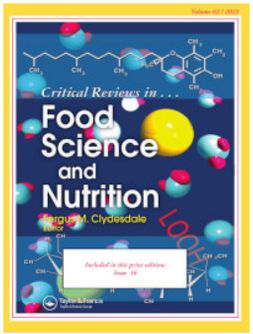

Taylor & Francis
Taylor & Francis Group

Critical Reviews in Food Science and Nutrition

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/bfsn20

Chemometric and sensometric techniques in enological data analysis

Mpho Mafata, Jeanne Brand, Andrei Medvedovici & Astrid Buica

To cite this article: Mpho Mafata, Jeanne Brand, Andrei Medvedovici & Astrid Buica (2022): Chemometric and sensometric techniques in enological data analysis, Critical Reviews in Food Science and Nutrition, DOI: 10.1080/10408398.2022.2089624

To link to this article: https://doi.org/10.1080/10408398.2022.2089624

Published online: 22 Jun 2022.

CRITICAL REVIEWS IN FOOD SCIENCE AND NUTRITION
https://doi.org/10.1080/10408398.2022.2089624

Taylor & Francis
Taylor & Francis Group

REVIEW

Chemometric and sensometric techniques in enological data analysis

Mpho Mafata[a,b], Jeanne Brand[a], Andrei Medvedovici[c] and Astrid Buica[a,b]

[a]South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Stellenbosch University, Stellenbosch, South Africa; [b]School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa; [c]Department of Analytical Chemistry, Faculty of Chemistry, University of Bucharest, Bucharest, Romania

ABSTRACT

Enological evaluations capture the chemical and sensory space of wine using different techniques; many sensory methods as well as a variety of analytical chemistry techniques contribute to the amount of information generated. Data fusion, especially integrating data sets, is important when working with complex systems. The success reported when trying to integrate different modalities is generally low and has been attributed to the lack of statistically considerate strategies focusing on the data handling process. Multiple stages of data handling must be carefully considered when dealing with multi-modal data. In this review, the different stages in the data analysis process were examined. The study revealed misconceptions surrounding the process and elucidated rules for purpose-driven approaches by examining the complexities of each stage and the impact the decisions made at each stage have on the resulting models. The two major modeling approaches are either supervised (discrimination, classification, prediction) or unsupervised (exploration). Supervised approaches were emphatic on the pre-processing steps and prioritized increasing performance. Unsupervised approaches were mostly used for preliminary steps. The review found aspects often neglected when it came to the data collection and capturing which in the end contributed to the low success in combining sensory and chemistry data.

KEYWORDS

Chemometrics; data analysis; data concatenation; data integration; data fusion; multivariate analysis; multi-modal; sensometrics

Introduction

Statistical analysis is used in applied sciences to evaluate experimental results and enhance the interpretation of their significance. The use of statistical analysis in Chemistry is referred to as *chemometrics* (Kowalski 1980). Chemometrics has been used in several different natural science fields, including food chemistry. In subsequent years, the term *sensometrics* was coined for the statistical analysis of sensory and consumer science data (Hunter et al. 1995). Chemometrics and sensometrics have been developed to handle large data, but the more information introduced into a model, the more complex assessing relationships between observations (*e.g.* samples) and variables (*e.g.* treatments) becomes (McKillup 2005). In such cases, multivariate data analysis tools that reduce the dimensionality of large data in order to highlight and visualize the important features describing the overall relationships are needed (Granato, Ônica Maria de Araújo Calado, and Jarvis 2014).

Data fusion (defined as combining and integrating different data sets) is important when working with complex systems such as natural products (White 1991; Cocchi 2019b). Data fusion systems provide holistic and comprehensive data models (Cocchi 2019a). These data models are holistic in the sense that they accommodate different perspectives (modalities) and comprehensive in that they create a representative picture of the entire natural system. Data integration systems are used in a wide variety of fields for information retention, interpretation, and decision-making (Borràs et al. 2015; Cocchi 2019a).

Enological studies evaluate the behavior of wine throughout the winemaking process under different treatment/conditions such as temperature (Serra-Cayuela et al. 2014; Mafata et al. 2019; Mafata, Buica, Du Toit, et al. 2018; Du Toit and Piquet 2016) and temporal changes (Pereira et al. 2011; Coetzee et al. 2016; Pereira et al. 2016). The field has advanced to use holistic measurements that capture various sensory and chemical responses to the treatments/conditions, resulting in the development of a variety of analytical chemistry techniques and several rapid sensory methods. More measurements result in generating more data and a more comprehensive product profile, but some methods may be redundant in the information they provide. It is thus important to use techniques that are appropriate and information-rich (Cocchi 2019b; Borràs et al. 2015). Evaluating the redundancy of measurements can be based on an understanding of the theoretical and practical principles behind each method.

