannari, E. M. (2018). *Statistical treatment of free sorting data by means of correspondence and cluster analyses*. https://doi.org/10.1016/j.foodqual.2018.01.011

Cayuela, J. A., Puertas, B., & Cantos-Villar, E. (2017). Assessing wine sensory attributes using Vis/NIR. *European Food Research and Technology*, *243*, 941–953. https://doi.org/10.1007/s00217-016-2807-9

Cocchi, M. (2019). *Data fusion methodology and applications* (M. Cocchi (ed.); Vol. 31).

De Carvalho Rocha, W. F., Do Prado, C. B., & Blonder, N. (2020). Comparison of chemometric problems in food analysis using non-linear methods. *Molecules*, *25*(13), 3025. https://doi.org/10.3390/molecules25133025

Engel, J., Gerretzen, J., Szyman´ska, E., Szyman´ska, S., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, M. C. (2013). Breaking with trends in pre-processing? *Trends in Analytical Chemistry*, *50*, 96–106. https://doi.org/10.1016/j.trac.2013.04.015

Figueiredo-González, M., Martínez-Carballo, E., Cancho-Grande, B., Santiago, J. L., Martínez, M. C., & Simal-Gándara, J. (2012). Pattern recognition of three Vitis vinifera L. red grapes varieties based on anthocyanin and flavonol profiles, with correlations between their biosynthesis pathways. *Food Chemistry*, *130*(1), 9–19. https://doi.org/10.1016/j.foodchem.2011.06.006

Garrido-Bañuelos, G., Buica, A., De Villiers, A., & Du Toit, W. J. (2019). Impact of Time , Oxygen and Different Anthocyanin to Tannin Ratios on the Precipitate and Extract Composition Using Liquid Chromatography-High Resolution Mass Spectrometry. *South African Journal for Enology and Viticulture*, *40*(1). https://doi.org/https://doi.org/10.21548/40-1-3249

Garrido-Delgado, R., Arce, L., Guamán, A. V., Pardo, A., Marco, S., & Valcárcel, M. (2011). Direct coupling of a gas–liquid separator to an ion mobility spectrometer for the classification of different white wines using chemometrics tools. *Talanta*, *84*(2), 471–479. https://doi.org/10.1016/J.TALANTA.2011.01.044

Ghasemi, J. B., Heidari, Z., & Jabbari, A. (2013). Toward a continuous wavelet transform-based search method for feature selection for classification of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, *127*, 185–194. https://doi.org/10.1016/j.chemolab.2013.06.008

Godelmann, R., Fang, F., Humpfer, E., Schütz, B., Bansbach, M., Schäfer, H., & Spraul, M. (2013). Targeted and nontargeted wine analysis by1H 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*(23), 5610–5619. https://doi.org/10.1021/jf400800d

Granato, D., de Araújo Calado, V. M., & Jarvis, B. (2014). Observations on the use of statistical methods in Food Science and Technology. *Food Research International*, *55*, 137–149. https://doi.org/10.1016/J.FOODRES.2013.10.024

Härdle, W. K., & Simar, L. (2015). Applied multivariate statistical analysis, fourth edition. In *Applied Multivariate Statistical Analysis, Fourth Edition*. https://doi.org/10.1007/978-3-662-45171-7

Khang, T. D., Vuong, N. D., Tran, M. K., & Fowler, M. (2020). Fuzzy C-means clustering algorithm with multiple fuzzification coefficients. *Algorithms*, *13*(13). https://doi.org/10.3390/A13070158

Lawrence, N. (2012). Volatile metabolic profiling of SA Chenin blanc fresh and fruity and rich and ripe wine styles : Development of analytical methods for flavour compounds (aroma and flavour) and application of chemometrics for resolution of complex analytical measurement. *MSc Thesis. Institute of Wine Biotechnology*, *March*.

