beverages

MDPI

Article

Exploration of Data Fusion Strategies Using Principal Component Analysis and Multiple Factor Analysis

Mpho Mafata [1,2,*], Jeanne Brand [2], Martin Kidd [3], Andrei Medvedovici [4] and Astrid Buica [1,2,*]

[1] School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch 7600, South Africa
[2] South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Stellenbosch University, Stellenbosch 7600, South Africa
[3] Center for Statistical Consultation, Stellenbosch University, Stellenbosch 7600, South Africa
[4] Faculty of Chemistry, University of Bucharest, 030018 Bucharest, Romania
* Correspondence: mafata@sun.ac.za (M.M.); astrid.buica@gmail.com (A.B.)

Abstract: In oenology, statistical analyses are used for descriptive purposes, mostly with separate sensory and chemistry data sets. Cases that combine them are mostly supervised, usually seeking to optimize discrimination, classification, or prediction power. Unsupervised methods are used as preliminary steps to achieving success in supervised models. However, there is potential for unsupervised methods to combine different data sets into comprehensive, information-rich models. This study detailed stepwise strategies for creating data fusion models using unsupervised techniques at different levels. Principal component analysis (PCA) and multiple factor analysis (MFA) were used to combine five data blocks (four chemistry and one sensory). The model efficiency and configurational similarity were evaluated using eigenvalues and regression vector (RV) coefficients, respectively. The MFA models were less efficient than PCA, having gradual distributions of eigenvalues across model dimensions. The MFA models were more representative than PCA, as indicated by high RV coefficients between MFA and each individual block. Therefore, MFA approaches were better suited for multi-modal data than PCA. This work approached data fusion systematically and showed the type of decisions that must be made and how to evaluate their consequences. Proper integration of data sets, instead of concatenation, is an important aspect to consider in multi-modal data fusion.

Keywords: wine storage; Chenin blanc; Sauvignon blanc; data concatenation; data fusion; multiblock; multi-modal; multivariate analysis

Citation: Mafata, M.; Brand, J.; Kidd, M.; Medvedovici, A.; Buica, A. Exploration of Data Fusion Strategies Using Principal Component Analysis and Multiple Factor Analysis. *Beverages* **2022**, *8*, 66. https://doi.org/10.3390/beverages8040066

Academic Editors: Alberto Mannu and Giacomo L. Petretto

Received: 29 August 2022
Accepted: 18 October 2022
Published: 21 October 2022

1. Introduction

The fields of metabolomics, engineering, and chemistry have a long history of working on data-orientated approaches for combining data sets from different sources, termed chemometric data fusion [1–3]. Chemometric data fusion has a long history in other fields, but its use in agricultural sciences is quite recent, even more so for the oenology field.

In order to compile a comprehensive account of the response of a wine to a certain phenomenon or influence, data from different sources are gathered; for example, wine can be profiled chemically and sensorially. Owing to the complexity of sensory data matrices, the two evaluation approaches (chemistry and sensory) are generally discussed separately from one another and similarities are inferred. This has been the case for wine authenticity studies in which several measurements are taken and discussed separately [4]. Although this works well for contained cases that have an application-based approach, cases that require the collection of multiple responses across different stimuli or time require a data-orientated approach. Combining data sets from different sources creates a comprehensive profile of the behaviour of a product (in this case, wine) in response to stimuli [5,6]. This is in alignment with the motivation for the fourth industrial revolution, which requires not just gathering large amounts of different data, but looking at the data in smarter ways.

Putting together chemistry and sensory data has its own set of challenges. Data outputs for analytical chemistry instruments have progressed with the development of standardized matrix arrangements for two- to four-dimensional data (e.g., hyphenated techniques, such as LC-MS/MS, used in wine metabolomics, which use multiple detectors with three-dimensional data outputs) [7]. This required the consolidation of statistical treatments (normalization of peak intensities) and alignment across different detectors. On the contrary, the complex and very often qualitative nature of sensory data is usually communicated through descriptive narratives. Although there are standardized statistical treatments for certain methods [8,9], there is still progress to be made to reach a consensus on standardized matrix arrangements and outputs that encourage data consolidation. Owing to the qualitative nature of many sensory evaluations, the assumptions made through the statistical treatment of data are being continuously debated and tend to be misconstrued as over-reaching or over-fitting [8].

Data fusion is defined not simply as putting together data sources, but rather as "integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source" isif.org (https://isif.org/, accessed on 29 August 2022). Data fusion is classified as low-level, mid-level, and high-level based on the increasing complexity of the models and depending on the number of steps between the capture of the raw data and the final fused model [10]. Low-level data fusion is the simplest form, which usually uses raw data with little pre-modelling processing. Issues and challenges related to pre-modelling processing have previously been described for sensory [8,11] and chemical analyses [12–14] in oenological applications. Low-level data fusion requires data sets to have compatible matrices, with a compatible matrix order (2D, 3D, etc.) and at least one of the dimensional array (observations or variables) being the same [10]. Low-level data fusion models are often used as a pre-modelling step in mid- and high-level data fusion. From the low-level model, the pre-modelling processing techniques used include the selection of variables (e.g., standardized deviates/coordinates) or features (e.g., model dimensions), and a new matrix is then modelled [15,16]. Mid-level data fusion is a systematic approach comprised of intermediate steps between the raw data and the final model. This may be due to differences in the matrix dimensions and directed goals requiring feature selections and/or pre-modelling processing. High-level, also called decision-level, data fusion, is the most complex and involves several directed steps. High-level data fusion strategies use both classical statistical analysis and machine learning techniques [17]. High-level supervised models have been used in oenology for the prediction and calibration of oenological processes, such as wine ageing [18]. Recently, machine learning techniques, such as text-mining for qualitative sensory data [19] and fuzzy logic [6,20], have been used for information mining in food applications.

In each of the three levels of data fusion, unsupervised modelling strategies in which the objective is data exploration may be used. These unsupervised modelling strategies look for patterns of grouping or similarity, or for the best-fit model. The objective of data fusion may have a specific target in mind, in which case supervised modelling strategies are used. In the field of oenology, most reported cases of data fusion are supervised, with unsupervised methods being used as preliminary explorative steps that work to refine the final model [15,17]. Supervised data fusion approaches are goal-orientated, targeting and selecting only certain features from data blocks related to the phenomenon under investigation, thereby reducing dimensionality and increasing the predictive, discriminant, or classification power [3]. In trying to refine these supervised models, the data that do not contribute to increasing the regression coefficients are discarded. Conversely, unsupervised data fusion approaches retain most of the information captured while reducing the dimensionality.

The most commonly used unsupervised data fusion methods in oenology are principal component analysis (PCA) and multiple factor analysis (MFA) [15,21]. PCA is one of the most popular multivariate statistical tools in applied science [22] which can be used for low or mid-level data fusion (by matrix concatenation) or as a pre-processing model.

The focus of PCA is efficiency, accomplished by reducing the dimensions of a data set into more manageable dimensions called principal components, which make it easier to visualize and interpret complex data [23]. These principal components standardize the raw data to capture the essence of the correlations or covariance between the variable and the observations (the common vectors). It is because of these functions that PCA is an appropriate low-level data fusion model in applied food science [15]. The disadvantages of PCA are its inability to handle data with high counts of zero or 'missing data', which can be an issue for certain sensory data and chemical instrument outputs [23,24]. In such cases, the raw data are revisited and pre-processed manually or through statistical exclusions of some data before being modelled again. This rigorous approach can result in overfitting/overcorrection that disregards the unsupervised intent of PCA modelling [15].