Data fusion approaches can be sectioned into four parts: input (what goes into the model), modeling (how data are treated), output (what comes out of the model), and interpretation (what it all means). The input involves acquisition and treatment of data to prepare it for modeling. The

CONTACT Astrid Buica astrid.buica@gmail.com

modeling is dependent on the research question and type of data acquired. The output refers to tables of calculations of model performance parameters and related figures of merit/evaluation parameters. Interpretation of models for the application involves the use of visual aids and evaluation parameters generated from the various model outputs used to evaluate the model performance. Evaluating the success of a data fusion model is based on the statistical significance of the figures of merit and on the motivations behind the data fusion (in the applied sense).

When the level of success reported when trying to integrate the multiple measurements is low, this can be attributed to a lack of statistically considerate strategies, highlighting a need for more sophisticated thinking behind the proposed strategies. The motivations behind data fusion can be problem-focused if the aim is to articulate the problem and analyze the problem space. This motivation leads to approaches that are unsupervised, explorative, and indirect - generally hypothesis-forming -, but can be used as a stage in hypothesis-testing approaches. The motivation can also be solution-centered in that it seeks to find the best possible answer to the problem. This motivation leads to the use of supervised and directed data analysis methods for prediction, classification, or discrimination, which are generally hypothesis testing. In both approaches, the appropriate method must be aligned to the motivation.

To evaluate the use of statistical strategies (especially data fusion) in Enology, a descriptive bibliometric search was performed including documents published in the past decade (2010 – 2020). The analysis used two credible academic databases, namely Scopus and the Commonwealth Agricultural Bureaux Index or Center for Agriculture and Biosciences International (CABI). The Scopus database was used because of its up-to-date, diverse index systems from many publishers. CABI was used because it specifically indexes agriculture, forestry, and related disciplines. The search string "(wine OR enology OR oenology) AND (data AND fusion)" was used, based on terms found in the title, author- or database-supplied keywords, and abstract.

As of October 2020, the CABI databased returned only 13 results for publications on data fusion. The Scopus search returned 279 results, of which 187 were research articles, 31 reviews, and 26 book chapters. The past ten years have seen a gradual increase in research publications using multivariate tools in wine-related research (an average increase of 20 articles per year). Most investigations that are Chemistry orientated refer to the use of statistics in wine science as '*chemometrics*', while those that focus on the enological application often refer to it as '*omics*' ('*metabolomics*' or '*wineomics*') (Moyano et al. 2019; Alañón, Pérez-Coello, and Marina 2015). For sensory investigations in enology, most publications use the term '*multivariate analysis*'; only some of the sensory investigations refer to the data handling as '*sensometrics*' (Guld et al. 2020; Cariou et al. 2018; Brand 2019). The use of the term '*data fusion*' explicitly (as author-supplied keyword) was returned only 20 times, with the focus of the approaches being split between application and statistics.

Statistical investigations in enology can focus either on a specific application or on methodology development. In most reported cases, both use statistical analysis for hypothesis testing (Granato and Ares 2013; Granato, Ônica Maria de Araújo Calado, and Jarvis 2014). The advantage of exploring approaches focused on hypothesis forming is that it can shed light on the underlying intricacies and difficulties of the data handling process in enology. In turn, this can underscore the aspects of the methodology that may need to be improved and can lead to better hypothesis-proving methods. In this context, the current literature review will examine the different stages of data fusion and elucidate the rules for data handling in enology. It will detail the differences between the chemometric and sensometric treatments of the data according to the literature, and comment on the impact decisions made at each stage have on the resulting data fusion model.