López-Rituerto, E., Savorani, F., Avenoza, A., Busto, J. H., Peregrina, J. M., & Engelsen, S. B. (2012). Investigations of la Rioja terroir for wine production using 1H NMR metabolomics. *Journal of Agricultural and Food Chemistry*, *60*(13), 3452–3461. https://doi.org/10.1021/jf204361d

Mafata, M., Brand, J., Panzeri, V., & Buica, A. (2020). Investigating the Concept of South African Old Vine Chenin blanc. *South African Journal of Enology & Viticulture*, *41*(2), 168–182. https://doi.org/10.21548/41-2-4018

Mascellani, A., Hoca, G., Babisz, M., Krska, P., Kloucek, P., & Havlik, J. (2021). H NMR chemometric models for classification of Czech wine type and variety. *FOOD CHEMISTRY*, *339*. https://doi.org/10.1016/j.foodchem.2020.127852

McKillup, S. (2005a). Statistics explained: An introductory guide for life scientists. In *Statistics Explained: An Introductory Guide for Life Scientists*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815935

McKillup, S. (2005b). Statistics Explained. In *Statistics Explained: An Introductory Guide for Life Scientists*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815935

Myhre, J. N., Mikalsen, K. Ø., Løkse, S., & Jenssen, R. (2018). Robust clustering using a kNN mode seeking ensemble R. *Pattern Recognition*, *76*, 491–505. https://doi.org/10.1016/j.patcog.2017.11.023

Old Vine project. (2020). *Old Vine Project Certification process – Old Vines Project*. Retrieved March 17, 2020, from http://oldvineproject.co.za/old-vine-project-certification-process/

Panzeri, V., Ipinge, H. N., & Buica, A. (2020). *Evaluation of South African Chenin blanc Wines Made From Six Different Trellising Systems Using a Chemical and Sensorial Approach*. *41*(2), 133–150.

Pereira, A. C., Reis, M. S., Saraiva, P. M., & Marques, J. C. (2010). Analysis and assessment of Madeira wine ageing over an extended time period through GC–MS and chemometric analysis. *Analytica Chimica Acta*, *660*(1–2), 8–21. https://doi.org/10.1016/J.ACA.2009.11.009

Radovanovic, A., Jovancicevic, B., Arsic, B., Radovanovic, B., & Bukarica, L. G. (2016). Application of non-supervised pattern recognition techniques to classify Cabernet Sauvignon wines from the Balkan region based on individual phenolic compounds. *Journal of Food Composition and Analysis*, *49*, 42–48. https://doi.org/10.1016/j.jfca.2016.04.001

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/J.TRAC.2009.07.007

Salkind, J., & Kristin, R. (2007). *Encyclopidia of Measurement and Statistics* (N. J. Salkind (ed.)). Sage.

Seisonen, S., Vene, K., & Koppel, K. (2016). The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. *Food Chemistry*, *210*, 530–540. https://doi.org/10.1016/j.foodchem.2016.04.134

Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Li Vigni, M., Marchetti, A., & Cocchi, M. (2014). A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemometrics and Intelligent Laboratory Systems*, *137*, 181–189. https://doi.org/10.1016/j.chemolab.2014.06.012

Stevenson, T. (2005). The-New-Sothebys-Wine-Encyclopedia. In *The Sotheby's wine Encyclopedia* (Fourth). Dorling Kindersley Limited.

Valente, C. C., Bauer, F. F., Venter, F., Watson, B., & Nieuwoudt, H. H. (2018). Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-23347-w

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: a review of new descriptive methods in food science. *International Journal of Food Science & Technology*, *47*(8), 1563–1578. https://doi.org/10.1111/j.1365-2621.2012.03022.x

Vannier, A., Brun, O. X., & Feinberg, M. H. (1999). Application of sensory analysis to champagne wine characterisation and discrimination. *Food Quality and Preference*, *10*, 101–107. https://doi.org/10.1016/S0950-3293(98)00047-0

Versari, A., Laurie, V. F., Ricci, A., Laghi, L., & Parpinello, G. P. (2014). Progress in authentication, typification and traceability of grapes and wines by chemometric approaches. *Food Research International*, *60*, 2–18. https://doi.org/10.1016/j.foodres.2014.02.007.