MFA is another popular multivariate tool in applied food science that goes beyond the simple matrix concatenation approach of PCA [21,25]. MFA is a generalized PCA with a multiblock data fusion approach that retains and standardizes each block before fusion, retaining the weight and contributions of the variables in each block to avoid any skewing by one data block [26]. MFA is part of the unsupervised multiblock category of data fusion methods, which include techniques such as common dimension analysis (CommDimm), and variations of parallel factor analyses (PARAfac). MFA has previously been used as a multi-modal data fusion tool in genomics [27], biotechnology [28], and sensory analyses [29,30]. In sensory analyses, multiblock techniques are used for combining data from ordinal data, such as Napping, that cannot be simply concatenated. Other familiar multiblock analyses commonly used for sensory data include STATIS, Distatis, and others [31]. Combinations of data sets (qualitative or quantitative, continuous or discreet, and categorical or integer) can thus be incorporated by pre-processing steps, such as scaling, and weighing can be conducted before applying multiblock modelling [29]. Rapid sensory methods that capture categorical data use CA and MCA as common multivariate methods [8].

The problems with conducting data fusion are three-fold:

- The choice of model can be difficult due to the large number of available techniques and their variants;
- The execution of some models is difficult due to the availability of software and may often require advanced programming skills. In addition to this, a lack of transparency when it comes to the different stages of data handling creates reproducibility issues among the science community;
- Evaluating the performance of unsupervised data models is often descriptive of the data, but does not include descriptions of the model.

The performance of unsupervised data fusion models is evaluated comparatively and descriptively by looking at the distribution of the explained variance over different dimensions and grouping of samples when using cluster analysis or confidence ellipses [21,29,32]. Recently, in order to compare the similarities between the sample configurations of different models, regression vector coefficients have been used [30,33–36].

This study investigated the differences between concatenation (PCA) and multiblock (MFA) data fusion strategies. These methods were chosen for their long-standing historical use across many fields, their ease of access across different data analysis packages, and their ease of execution, which increases reproducibility for interested scientists. These methods were additionally chosen for their ability to incorporate both common and unique information in an unsupervised approach, which allows for further inferences, especially when deciding to create predictive models for the sensory data captured. This study was focused on elucidating the phases of data analysis and detailed the rationale behind the different steps of data fusion, from data set curation to the evaluation of the final fused models. The data used in this study were based on the response of white wine to different storage conditions [34]. Twelve sample sets, with seven samples per data set, a total of 84 individual samples, were used. The data were captured and grouped into five blocks: antioxidant-related parameters (ARP), volatile compounds composition (VCC), UV-Vis

spectrum (UV-Vis), infra-red spectrum (IR), and sensory. The purpose of building these models was to create efficient, comprehensive, and representative data fusion models. The performance of the models was evaluated by examining the distribution of the percentage explained variance (%EV) and the slope of the exponential decay of the eigenvalues across the different model dimensions as measures of information distribution. Comparisons between model sample configurations were made using pair-wise regression vector (RV) coefficients. Issues surrounding model efficiency and redundancies between data blocks, and the representativeness of the data fusion model will be discussed.

2. Materials and Methods

2.1. Experimental Design

The materials and methods related to winemaking, wine treatments, sensory evaluation, and chemical analysis (oenological parameters, thiols, glutathione, and major volatiles) were previously published by Mafata et al. [34]. In brief, the experiment focused on the stability of wines at various temperatures and for different time periods. The samples belonged to two cultivars (Chenin Blanc and Sauvignon Blanc) from six wineries each (twelve sample sets in total). The sample sets can be identified by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB). Each sample set consisted of seven wines corresponding to the experimental storage conditions (i.e., no storage time/control, three and nine months of storage; and three temperatures: 15 °C, 25 °C, and uncontrolled ambient temperature).

2.2. Sensory Data Methodology

The descriptive part of the sensory data methodology (panel parameters and instructions) was previously published in Mafata et al. [34]. For the purpose of the current study, some relevant aspects are described here. The sensory method chosen for this experiment was Pivot© Profile (PP) [37]. PP is a verbal, reference-based method that collects information about the attributes, per sample, relative to the pivot [8]; in this case, the control sample. The data were captured as a rating of either +1 (more than the pivot) or −1 (less than the pivot), and for attributes that were not mentioned, a rating of zero was given. The raw data were captured per data set, with judges and repeats kept separate and not further concatenated [38].

Linguistic and semantic reductions of terms were performed manually, resulting in a total of 200 attributes. Statistical consolidation was then performed separately for each sample set. Each attribute was summed across the judges and repeats, translated into positive ratings, and zero-sum terms were excluded. Positive translation was conducted to convert the data from rating to frequency so that the modelling could be conducted by CA [37]. Terms with less than 5% citations were removed, resulting in 29 to 36 attributes per sample set.

2.3. Chemical Data Collection and Capturing

The chemical data were categorized into volatiles (VCC data set: thiols and major volatiles) and antioxidant-related parameters (ARP data set: colour intensity (CI, A_{520} + A_{420}), colour hue (CH, A_{520}/A_{420}), total phenolics (A_{280}), hydroxycinnamic acids (A_{320}) and browning (A_{420}), CIElab parameters, glutathione, and total and free sulphur dioxide) were discussed previously [34]. Ultraviolet-visible light spectrophotometric scans (UV-Vis data set) were run from 280 nm to 780 nm (in 1 nm increments) in triplicate on a Thermo Scientific Multiskan GO (Waltham, MA, USA) 1510-02586 microplate spectrophotometer. Infra-red spectra measurements (IR data set), in the mid-infrared range (4000–600 cm^{-1}), were collected using an Alpha-P ATR FT-MIR spectrometer (Bruker Optics, Ettlingen, Germany). Each sample was scanned at a resolution of 4 cm^{-1} and at a scanning velocity of 7.5 kHz, and then averaged over 64 scans to give a final reading. Instrumental control and data capture were carried out using OPUS software (OPUS v. 7.0 for Microsoft, Bruker Optics, Ettlingen, Germany).

2.4. Statistical Analysis

Multivariate analysis was performed separately for each winery and each cultivar, and each sample set consisted of seven wines (2.1). The data were divided into five blocks based on the properties and modality of acquisition: volatile compounds (VCC), antioxidant-related parameters (ARP), UV-Vis spectra (UV-Vis), infra-red spectra (IR), and sensory data (Table 1). In other words, each block consisted of twelve data sets, and each data set contributed to five blocks. The raw sensory data were subjected to correspondence analysis (CA) and the standardized deviates/residuals matrix was used for data fusion. All PCA analyses in this study were based on the generalized Pearson correlation coefficient with standard univariate scaling applied to all measurements before modelling. MFA was performed on the correlation matrices of the chemistry data sets (observations vs. variables) and the latent variables of the sensory data. The data blocks were first standardized by PCA and then MFA was performed [26]. For each model, an exponential decay curve was plotted using eigenvalues for each dimension and the slope was calculated using Microsoft Excel (Excel Office 365, version 2002, Microsoft Corp., United States). Configurational similarities for all score plots were calculated using pair-wise regression vector (RV) coefficients [33] and inferred topology (iTOP) RV between the PCA and MFA data fusion models [39] (Figure 1). Statistical calculations and modelling were performed using R and visualised using the packages "FactoMineR" for MFA and "UBbipl" for PCA https://www.R-project.org/, accessed on 29 August 2022 (R Foundation for Statistical Computing, Vienna, Austria). Additional modelling was also conducted using Statistica™ 13 (TIBCO Software Inc., Palo Alto, CA, USA).