Evaluation in enology and rationale behind the movement toward multivariate statistical analysis (MVA)

Categories of chemistry and sensory methods

The major modes of evaluation in enology are chemical and sensorial. Chemical methods can be broadly categorized under targeted and untargeted (Alañón, Pérez-Coello, and Marina 2015). Targeted methods produce discreet measurements (usually concentrations of compounds translated mathematically from the detector response), whereas untargeted methods can produce continuous or discreet measurements (*e.g.,* full chromatograms *vs* peak areas). An analysis technique can be used in either targeted or untargeted manner depending on the research question (Godelmann et al. 2013). Chemical data used in enology can be further sub-categorized into volatile and nonvolatile compounds, broadly corresponding to analysis done in liquid or gas phase, and can be linked to sensory stimulation. Investigations that use such categories do so with the intent to link the sensory perception to the chemical composition (Lapalus 2016; Borràs et al. 2015) of a sample or a set of samples. The compounds can be further sub-categorized according to their chemical properties, linked to the size of the compounds and their functional groups. Untargeted analysis methods are widely used for authentication applications (Ríos-Reina et al. 2019; Alañón, Pérez-Coello, and Marina 2015; Borràs et al. 2015). Untargeted statistical techniques use supervised data models for prediction or classification of samples (Versari et al. 2014) and few have attempted to use untargeted analysis techniques to predict sensory data (Niimi, Tomic, et al. 2018).

Sensory methods can be categorized based on the information collected and the manner of execution, which has implications on the psychological aspects of the methodology (Valentin et al. 2012). The broadest categories are verbal *vs* non-verbal methods and single *vs* multiple presentations (Brand 2019). Verbal methods use attributes (sensory

descriptors) to describe the samples and/or the relationship between samples. Descriptive Analysis (DA) and its variants are the most widely used verbal methods (Campo et al. 2010; Murray, Delahunty, and Baxter 2001; Torrens et al. 2010). The development of rapid methods saw the use of other verbal methods such as check-all-that-apply (CATA) and flash profiling (Fleming, Ziegler, and Hayes 2015; Ares et al. 2010) and non-verbal methods such as rating, which measures a single sensory character of each sample (Ballester et al. 2008). Methods with multiple presentations can be similarity-based such as sorting and Projective Mapping (PM) or reference-based such as Pivot© Profiling (Valentin et al. 2012). Mixed method approaches use a combination of verbal or non-verbal aspects where one task is primary, while the other is secondary (Brand 2019). An example of a mixed method is sorting (primary) with a descriptive element to the grouping (Valentin et al. 2012; Mafata, Buica, Du Toit, et al. 2018; Ballester et al. 2008).

Statistical approaches in evaluating enological experiments

Advances in statistical data handling techniques have naturally progressed to analyze more variables simultaneously from univariate, bivariate, multivariate, to what is sometimes called megavariate data analysis (Eriksson et al. 2006). Univariate analysis looks at the variation in one or two variables across samples. Looking simultaneously at more than three variables created the need for multivariate techniques (McKillup 2005). Univariate data treatment is still important even in the context of multivariate analysis (MVA) and can be used to look deeper into the MVA results (Granato, Ônica Maria de Araújo Calado, and Jarvis 2014). Megavariate is often used for advanced multivariate techniques that use multiple sets of data acquired from different sources, requiring specialized statistical treatment (Eriksson et al. 2006). The multiple data sets are designated as blocks and used in multiblock and data fusion approaches (Cocchi 2019b). Each have their merit, but the reasoning is, when looking at evaluating complex systems like natural products, holistic approaches must be taken on all fronts: methodology, execution, and data analysis.

MVA is becoming more common in enological approaches mainly due to an increase in the number of methods from both Chemistry and Sensory (Alañón, Pérez-Coello, and Marina 2015). Chemistry methods have increased in numbers and sophistication, in accordance with advances in technological and computing power (Gagolewski 2012; Alañón, Pérez-Coello, and Marina 2015; Borràs et al. 2015). The variety of methods have increased, leading to opportunities in measuring more wine-related chemical compounds. The increase in the number of sensory methods was due to the need to address shortcomings in the already existing methodologies, related to differences in panels used for evaluation, the time, and cost of the analysis (Varela and Ares 2014; Valentin et al. 2012). Several rapid methods have recently been developed and have resulted in works using several sensory methods in a single study, something that was not always possible due to the limitations previously mentioned (Ballester et al. 2013; Hayward et al. 2020).