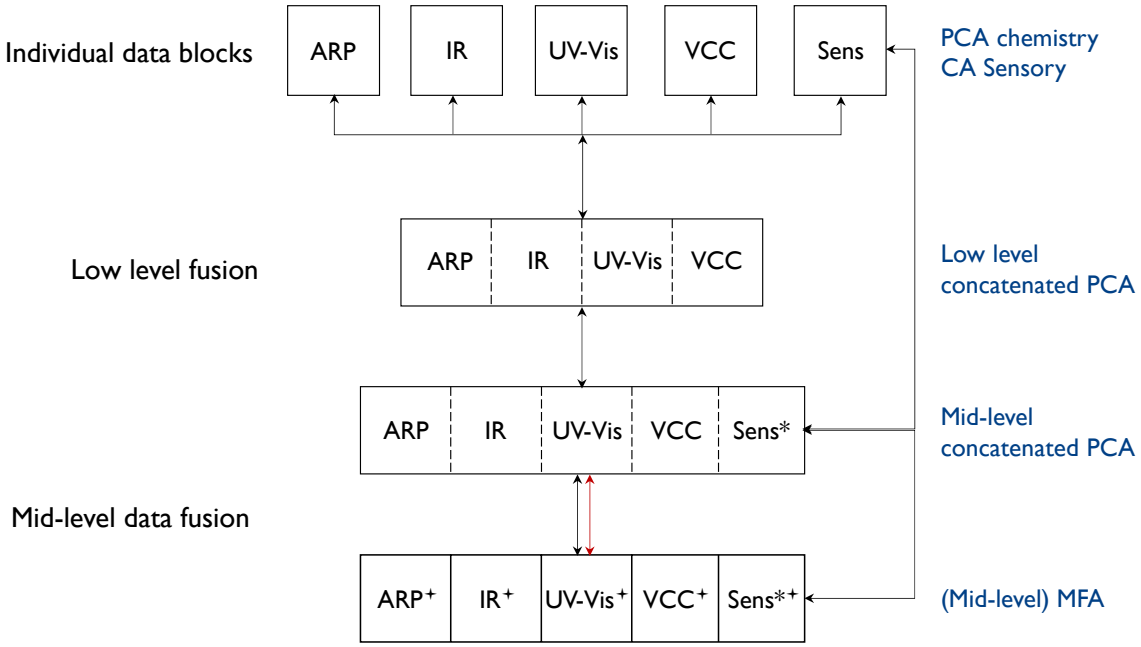

Figure 1. Structure of the data fusion strategy and model comparisons. Black arrows indicate the pairs used for RV coefficient calculation and the red arrow indicates iTOP RV. *: the data block was pre-processed, ✦: the data block was weighed. Abbreviations and further details are provided in the text.

Table 1. Summary of the five blocks, low-level, and mid-level data fusion approaches using multivariate analysis.

Level	Blocks	Input					Pre-Processing	Modelling		Model Output			
		Description	Value Type	Matrix Type	Row	Column		Modelled Matrix	Model Type	Output Matrix Type	Output Matrix Row	Output Matrix Column	Model Performance Parameters
Individual Data Blocks	ARP	Concentrations, absorbance values	Discreet	Correlation	Samples	Concentrations, AU	None	Raw data	PCA	Scores	Samples	Principal components	EV%, eigenvalue, decay slope R2
	IR	Spectral, reflectance	Continuous	Continuous	Samples	Wavenumber	None *	Raw data	PCA	Scores	Samples	Principal components	
	UV-Vis	Spectral, absorbance	Continuous	Continuous	Samples	Absorbance wavelengths	None	Raw data	PCA	Scores	Samples	Principal components	
	VCC	Concentrations	Discreet	Correlation	Samples	Concentrations	None	Raw data	PCA	Scores	Samples	Principal components	
	Sensory	Pivot profile reference-based method	Discreet	Rating	Samples	Ratings (−1, 0, 1)	Conversion to frequency matrix	Positive FC	CA	Scores	Samples	Factors	
										Standardized deviates	Samples	Variables	
Low	ARP + IR + UV-Vis + VCC	Block concatenation	Mixed	Mixed	Samples	See individual data blocks	Matrix concatenation	Concatenated matrix	PCA	Scores	Samples	Principal components	EV%, eigenvalue, decay slope R2
Mid	ARP + IR + UV-Vis + VCC + Sensory ‡	Block concatenation	Mixed	Mixed	Samples	See individual data blocks except ‡	Matrix concatenation	Concatenated matrix	PCA	Scores	Samples	Principal components	EV%, eigenvalue, decay slope R2
	ARP + IR + UV-Vis + VCC + Sensory ‡	Blocks	Mixed	Multiblock	Samples	See individual data blocks except ‡	PCA per block on raw data except ‡	Multiblock standardized deviates from individual PCA	MFA	Scores	Samples	MFA dimensions	EV%, eigenvalue, decay slope R2
										Loadings	Blocks	MFA dimensions	

* See text for discussion and details. Abbreviations: PCA—principal component analysis, CA—correspondence analysis, MFA—multiple factor analysis, and FC—frequency of citation.
‡ Sensory: CA standardized deviates.

3. Results

The fusion of the five data blocks in this study (VCC, ARP, IR, UV-Vis, and sensory) was unsupervised and explorative, from low-level to mid-level data fusion strategies with increasing complexity (Table 1). This section is arranged according to both the complexity of the conceptualisation of the approach, as well as the operational order taken in fusing the data blocks.

3.1. Curation of Data Blocks

3.1.1. Assessment of Pre-Modelling Processing

It is important to first inspect the data blocks specifically for the purposes of data fusion, as this will dictate which type and which level of fusion are needed; the decisions taken might be different to the ones when data fusion is not the purpose [40]. When looking at pre-modelling processing methods in view of data fusion, two criteria were considered in this study, namely matrix compatibility and signal correction.

Matrix compatibility is an important eligibility criterion for low-level data fusion strategies [41]. If matrices are incompatible, then pre-modelling processing must be performed. The chemistry data sets (ARP, VCC, UV-Vis, and IR) were captured as compatible correlation matrices (Table 1) and, thus, could be combined using either low-level or higher-level data fusion strategies. In contrast, in order to obtain a compatible matrix for the sensory data, the standardized deviates matrix was obtained from the CA model (Table 1). The raw sensory data were captured as rating data, the matrix of which consisted of 0, 1, and -1 ratings. These types of data sets cannot be modelled using PCA as they contain large counts of zero measurements [22].

With regards to signal correction, spectral pre-processing is often considered for UV-Vis and IR spectral data blocks and included as toolkits in most software [42,43]. In the case of the UV-Vis data block, high model efficiency (%EV) was taken as a good indicator for proceeding with the raw UV-Vis data for fusion without the need for pre-processing.

As IR had lower %EV and pair-wise RV coefficients (i.e., between the scores of the PCA models with and without pre-processing), pre-modelling processing was considered in order to improve the performance parameters. Infra-red spectral data are prone to spectral irregularities, which are categorised under two phenomena, namely scattering and baseline irregularities [12]. The mathematical conversions performed to correct these phenomena fell under these two categories. Infra-red data regularly use multiplicative scatter correction (MSC) for scattering, first derivative transformations for baseline corrections, and combinations of the two [12]. In this section, the raw data, MSC, first derivative, and combinations of MSC with the first derivative were investigated as potential methods of pre-processing infra-red data.