From an applied perspective, multivariate statistical analyses can be categorized under supervised and unsupervised methods (Sohail and Arif 2020). The motivation behind supervised methods is to target a specific outcome from the analysis, whether it be a grouping of samples according to similarities (classification) or differences (discrimination), or prediction. Unsupervised methods look for inherent patterns in the data without imposing a specific targeted outcome. Both approaches look to lower the number of dimensions and find the best-fit model for the purpose of the experiment (McKillup 2005; Sohail and Arif 2020). From a theoretical perspective, multivariate methods can be categorized as parametric (classical approach) and non-parametric (non-classical/advanced approach) (Härdle and Simar 2015; McKillup 2005). Classical approaches assume a normal distribution of data around an average and fit the data according to how similar they are to this mean. Classical approaches include grouping (cluster analyses), regressions (least squares), similarity/dissimilarity (correspondence and generalized correspondence analyses). Non-parametric analyses such as machine learning techniques do not assume normal distribution or a fixed average (Härdle and Simar 2015).

To date, the literature surveyed indicate that research in enology frequently seeks to understand what drives/contributes to predictions and classification, and hence use supervised data analyses to find the discriminating markers (Brand, Panzeri, and Buica 2020). Advanced data handling techniques such as *k*-nearest neighbors (kNN) have provided a good starting point to dig deeper into these types questions (Carvalho Rocha, Prado, and Blonder 2020). In enology and sensory research, Artificial Intelligence (AI) applications have been used in supervised strategies (Valente et al. 2018; Carvalho Rocha, Prado, and Blonder 2020). Unsupervised advanced strategies as well as other simple machine learning strategies have seldom been explored indicating a possible lack of confidence in using these data analysis approaches (Carvalho Rocha, Prado, and Blonder 2020).

Model input

Only what has been captured can be modeled; therefore, data collection and capturing are of outmost importance. The collection of data refers to the acquisition of the data related to the method and/or technique applied based on the experimental design. An experimental design that consolidates the sensory and chemistry data is most advised for data-orientated approaches in enology. Before the data captured can be modeled, several decision steps concerning the pre-modeling processes of the data and the modeling specifications must be taken. Several standardized pre-modeling processes have been developed for chemistry, but few are available for sensory data. Furthermore, due to the focus being mainly on the application, model specifications are seldom discussed in the literature, which creates

a gap in knowledge from the statistical handling of enological data perspective. There is an imbalance of greater detailing of the strategy behind the method compared to the data modeling. This section will cover how to convey important specifications and create a complete methodology based on important aspects of the data input stage.

Data collection and capturing

Prior to collecting and capturing experimental data, an intelligent design must be planned. An experimental design based on statistics determines the experimental execution and the data handling tools to be ultimately used (Yu, Low, and Zhou 2018; McKillup 2005). Several experimental designs have been development from a statistical perspective (Ferreira 2019), as well as for natural sciences perspective, including chemistry (Kreutz and Timmer 2009). Recently, design of experiments (DOEs) that are particularly sensitive to the structure and premise of sensory methods have been reviewed (Yu, Low, and Zhou 2018). DOEs are important in the natural sciences since they consider multiple (potentially) influential factors which may not always be possible to take into account for every experiment. Planning an intelligent DOE increases the chances of successful experimental outputs and data modeling, thus it is important to take time and create a DOE that is aligned with the research question.

Analytical chemistry instruments can have a single acquisition mode or multiple acquisition modes in which case they become hyphenated (Alañón, Pérez-Coello, and Marina 2015). Hyphenated techniques measure several responses and capture them in a conjugated (syn.: coupled/connected) manner. Software coupled to hyphenated techniques may capture the responses in independent channels and/or in a conjugated matrix. For example, in liquid chromatography coupled with mass spectrometry (LC/MS) the data can be extracted as a chromatogram or a matrix (Versari et al. 2014). A chromatogram can be extracted in two modes, selected ion monitoring (IEC - ion extracted chromatogram; SIM is a special way of exploiting the mass analyzer in order to monitor a single m/z channel) or total ion current (TIC - resulting from the full scan exploitation of the mass analyzer) which are two-dimensional representation of the retention time (RT) *vs* ion abundance. The matrix is extracted as RT_mass-to-charge pair (RT_m/z) *vs* ion abundance for each channel. The software generates automated outputs that, even in the case of hyphenated techniques, can provide the user with choices as to which information to capture. The hyphenated instruments are set-up in such a way that there is a single output, in which the different channels are captured as a single matrix aligned across a common array/dimension, usually the retention time. This is the case of multiple detectors such as fluorescence followed by MS (Terblanche 2017), or UV-Vis (diode-array detection/DAD or photodiode-array detection/PDA) followed by MS (Trikas et al. 2016).

Sensory data collection is related not only to the category of the method; the specific instructions given to a panel are also important. Instructions must be clear and unambiguous to collect relevant data which is compatible with the experimental design. Some methods may have verbal and non-verbal aspects; one aspect will constitute the primary objective while the other will be secondary. For methods using more than one task, panel fatigue must also be considered. Since sensory data cannot always be captured automatically, it is important to keep the different elements (panels, sessions, flights, judges, samples, attributes, and repeats) and the different aspects (verbal and non-verbal) separate until the consolidation stage, in order to have an accurate record of the raw data.

Recent developments of rapid sensory methods can be likened to hyphenated chemistry methods since they do measurements in several different ways in one evaluation session (mixed methods). These methods result in increased data generation and informational value which can be gained. Barriers to this "hyphenated" consideration of sensory data is the number of samples that can be evaluated in one flight or session, due to panel fatigue. A common approach is multiple sessions with multiple/different methods. To ensure compatibility between the methods, there needs to be alignment along at least one dimension, usually the samples. Sensory methods which are directed (*e.g.* Descriptive Analysis, DA) rarely require data cleaning and consolidation since the attributes chosen are carefully selected through trained panels or sensory screening (Chollet, Valentin, and Abdi 2005; Faye et al. 2013; Makhotkina, Pineau, and Kilmartin 2012). In most sensory methods that are undirected (*e.g.,* free-sorting and word association), some manual cleaning of results is needed; these aspects will be discussed in the next section.

Pre-modeling processing and transformations

Data pre-processing can be done automatically, manually, or based on statistical reasoning. In order to model data, it first needs to be fitted into the same scale (usually into a normal distribution) to limit any bias in calculations and models (McKillup 2005). Chemistry data sets are generally pre-processed automatically based on certain mathematical reasoning. Sensory data is generally first pre-processed manually even if the data collection is done automatically. Statistical pre-processing methods such as centering and/or scaling are done for both chemistry and sensory data before modeling (McKillup 2005).

Chemical data processing is such that the data standardization can be obtained after the acquisition. The pre-processing of chemistry data is related to the modes (types) of acquisition and the dimensionality (Salkind and Kristi; Deneulin and Bavaud 2016). Targeted analyses tend to produce data sets with smaller dimensions/variables than untargeted data sets and generally are not pre-processed (Jansen et al. 2013). Targeted data can, however, be converted to different units of measurement or indices. For example, measurements of phenolics can use UV-Vis spectrophotometric absorbance units at different wavelengths, equivalents to appropriate standards such as gallic acid, or

can be measured using indices such as CIELab or color density (OIV 2006; Waterhouse 2002; Ribéreau-Gayon et al. 2006). Untargeted data sets often have associated pre-processing methods such as those developed for IR, NMR, Raman spectroscopy, and UV-Vis (Rinnan, Berg, and Engelsen 2009; Campos and Reis 2020). Untargeted data sets have particularities related to their acquisition, and the nature of the sample for which the pre-processing is done to address issues such as baseline offset and noise, and saturated peaks often seen in NMR, IR, and UV-Vis spectra (Rinnan, Berg, and Engelsen 2009; Jansen et al. 2013).

Sensory data cleaning involves linguistic and semantic reduction through consolidation, concatenation, and sometimes deletion. Analyses such as DA, that use trained/analytical panels in which the attributes are chosen in such a way that they are representative of the group of samples, do not require data cleaning/pre-processing (Murray, Delahunty, and Baxter 2001). Although no standardized rules for the consolidation of attributes exists, there is a theoretical framework (Valentin et al. 2012). Depending on the acquisition method, the general sensory components are color/appearance, aroma, taste, mouthfeel/trigeminal sensations (Valentin et al. 2012). Further sub-categorization from this point becomes complex; it can be based for example on certain foodstuff groupings (*e.g.*, 'lemon', 'lime', 'orange', 'clementine' belong to 'citrus') or on common sources for the sensation (*e.g.,* 'woody', 'planky', 'oaky', 'coconut' are related to wood contact). Adjectives which give not only a specific descriptor (*e.g.*, 'apple'), but further describe it (*e.g.,* 'yellow', 'green', 'overripe', 'baked') are often kept separate because they create a new attribute. This aspect is often not standardized, even though comprehensive lists exist, often in the form of aroma or mouthfeel wheels (Pickering and Demiglio 2008; Gawel, Oberholster, and Francis 2000; Lawless and Civille 2013).