The impact of the transformations on the efficiency of the PCA models was evaluated by %EV (Table 2) and any effect on the sample set configuration was evaluated through pairwise RV coefficients between the PCA models of the raw and the transformed data (Supplementary Table S1). The raw data produced PCA models with the highest efficiency, with average cumulative %EV for the first two principal components of 84 ± 9 for CB and 70 ± 6 for SB. All other pre-processing transformations lowered the efficiency of the models, with some exceptions; MSC increased the efficiency of the PCA models of the PDB and KZC CB sample sets by 7% and 4%, respectively. The KZC CB sample set model efficiency was increased by the pre-processing methods, except for the first derivative transformation. KZC had the second highest %EV of all the wineries; the 4% increase was thus relatively negligible, and inspection of the spectra showed no obvious faults.

Table 2. Cumulative percentage explained variance for the first two principal components of the infrared raw data and its mathematical transformations.

	Data Set	Raw	1st Deriv	MSC	1st Deriv + MSC	MSC + 1st Deriv
Chenin Blanc	AVN	82	52	73	51	53
	CDB	72	57	61	52	52
	DTK	97	62	97	72	73
	FRV	76	52	68	50	53
	KZC	96	79	100	100	100
	PDB	81	54	88	50	60
	Average	84	59	81	63	65
	Stdev	9	9	15	18	17
Sauvignon Blanc	AVN	72	43	55	40	39
	CDB	74	54	63	51	51
	DTK	63	45	46	39	40
	FRV	74	50	51	40	41
	KZC	62	43	51	38	39
	PDB	77	45	52	38	39
	Average	70	47	53	41	42
	Stdev	6	4	5	5	4
Overall	Low	62	43	46	38	39
	High	97	79	100	100	100
	Average	77	53	67	52	53
	Stdev	10	10	18	17	17

MSC—multiplicative scatter correction, 1st deriv—first derivative. The data sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB).

The RV coefficients showed high configurational similarities between the different pre-processed models and the raw data, except for the KZC CB sample set (Table S1), meaning that, generally, the pre-processing had very little effect on the sample configuration. The raw data set had the lowest RV coefficients, ranging from 0.73 to 0.95 for CB (0.84 $\pm$ 0.06, ave $\pm$ SD) and 0.70 to 0.90 for SB (0.78 $\pm$ 0.06). This means that the configurations of the transformed spectra were more similar to each other than to the raw data. However, there was a negligible difference in the configurations, with a maximum 15% increase in the RV coefficients on average.

For the KZC CB sample set, the RV coefficients between the MSC and the raw data (0.37) and 1st deriv (0.44) were the lowest. Overall, the MSC transformation resulted in increased model efficiency (%EV) and relatively unique sample configurations (low RV coefficients) in the KZC CB sample set. If the purposes of the data fusion in this study were to gather information that would increase the discrimination power between the sample sets, the MSC pre-processing would be suitable. As this study was explorative and unsupervised, such measures were not considered necessary and the decision was made to continue with the raw data for data fusion.

3.1.2. Performance of Individual Block Models

The chemistry data blocks each had a set number of variables (UV-Vis, 501 wavelengths; ARP, 14 parameters; VCC, 34 compounds; and IR, 879 wavenumbers); the sensory data had a varying number of variables as the number of attributes was different for each data set after pre-processing. A comparative exploration of the models' packing efficiency was conducted using the %EV (Supplementary Table S2) and the configurational similarity of the scores (seven samples per set) was calculated through pairwise RV coefficients between the data sets (Supplementary Table S3). Overall, the UV-Vis models were the most efficient, with cumulative %EV ranging from 78 to 99, and an average of 91 $\pm$ 7 for the first two PCs. ARP was the second most efficient (75 to 94%EV, 84 $\pm$ 5), followed by IR (64 to 98%EV, 78 $\pm$ 10) and VCC (72 to 83%EV, 77 $\pm$ 3). Sensory had the lowest cumulative %EV (55 to 78%EV, 68 $\pm$ 6) for the first two dimensions of the CA, which is an inherent characteristic of holistic techniques, such as sensory analysis [8].

The sample configurations of UV-Vis and ARP were the most similar, with RV coefficients ranging from 0.78 to 0.93 for CB and 0.73 to 0.93 for SB. This is understandable as compounds with antioxidant properties can absorb UV-Vis energy [44]. Additionally, the CIE lab and other colour indices listed in the ARP data block were calculated from specific measurements in the UV-Vis spectrum. The RV coefficients for ARP vs. VCC were the second highest, ranging from 0.55 to 0.83 for CB and 0.45 to 0.82 for SB. The RV coefficients for UV-Vis vs. VCC were lower compared with those for ARP vs. VCC, ranging from 0.31 to 0.81 for CB and 0.33 to 0.62 for SB. The RV coefficients were very low between IR and the other chemistry data blocks (UV-Vis, ARP, and VCC), ranging from 0.10 to 0.71 for CB and 0.21 to 0.79 for SB. The RV coefficients between IR and sensory were higher, ranging from 0.38 to 0.86 for CB and 0.51 to 0.72 for SB. The RV coefficients between sensory and UV-Vis were poor, ranging from 0.59 to 0.74 for CB and 0.43 to 0.79 for SB. The RV coefficients were higher between sensory and VCC, ranging from 0.60 to 0.87 for CB and 0.60 to 0.85 for SB. As the sensory method evaluated only the aroma of the wines, it is understandable that it resulted in higher configurational similarity with the VCC data set.

3.2. Low-Level Data Fusion

Low-level fusion involves the simple concatenation of raw data with compatible matrix dimensions [10,45]. The ARP, VCC, UV-Vis, and IR data blocks had compatible observations vs. variable correlation matrices, and thus could be fused using low-level strategies. In order to integrate the sensory and chemistry data, a mid-level data fusion strategy had to be employed; this is explored in the next section. The four chemistry data blocks were first concatenated into one correlation matrix of seven observations (for each sample set) vs. 1428 variables (corresponding to the sum of variables for the chemistry data blocks) and modelled by PCA.

It has previously been shown that the individual models for the four chemistry data blocks were highly efficient, with most of the explained variance captured within the first two principal components (Section 3.1.2, Supplementary Table S2). Comparatively, the low-level PCA fusion model was less efficient (Table 3); hence, a more in-depth exploration of the data distribution was needed to assess the model's performance. The overall stress in the model and the slope of the exponential decay in the stress across the principal components (Table 3) were used to evaluate the model's efficiency [22].

The 1428 variables were fitted over six principal components and the stress was fitted onto an exponential curve with R^2 values between 0.81 and 0.99. CB had more efficient models compared with SB, as measured by the slope, which ranged from 0.44 to 0.88 for CB and 0.38 to 0.55 for SB (Table 3). Approximately 80% of the explained variance was achieved within the first three principal components, indicating lower efficiency than the individual models (Supplementary Table S2). This is characteristic of multimodal data fusion due to the increased number of variables and the different types of data sources [3]. The KZC CB data set had the highest performance indicators again, with a slope of 0.87 (R^2 = 0.95) and a cumulative %EV of 89 for the first two principal components (Table 3).

Due to the concatenated (one matrix) nature of the PCA data fusion strategy, it is difficult to attribute the performance of the model to any one of the data blocks. In order to try and address the issue of redundancy between the data blocks in this low-level strategy, the sample configurations resulting from the PCA on the concatenated data were compared with the individual data sets' PCAs using RV coefficients (Supplementary Table S4). Although the KZC CB sample set was previously an exception in the individual PCA models, the low-level PCA data fusion model was not, as it had similar RV coefficient patterns described for the other sample sets.

Table 3. Performance parameters and stress distribution for low-level data fusion.

					Cumulative %EV per PC					
Cultivar	**Data Set**	**Total Stress (Eigenvalue)**	**Slope**	**R^2**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**
Chenin Blanc	AVN	589	0.55	0.989	41	68	84	92	97	100
	CDB	591	0.46	0.970	41	69	82	90	95	100
	DTK	742	0.88	0.966	52	85	93	98	99	100
	FRV	688	0.44	0.926	48	69	81	88	95	100
	KZC	962	0.87	0.947	67	89	95	98	99	100
	PDB	837	0.56	0.910	59	78	86	92	97	100
Sauvignon Blanc	AVN	617	0.47	0.962	43	70	82	90	95	100
	CDB	716	0.54	0.966	50	74	84	92	97	100
	DTK	541	0.38	0.932	38	65	78	86	93	100
	FRV	556	0.55	0.934	39	76	85	92	97	100
	KZC	800	0.41	0.813	56	70	79	88	95	100
	PDB	653	0.46	0.946	46	70	82	89	95	100
Range	Min	541	0.38	0.813	38	65	78	86	93	100
	Max	962	0.88	0.989	67	89	95	98	99	100

Data fusion of ARP, VCC, UV-Vis, and IR chemical data sets submitted to PCA. Abbreviations: VCC—volatile compounds composition, ARP—antioxidant-related parameters, UV-Vis—ultraviolet-visible spectra, IR—infrared spectra, PCA—principal component analysis, and %EV—percentage explained variation. The data sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB).

It may be misconstrued that the concatenated model is likely to be skewed by the most variable dense data block, in this case, the IR (879 variables); as this data block had the highest RV coefficients (IR vs. low-level PCA), the hypothesis seemed to have some support. IR vs. low-level PCA had RV coefficients ranging from 0.88 to 0.96 for the CB data sets and 0.83 to 0.95 for the SB data sets. As previously discussed, the sample configuration of the IR data block was different from the other data blocks (Section 3.1.2, Supplementary Table S3). A look at the RV coefficients between the low-level PCA model and the other data blocks showed that the sample configurations were mainly case-specific and no one-fits-all generalization of the patterns could be applied for the VCC and ARP data sets. Although the UV-Vis data block had the second-highest number of variables (507), it did not always have the second-highest RV coefficient. This meant that the number of variables was not the most influential factor affecting the sample configuration of the fusion model, but rather the amount of information the technique carried. As previously discussed, IR is an information-rich technique and infra-red activity is a more general property of organic molecules than UV-Vis [46].

3.3. Mid-Level Data Fusion

3.3.1. Principal Component Analysis (PCA)

In order to incorporate the sensory results into fused models, the data had to be in a format compatible with the rest of the data blocks [3,23]. To achieve this, the standardized deviates (standardized co-ordinates) from the CA model of sensory data were used. These were added to the chemistry data blocks by concatenation and the new matrix was modelled by (mid-level) PCA (Table 1). The distribution of the stress and performance indicators of the model are listed in Table 4. As expected, the increased dimensionality due to the addition of sensory data resulted in decreased model efficiency compared with both the individual data blocks and the low-level fusion PCA. The CB models were the most efficient, with the slopes of the exponential decay curves ranging from 0.43 to 0.83 ($R^2 > 0.90$) compared with those for SB ranging from 0.37 to 0.53 ($R^2 > 0.80$). The KZC CB mid-level PCA model was the most efficient, with a slope of 0.82 ($R^2 = 0.94$) and a cumulative %EV of 89 for the first two principal components.

Table 4. Performance indicators and stress distribution of the mid-level PCA data fusion models.

						Cumulative %EV per PC					
	Data Set	**Observations**	**Total Stress (Eigenvalue)**	**Slope**	**R^2**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**
Chenin Blanc	AVN	1458	601	0.45	0.97	41	68	81	89	95	100
	CDB	1463	595	0.53	0.99	41	67	83	91	97	100
	DTK	1461	747	0.83	0.96	51	84	92	97	99	100
	FRV	1463	698	0.43	0.92	48	68	80	88	95	100
	KZC	1458	968	0.82	0.94	66	89	95	97	99	100
	PDB	1461	847	0.54	0.90	58	77	85	91	97	100
Sauvignon Blanc	AVN	1459	661	0.45	0.94	45	69	81	89	95	100
	CDB	1457	721	0.53	0.97	50	73	84	92	97	100
	DTK	1464	544	0.37	0.93	37	64	77	86	93	100
	FRV	1463	561	0.53	0.93	38	75	84	92	97	100
	KZC	1464	805	0.40	0.80	55	69	78	87	95	100
	PDB	1458	661	0.45	0.94	45	69	81	89	95	100
Range	Min	1457	544	0.37	0.80	37	64	77	86	93	100
	Max	1464	968	0.83	0.99	66	89	95	97	99	100

The data sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB).

In the data curation section (Section 3.1.2), it was noted that the sensory data model was the least efficient, having the lowest cumulative %EV of all of the data blocks. However, the concatenation of the sensory data with the chemistry data sets did not lower the cumulative %EV compared with the low-level PCA. On average, the cumulative %EV across the model dimensions decreased by 1% from low-level to mid-level PCA. As the addition of the sensory data block was valuable to the overall information, the compromise in model efficiency was acceptable.

The similarities in the sample configurations were again assessed using RV coefficients (Table 5). The addition of sensory data resulted in lower RV coefficients between the mid-level PCA and the PCA for individual blocks compared with the RV coefficients between the low-level data fusion PCA and the PCA for individual data blocks.

As the fusion model is a composition of different data blocks originating from measurements of the different properties of wine, a resulting model that has a unique sample configuration was expected. Although the within-model redundancy cannot be calculated for a concatenated matrix, the RV coefficients (mid-level PCA vs. individual blocks range 0.52–0.88) could be considered an indicator of relatively low redundancy. The exception was, once more, the IR data block. As previously discussed (Section 3.1.2), the IR data block provided the most unique sample configuration pattern compared with the other data blocks, indicated by the low RV coefficients (IR vs. other data blocks, Supplementary Table S3). The IR sample configuration was the most similar to that of the mid-level PCA fusion model, indicated by high RV coefficients (mid-level PCA vs. IR) ranging from 0.88 to 0.96 for the CB data sets and 0.83 to 0.96 for the SB data sets. The pattern of RV coefficients between the mid-level PCA data fusion model and the other individual data blocks could not be generalized. The patterns were case-specific and unique for each data set. Much like the IR data block, due to its nature, the sensory data contained a unique profile of the wine, but, although the sensory experience is holistic, given the method used, the data captured were not. Pivot profiling is a comparative method, and not a profiling method, such as CATA, that includes a comprehensive list of attributes [8]. Compared with IR, which is unique and information-rich, the sensory data in this case were unique, but not as information-rich.

Table 5. Pairwise RV coefficients ($p \leq 0.01$) for the PCA/CA of individual blocks vs. the mid-level PCA fused model.