In practice, the approach is from the lowest level upwards or a bottom-up approach (synonyms, lemmatization, and grouping) where a descriptor can be eliminated due to low frequency of citation by a limited number of judges. Sensory methods are developed together with appropriate statistical analyses, which factor in the manner (verbal or non-verbal) and execution (single or multiple presentation) of the task (Valentin et al. 2012). The statistical pre-processing may involve concatenation, merging different blocks such as sessions, verbal and non-verbal aspects, and tasting repeats (Cardello et al. 1982). Another element to consider is the panel used: expert *vs* consumer *vs* trained (analytical). When considering the semantic consolidation, differences among the panel members can change the meaning of the attributes due to their different use and understanding of the lexicon, for example the meaning of texture (Chrea et al. 2005; Deneulin and Bavaud 2016) and perception of minerality (Ballester et al. 2013).

Statistical consolidation of intensity and frequency-based data includes imposing a limit on the intensity or frequency and/or a cutoff for the number of citations per attribute. Caution needs to be taken when considering the rules for consolidating the data. The difficulties and intricacies mentioned show case specificity of sensory data consolidation, emphasizing the reasons why it is difficult to standardize. Due to this, it is accepted that the semantic consolidation must be done in agreement by at least three specialists. It takes knowledge and experience to evaluate when exclusion of data constitutes data cleaning or a loss in information, for both chemistry and sensory data pre-processing.

Data modeling and performance parameters

When choosing how to model data, decisions are made based on the experimental question from which the design of experiment is derived and the data generated. The main aspects of choosing which data modeling to use is based on hypothesis testing or hypothesis formulating intent. The choice involves ensuring matrix compatibility and supervised or unsupervised purpose. The chosen model must be able to properly address the research question, therefore the steps of data collection, capturing, and pre-processing must be executed in consideration of the modeling. Depending on the type of data available, some modeling opportunities may not be possible. In some cases, a pre-processing step may be enough to address issues related to compatibility between data matrix type and model, but a conversion into a compatible format may not always be appropriate. Matrix compatibility concerns the type of data (values), the matrix dimensions, variability, and repeats, which will influence the modeling that can be done. Large data sets with high sample numbers (distinguishable samples, not including repeats), high sample variability, large number and diverse natures of measurements, can be modeled in different ways depending on the research question. Such a design is desirable for complex systems since the same data can be used to mine different information using various modeling tools.

As previously mentioned, the algorithms that are used to model data can be either supervised or unsupervised (Sohail and Arif 2020). The mathematical aspects related to these models will not be covered in this review, which will take a process-centered look at the aspects of modeling from an application perspective. In order to apply supervised models, the sample size must be large enough and contain enough variability to allow for classification, discrimination, or prediction. These two factors (number and variability in samples) have been shown to impact the performance of the supervised models. Unsupervised models require a good sample size but not such an extensive variation in the data set. The main requirement in unsupervised data models is compatibility between the matrix and the type of model desired.

Matrix compatibility

Chemical data generally has standardized outputs in compatible matrices, making various data modeling opportunities possible. Chemical instrumental analyses output data sets with single array correlation matrices of observations *vs*

measurements. In the case of hyphenated techniques, depending on the number of modes, instruments output multiple array matrices. Even given the differences in number of arrays, the modes are still compatible if one of the arrays is kept similar and the values are normalized or scaled (*e.g.,* LC-FLD-MS, where due to the serial setup there will be a constant delay between the retention time in the FLD chromatogram and the retention time in the MS chromatogram). The distribution of data in a discreet data set *vs* a continuous data set is different, making it difficult to combine the two. Since the data is scaled before modeling, the assumption in statistical context is that the distribution of the two is the same. Therefore, continuous data is often scaled differently from discreet data sets; to combine them, they are first scaled separately and then combined.