		Chenin Blanc	Sauvignon Blanc
AVN	IR	↑0.90	↗0.88
	ARP	↘0.67	↘0.61
	VCC	↗0.80	↓0.45
	UV-Vis	→0.76	↗0.82
	Sensory	→0.73	↘0.68
CDB	IR	↗0.88	↗0.86
	ARP	→0.75	↗0.83
	VCC	↘0.60	→0.72
	UV-Vis	→0.78	→0.76
	Sensory	↗0.84	→0.74
DTK	IR	↗0.88	↑0.93
	ARP	↘0.57	↘0.68
	VCC	↘0.65	↘0.64
	UV-Vis	↘0.56	↗0.86
	Sensory	↘0.66	→0.73
FRV	IR	↑0.95	↗0.83
	ARP	↗0.85	→0.75
	VCC	→0.74	→0.72
	UV-Vis	↗0.86	→0.72
	Sensory	↗0.84	→0.71
KZC	IR	↑0.96	↑0.96
	ARP	↘0.53	→0.79
	VCC	↘0.52	↓0.46
	UV-Vis	→0.78	↑0.93
	Sensory	↘0.63	↘0.61
PDB	IR	↑0.92	↑0.93
	ARP	↗0.88	↗0.86
	VCC	↘0.69	↘0.55
	UV-Vis	↗0.88	↗0.86
	Sensory	↗0.82	→0.75

The data sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB). Arrows indicate RV values higher than (diagonally upward), equivalent to (horizontal), or lower than (diagonally downward) 0.7. Significantly higher values are in green (vertically up) and significantly lower values are in red (vertically down).

3.3.2. Multiple Factor Analysis (MFA)

Unlike PCA, which aims to reduce the dimensionality and produce the most efficient model, the MFA seeks to create/build the most representative model of the relationships between blocks of data [26]. The figures of merit related to the performance of the MFA models are shown in Table 6. As a multiblock analysis, the stress calculated on the MFA was relative to the different blocks, and not the individual variables within each block [26]; as such, the eigenvalues were lower than those of the PCA data fusion models. The exponential decay curves had R^2 values ranging from 0.84 to 0.99, except for PDB SB, which had an R^2 of 0.71. Generally, the models had low efficiency; CB had higher efficiency, as indicated by the slopes (0.35 to 0.47), than SB (0.27 to 0.37). For all models, the stress was distributed gradually across the different dimensions, with less than 80%EV accumulated over the first three dimensions.

Table 6. Stress distribution over components (C) in the MFA data fusion.

					Cumulative %EV					
	Sample Set	**Total Stress (Eigenvalue)**	**Slope**	**R^2**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Chenin Blanc	AVN	9.8952	0.43	0.96	40	67	79	89	95	100
	CDB	9.7308	0.35	0.94	40	61	75	86	93	100
	DTK	9.0578	0.43	0.99	41	64	78	89	96	100
	FRV	9.7637	0.38	0.96	42	63	77	87	94	100
	KZC	8.9517	0.47	0.97	42	68	82	90	96	100
	PDB	9.0879	0.37	0.84	49	64	76	86	95	100
Sauvignon Blanc	AVN	9.3392	0.30	0.85	39	60	72	82	92	100
	CDB	10.6642	0.37	0.99	33	58	76	86	94	100
	DTK	10.3854	0.33	0.94	38	57	75	85	93	100
	FRV	9.3328	0.35	0.93	39	63	75	85	94	100
	KZC	10.7258	0.32	0.98	33	58	73	84	93	100
	PDB	9.21612	0.27	0.71	43	56	70	82	93	100
Range	Min	8.9517	0.27	0.71	33	56	70	82	92	100
	Max	10.7258	0.47	0.99	49	68	82	90	96	100

The data sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB).

An MFA model generates new weights for the different data blocks relative to each other and can thus show the correlations between different data blocks. This means that the MFA sample configuration was most representative of all the data blocks and was not skewed by any individual data block (as might be the case with low/mid-level data fusion PCA). RV coefficient values could be calculated between the sample configurations of the data blocks after weighing (Supplementary Table S5). The RV coefficients for the MFA (vs. individual data blocks) were higher than those of the mid-level data fusion PCA (vs. individual data blocks), ranging from 0.52 to 0.95 for CB and 0.64 to 0.92 for SB. The RV coefficients between MFA and the IR data block (ranging from 0.55 to 0.88 for CB and 0.64 to 0.87 for SB) were lower compared with those of the other data blocks (ranging from 0.76 to 0.95 for CB and 0.69 to 0.92 for SB). This is unlike the results for the PCA data fusion models (low and mid-level), in which the RV coefficients between the PCA data fusion model and the IR data block were the highest compared with the other data blocks (Table 5 and Supplementary Table S4). This is indicative of how the number and nature of the variables from the IR data block had a skewing effect on the PCA data fusion models. This means that, in the concatenated matrices, the IR data block influenced the sample configuration the most. This could not be directly demonstrated in the case of PCA due to the nature of the statistical analysis.

The sample configurations of the mid-level PCA and MFA fusion models were calculated using the conventional RV coefficient and inferred topology (iTOP) calculation of the RV (Table 7). The inferred topology (iTOP) RV reportedly takes into account the redundancy between data blocks and skewing by any one data block [39]. Although the iTOP RV coefficients were slightly lower than the conventional RV coefficient, they were similar. All RV coefficients were higher than 0.70, indicating very high similarity between the two approaches (iTOP vs. conventional), but, as the two data fusion models contained the same original data, this was expected. The burden now shifts to the 30% dissimilarity between the data fusion approaches.

3.4. General Discussion

The case chosen to illustrate the stepwise approach to data fusion had its particularities originating from the type of sensory method that generated the data and the fact that the data sets were first considered separately due to the original experimental design. However, these types of results are quite common in wine evaluation, where one or more analytical chemistry techniques are used in addition to (usually) one sensory method.

Table 7. Pairwise RV coefficients ($p \leq 0.01$) between PCA and MFA for the mid-level data fusion.

	SAMPLE SET	RV	ITOP RV
Chenin Blanc	AVN	↗0.82	→0.70
	CDB	↗0.82	→0.75
	DTK	↗0.80	↘0.62
	FRV	↑0.94	↑0.93
	KZC	↗0.85	↗0.80
	PDB	↑0.96	↑0.95
Sauvignon Blanc	AVN	→0.78	→0.77
	CDB	↑0.93	↑0.92
	DTK	↗0.81	→0.70
	FRV	↗0.89	↗0.85
	KZC	↗0.84	→0.79
	PDB	↗0.81	↗0.81

PCA—principal component analysis, MFA—multiple factor analysis, RV—regression vector, and iTOP—inference topometry. The sample sets are defined by three letters corresponding to each winery (i.e., AVN, CDB, DTK, FRV, KZC, and PDB). Arrows indicate RV values higher than (diagonally upward), equivalent to (horizontal), or lower than (diagonally downward) 0.7. Significantly higher values are in green (vertically up) and significantly lower values are in red (vertically down).