In sensory, methods are frequently developed with the statistics as part of the design of experiments (Valentin et al. 2012; Yu, Low, and Zhou 2018). As previously discussed, the execution has implications on the data analysis. The sensory matrix captured is dependent on the method, including coordinates (*e.g.,* Projective Mapping), frequency (*e.g.,* sorting, CATA), and correlation matrices (*e.g.,* RATA and DA) (Valentin et al. 2012). For methods that have two or more tasks, such as a sorting experiment with an additional verbal task, the data can be captured with two different matrices; sorting data can be captured as a co-occurrence matrix of samples as well as a correlation matrix of samples *vs* attributes (Valentin et al. 2012). For Projective Mapping with Ultra Flash Profile, the data is captured as (*x,y*) coordinates for the position of the samples on the map, and frequency of citation for the sample description (Garrido-Bañuelos et al. 2020; Hayward et al. 2020). The implication is that the matrices are then modeled differently based on the different types of matrices captured.

Unsupervised modelling

Unsupervised models are used to investigate inherent trends in the data without imposing any restrictions. These models mainly look for trends based on correlation or covariance, from which groupings can be found based on similarities or differences between samples. Unsupervised models are used for general exploration, pre-processing, or as a preceding step to supervised modeling or data fusion (Gagolewski 2012; Vera et al. 2011; Lahat, Adali, and Jutten 2015; Cocchi 2019a).

Since most chemical analysis output correlation matrices, the most common unsupervised MVA tool used in enology is principal component analysis (PCA), often accompanied by hierarchical cluster analysis (HCA). Correspondence analysis (CA) and multiple correspondence analysis (MCA) are generalized PCA used for categorical/frequency data where many counts of zero are present (Johs and Johs 2018; Valentin et al. 2012; McKillup 2005). Other common unsupervised data modeling tools used in enology include multidimensional scaling (MDS), multiple factor analysis (MFA), that can also be accompanied by HCA (Dien and Pagès 2003; Pagès 2004; Kruskal 1977; Salkind 2007). PCA is commonly used for chemistry data because of the matrix compatibility, whereas due to the types of matrices in sensory science, the other modeling tools mentioned are more appropriate (Valentin et al. 2012).

Supervised modelling

Supervised models are used for classification, discrimination, or prediction; these models are based on a measurable trend/regression which distinguishes one set of samples or variables from another (Sohail and Arif 2020). The models then find the best-fit function (regression) which represents the trend. Classification models look at similarities within a group based on the relationships between variables. Discrimination models look at the differences between the regressions of each class. Both types of models have been used in enology to classify samples according to qualitative aspects such as regionality, cultivar, and wine styles among others (Cuadros-Inostroza et al. 2010; Makris, Kallithraka, and Mamalos 2006; Edelmann et al. 2001).

Prediction models are similar but look at groups of variables instead of sample sets; all the samples should ideally have a similar variable correlation to the overall regression. These models have a calibration, validation, and prediction stage. A set of samples is used as a calibration set to build a regression which is representative of the common relationship between all variables. Another group of new or existing samples is used to validate or cross-validate the calibration model. There are different ways to validate the calibration model (Petrovic, Luis Aleixandre-Tudo, et al. 2019; Moyano et al. 2019; Jansen et al. 2013). The prediction set contains new observations (*unknown samples*) for which its membership to one of the calibrated classes can be predicted. Prediction models can also use the calibration set to predict an index such as predicting total antioxidant capacity (TAC) (Versari et al. 2010) or yeast assimilable nitrogen (YAN) (Petrovic, Luis Aleixandre-Tudo, et al. 2019) using untargeted infrared spectra. The variables (*e.g.*, spectral data) used in the calibration set have an already known correlation for which an index (*e.g.*, TAC, YAN) can be calculated. The calibration is then validated and used for the prediction of the index of an unknown sample.

Most supervised modeling in enology have used least squares for classification (Borràs et al., 2016; Silvestri et al. 2014; Vera et al. 2011) and discrimination (Vera et al. 2011). Some prediction models in enology have attempted to predict a set of sensory variables using chemical variables, with minimal success. The rationale here is that the sensory perception is caused by the presence of certain compounds, such as aroma derived from volatile compounds and thus a correlation can be calculated between the two types of data. The difficulty lies in that sensory analysis is holistic while the chemical analysis was based on samples that were altered through the sample preparation stage. Important interactions in the wine matrix are thus removed. Some attempts have then moved toward noninvasive sample preparations, untargeted chemical analysis, and data fusion strategies for coupling and ultimately

predicting sensory perception from chemistry data (Seisonen, Vene, and Koppel 2016; Brand, Panzeri, and Buica 2020). Additionally, to address this shortcoming, studies have advocated for the use of advanced techniques such as artificial intelligence and machine learning (Seisonen, Vene, and Koppel 2016).