Different steps and levels of data modelling for the purposes of data fusion have been presented, from individual data blocks, low-level, and mid-level data fusion to multiblock data fusion. In assessing the different models, it is important to use multiple evaluation parameters that take into account different aspects of the models. In this study, the performance of the models was evaluated by looking at the distribution of the data over all dimensions and the rate of eigenvalue decay as indicators of model efficiency. This was because, for multiple data blocks and data fusion strategies, the orthodox use of %EV as a sole indicator of model performance is not appropriate. This tactic is used for choosing dimensions/factors on which to run pattern recognition analysis. Thus, this tactic does not appropriately evaluate the overall model performance. The RV coefficients were used to evaluate the representativeness of the fusion models and evaluate redundancy in the cases where other parameters could not be used. The RV coefficients were used because the comparison of various sample sets that consist of various data blocks using sample scores and variable loadings is not appropriate. The approach could not provide meaningful comparisons of information loss and information extraction for multiple data sets. This was especially true for data fusion models that had large data plots (greater than a thousand) that include untargeted spectral signals.

Low-level data fusion is generally appropriate for data blocks with only a small number of variables, as finding patterns in correlations between a large number of variables can be tedious and the visual aids offer very little assistance with the complex interpretations [23]. The low-level and mid-level PCA fusion models did not offer any information on the within-model correlations between data blocks. Although the models were highly efficient, they were not representative. Owing to the incompatibility of the sensory data matrix with the four chemistry data blocks, low-level PCA data fusion was not as comprehensive as the mid-level strategies. Although the addition of the sensory data block resulted in slightly lower model efficiency, the sensory aspect added to the overall informational value and comprehensiveness of the data fusion model; thus, a compromise in model efficiency must be made. For cases where the model efficiency was drastically lowered by the inclusion of a data block, the influence of the additional block must be further investigated. This can be achieved by revisiting the pre-modelling processing to "clean" the data.

Mid-level PCA data fusion models were skewed by the information-dense IR data block. This was revealed by the lower RV coefficients for mid-level PCA vs. individual blocks compared with mid-level PCA vs. the IR data block. The mid-level PCA sample configuration was thus unrepresentative of all of the blocks. The mid-level MFA models were less efficient than the PCA models, but were more representative of the commonality between data blocks, indicated by high RV coefficients (the models had sample configu-

rations more representative of all of the data blocks). Although the PCA fusion models were highly efficient (rate of eigenvalue decay), this was rather indicative of the overfitting of the data, as the models were also found to be unrepresentative due to skewing by the information-rich IR data block. Hence, by comparison, MFA proved to be less biased and more representative of the individual data blocks. Thus, multiblock approaches were determined to be more appropriate data fusion methods compared with concatenation.

4. Conclusions

The aim of this study was to explore and compare PCA (low-level and mid-level) and MFA data fusion approaches. The study evaluated model efficiency (%EV and slope of the exponential decay in stress) and model representativeness (within-model and between-model pairwise RV coefficients). Using these parameters, issues of overfitting of data and redundancy between the different data blocks were inferred. Adding more data, especially data of a different nature/origin, resulted in a decrease in model efficiency. As the addition of more data of different variations is the motivation of data fusion, the model efficiency was found to act as an ineffective evaluation parameter for data fusion models. The RV coefficients were a more effective parameter for evaluating data fusion model performance. However, RV could not be used within low/mid-level PCA data fusion models; only the MFA multiblock strategy offered this feature. It is for these reasons that, for multivariate and distinct data sets, such as those presented in this study, MFA should be considered an appropriate option for unsupervised data fusion.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/beverages8040066/s1, Table S1: Pairwise regression vector (RV) coefficients ($p \leq 0.01$) for score configuration using IR raw data vs. its mathematical transformations using multiplicative scatter correction (MSC), first derivative (1st deriv), and their combinations before Principal Component Analysis (PCA). Table S2: Cumulative percentage explained variance (%EV) of the first two principal components of the PCA (VCC, ARP, UV-Vis, and IR) and first two dimensions of the CA (sensory) for individual data sets. Table S3: Pairwise RV coefficients ($p \leq 0.01$) for the scores of the individual data blocks. Table S4: Pairwise RV coefficients ($p \leq 0.01$) for the PCA scores of the chemistry data blocks vs. low-level PCA fused model. Table S5: Pairwise RV coefficients ($p \leq 0.01$) between MFA and individual data blocks PCA/CA.

Author Contributions: Conceptualization, A.B.; methodology, A.B., J.B. and M.K.; software, M.K. and J.B.; validation, M.K. and J.B.; formal analysis, M.M.; investigation, M.M.; resources, A.B.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, A.B., J.B., M.M., M.K. and A.M.; visualization, M.K. and M.M.; supervision, A.B. and A.M.; project administration, A.B.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation (NRF) of South Africa, grant number 113761.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors. The data are not publicly available due to them being institutional intellectual property.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CA, correspondence analysis; PCA, principal component analysis; MFA, multiple factor analysis; %EV, percentage explained variance; MSC, multiplicative scatter correction; 1st deriv, first derivative; IR, infra-red; UV-Vis, ultra-violet visible light; ARP, antioxidant-related parameters; VCC, volatile compound composition; RV, regression vector coefficient; iTOP, inferring topology RV.

References

1. Gagolewski, M. *Data Fusion. Theory, Methods, and Applications*; Hryniewicz, O., Mielniczuk, J., Penczek, W., Waniewski, J., Eds.; Institute of Computer Science, Polish Academy of Sciences: Warsaw, Poland, 2015; ISBN 9788363159207.
2. Lahat, D.; Adalı, T.; Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. *Inst. Electr. Electron. Eng.* **2015**, *103*, 1449–1477. [CrossRef]
3. Cocchi, M. Data fusion methodology and applications. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 1–370. ISBN 9780444639844.
4. Arvanitoyannis, I.S.; Katsota, M.N.; Psarra, E.P.; Soufleros, E.H.; Kallithraka, S. Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends Food Sci. Technol.* **1999**, *10*, 321–336. [CrossRef]
5. Iorgulescu, E.; Voicu, V.A.; Sârbu, C.; Tache, F.; Albu, F.; Medvedovici, A. Experimental variability and data pre-processing as factors affecting the discrimination power of some chemometric approaches (PCA, CA and a new algorithm based on linear regression) applied to (+/-)ESI/MS and RPLC/UV data: Application on green tea extrac. *Talanta* **2016**, *155*, 133–144. [CrossRef] [PubMed]
6. Silvestri, M.; Elia, A.; Bertelli, D.; Salvatore, E.; Durante, C.; Li Vigni, M.; Marchetti, A.; Cocchi, M. A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 181–189. [CrossRef]
7. Alañón, M.; Pérez-Coello, M.; Marina, M. Wine science in the metabolomics era. *Trends Anal. Chem.* **2015**, *74*, 1–20. [CrossRef]
8. Valentin, D.; Chollet, S.; Lelièvre, M.; Abdi, H. Quick and dirty but still pretty good: A review of new descriptive methods in food science. *Int. J. Food Sci. Technol.* **2012**, *47*, 1563–1578. [CrossRef]
9. Granato, D.; de Araújo Calado, V.M.; Jarvis, B. Observations on the use of statistical methods in Food Science and Technology. *Food Res. Int.* **2014**, *55*, 137–149. [CrossRef]
10. Cocchi, M. Introduction: Ways and Means to Deal With Data From Multiple Sources. In *Data Handling in Science and Technology*; Elsevier Ltd.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 1–26.
11. Brand, J. Rapid Sensory Profiling Methods for Wine: Workflow Optimisation for Research and Industry Applications. Ph.D. Thesis, Stellenbosch University, Stellenbosch, South Africa, 2019.
12. Rinnan, Å.; van den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
13. López-Rituerto, E.; Savorani, F.; Avenoza, A.; Busto, J.H.; Peregrina, J.M.; Engelsen, S.B. Investigations of la Rioja terroir for wine production using 1H NMR metabolomics. *J. Agric. Food Chem.* **2012**, *60*, 3452–3461. [CrossRef]
14. Ragone, R.; Crupi, P.; Piccinonna, S.; Bergamini, C.; Mazzone, F.; Fanizzi, F.P.; Schena, F.P.; Antonacci, D. Classification and Chemometric Study of Southern Italy Monovarietal Wines Based on NMR and HPLC-DAD-MS. *Food Sci. Biotechnol.* **2015**, *24*, 817–826. [CrossRef]
15. Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. Data fusion methodologies for food and beverage authentication and quality assessment-A review. *Anal. Chim. Acta* **2015**, *891*, 1–14. [CrossRef] [PubMed]
16. Brand, J.; Panzeri, V.; Buica, A. Wine quality drivers: A case study on South African chenin blanc and pinotage wines. *Foods* **2020**, *9*, 805. [CrossRef] [PubMed]
17. Biancolillo, A.; Boqué, R.; Cocchi, M.; Marini, F. Data Fusion Strategies in Food Analysis. In *Data Handling in Science and Technology*; Elsevier Ltd.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 271–310.
18. Pereira, A.C.; Carvalho, M.J.; Miranda, A.; Leça, J.M.; Pereira, V.; Albuquerque, F.; Marques, J.C.; Reis, M.S. Modelling the ageing process: A novel strategy to analyze the wine evolution towards the expected features. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 176–184. [CrossRef]
19. Valente, C.C.; Bauer, F.F.; Venter, F.; Watson, B.; Nieuwoudt, H.H. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. *Sci. Rep.* **2018**, *8*, 4987. [CrossRef] [PubMed]
20. Ballabio, D.; Todeschini, R.; Consonni, V. Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data. In *Data Handling in Science and Technology*; Elsevier Ltd.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 129–155.
21. Pagés, J.; Husson, F. Multiple factor analysis with confidence ellipses: A methodology to study the relationships between sensory and instrumental data. *J. Chemom.* **2005**, *19*, 138–144. [CrossRef]
22. Salkind, J.; Kristin, R. *Encyclopidia of Measurement and Statistics*; Salkind, N.J., Ed.; Sage: Newcastle upon Tyne, UK, 2007; ISBN 9781412916110.
23. McKillup, S. *Statistics Explained: An Introductory Guide for Life Scientists*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2012; ISBN 9781107005518.
24. Borgognone, M.G.; Bussi, J.; Hough, G. Principal component analysis in sensory analysis: Covariance or correlation matrix? *Food Qual. Prefer.* **2001**, *12*, 323–326. [CrossRef]
25. Pagès, J. Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Qual. Prefer.* **2005**, *16*, 642–649. [CrossRef]
26. Abdi, H.; Valentin, D. Multiple Factor Analysis (MFA). *Encycl. Meas. Stat.* **2007**, *1*, 657–663.
27. de Tayrac, M.; Lê, S.; Aubry, M.; Mosser, J.; Husson, F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genom.* **2009**, *10*, 32. [CrossRef]

28. Baldwin, E.; Han, J.; Luo, W.; Zhou, J.; An, L.; Liu, J.; Zhang, H.H.; Li, H. On fusion methods for knowledge discovery from multi-omics datasets. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 509–517. [CrossRef]
29. Pagès, J. Multiple factor analysis: Main features and application to sensory data. *Rev. Colomb. Estad.* **2004**, *27*, 1–26.
30. Cadena, R.S.; Cruz, A.G.; Netto, R.R.; Castro, W.F.; Faria, J.D.A.F.; Bolini, H.M.A. Sensory profile and physicochemical characteristics of mango nectar sweetened with high intensity sweeteners throughout storage time. *Food Res. Int.* **2013**, *54*, 1670–1679. [CrossRef]
31. Mafata, M.; Brand, J.; Medvedovici, A.; Buica, A.; Mafata, M.; Brand, J.; Medvedovici, A.; Buica, A. Chemometric and sensometric techniques in enological data analysis. *Crit. Rev. Food Sci. Nutr.* **2022**, 1–15. [CrossRef]
32. Le Dien, S.; Pagès, J. Hierarchical Multiple Factor Analysis: Application to the comparison of sensory profiles. *Food Qual. Prefer.* **2003**, *14*, 397–403. [CrossRef]
33. Abdi, H. RV Coefficient and Congruence Coefficient. *Encycl. Meas. Stat.* **2007**, *1*, 849–853.
34. Mafata, M.; Brand, J.; Panzeri, V.; Kidd, M.; Buica, A. A multivariate approach to evaluating the chemical and sensorial evolution of South African Sauvignon Blanc and Chenin Blanc wines under different bottle storage conditions. *Food Res. Int.* **2019**, *125*, 108515. [CrossRef] [PubMed]
35. Antúnez, L.; Salvador, A.; de Saldamando, L.; Varela, P.; Giménez, A.; Ares, G. Evaluation of Data Aggregation in Polarized Sensory Positioning. *J. Sens. Stud.* **2015**, *30*, 46–55. [CrossRef]
36. Fleming, E.E.; Ziegler, G.R.; Hayes, J.E. Check-all-that-apply (CATA), sorting, and polarized sensory positioning (PSP) with astringent stimuli. *Food Qual. Prefer.* **2015**, *45*, 41–49. [CrossRef]
37. Thuillier, B.; Valentin, D.; Marchal, R.; Dacremont, C. Pivot© profile: A new descriptive method based on free description. *Food Qual. Prefer.* **2015**, *42*, 66–77. [CrossRef]
38. Lelièvre-Desmas, M.; Valentin, D.; Chollet, S. Pivot profile method: What is the influence of the pivot and product space? *Food Qual. Prefer.* **2017**, *61*, 6–14. [CrossRef]
39. Aben, N.; Westerhuis, J.A.; Song, Y.; Kiers, H.A.L.; Michaut, M.; Smilde, A.K.; Wessels, L.F.A. iTOP: Inferring the topology of omics data. *Bioinformatics* **2018**, *34*, 988–996. [CrossRef] [PubMed]
40. Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Downey, G.; Blanchet, L.; Buydens, L.M. Breaking with trends in pre-processing? *Trends Anal. Chem.* **2013**, *50*, 96–106. [CrossRef]
41. Smilde, A.K.; Van Mechelen, I. A Framework for Low-Level Data Fusion. In *Data Handling in Science and Technology*; Elsevier Ltd.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 27–50.
42. Umetrics, M. User Guide to SIMCA 13. *Umetrics* **2012**, *13*, 1–661. [CrossRef]
43. Gishen, M.; Dambergs, R.G.; Cozzolino, D. Grape and wine analysis-enhancing the power of spectroscopy with chemometrics. *Aust. J. Grape Wine Res.* **2005**, *11*, 296–305. [CrossRef]
44. Stevenson, T. *The-New-Sothebys-Wine-Encyclopedia*, 4th ed.; Dorling Kindersley Limited: London, UK, 2005.
45. Ríos-Reina, R.; Callejón, R.M.; Savorani, F.; Amigo, J.M.; Cocchi, M. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta* **2019**, *198*, 560–572. [CrossRef] [PubMed]
46. Robinson, J.W. *Practical Handbook of Spectroscopy*; CRC Press: Boca Raton, FL, USA, 2017; ISBN 9781351422789.