Performance parameters and model optimization

All models generated through unsupervised and supervised techniques can be evaluated using various performance parameters (model diagnostics) that are based on the size, distribution, and purpose of the model (Härdle and Simar 2015; Salkind and Kristi). This section will address the parameters most often reported in enological applications.

Although specific for every model, performance parameters include measurements of the model fit (*e.g.,* regression coefficient, R^2 and root mean square of error in calibration, RMSEC), prediction power (*e.g.,* Q^2, root mean square of deviation/prediction – RMSD/P and validation RMSV), outliers (*e.g.,* distance to model in X variables, DmodX, and misclassification tables), and residuals (Eriksson et al. 2006; Wheeloc; McKillup 2005; Salkind and Kristi; Härdle and Simar 2015).

Many of the performance parameters related to the model fit are calculated from the stress of the model, for example the eigenvalue used for analysis such as MFA and PCA, and Kruskal's stress used for MDS (Robinson et al. 2014; Kruskal 1977; McKillup 2005; Salkind and Kristi; Härdle and Simar 2015). The stress is a relative measure of the total explained variation in the model (McKillup 2005; Salkind and Kristi). The distribution of the stress across the several dimensions (*e.g.,* principal components for PCA, dimensions for MDS and CA, and factors for MFA and GFA) that the model is fitted over, is a relative measure of the efficiency of the model.

The efficiency of a model is often described using a scree plot. The scree plot describes the decay of the stress and the cumulative explained variance (McKillup 2005). This efficiency is often expressed as the cumulative percentage explained variance (%EV) (McKillup 2005). The %EV is the most communicated performance parameter for multivariate analyses such as PCA, CA, and MFA in enology (Valentin et al. 2012; Alañón, Pérez-Coello, and Marina 2015; Valente et al. 2018). The %EV is mostly used for unsupervised techniques, supervised techniques tend to report the goodness-of-fit for calibration (using R^2 and RMSC), validation (RMSEV), and prediction (RMSP) using other performance parameters.

Studies mostly use the first two dimensions to evaluate performance since they contain the highest %EV. Chemistry data models generally contain high %EV for the first two dimensions but sensory data usually contain less, depending on the sensory method. For example, DA and RATA have %EV similar to chemistry data because, similar to chemistry, their data is based on intensity (Valentin et al. 2012; Brand 2019). Other sensory methods such as sorting, Pivot© Profile, and Projective Mapping have lower %EV because the data is not intensity- but rather frequency-based or ordinal (Valentin et al. 2012; Brand 2019).

Targeted chemical data generally has lower %EV in the first few dimensions compared to untargeted analysis. Targeted data analyses have a lower number of variables than untargeted data. Increasing the number of variables generally results in increased %EV for the first few dimensions (McKillup 2005). Although, since untargeted analyses can also include a significant amount of noise captured, data that is not pre-processed can have a low %EV compared to processed data (Rinnan, Berg, and Engelsen 2009). Additionally, the inclusion or exclusion of certain variables can result in a change in efficiency of the model (*i.e.,* increase or decrease in the %EV) (McKillup 2005). Adding variables of different sources or which measure different stimuli increases the stress in a model resulting in a broader distribution of the stress over the dimensions and thus lowering the %EV over the first dimensions (McKillup 2005). This is often observed in data fusion and multi-modal strategies (Lahat, Adali, and Jutten 2015; Borràs et al. 2015). When the %EV for the first two dimensions is low, the efficiency of the model can be communicated by looking at the first three dimensions (Parr et al. 2015), narrating the distribution of the %EV throughout the entire model, and/or by calculating the steepness of the slope in the scree plot (Mafata et al. 2020). The variables' contribution to the %EV of each dimension can be seen in the contributions table, sometimes presented also as a bar graph output in multivariate analysis toolkits. If variables' values remain relatively unchanged throughout an experiment, these variables will not greatly influence the %EV and will often lie close to the zero-point intersection (origin) of the Cartesian plots (McKillup 2005).

Cluster analyses calculate groupings based on similarity or dissimilarity, which can be done in an agglomerative or hierarchical manner. The distance similarity matrix is calculated based on the proximity/distance of samples. The coefficients of these distances are calculated based on a variety of algorithms; for example, they can be based on weighted distance for unfitted data or given by the Euclidean distance in fitted data (Härdle and Simar 2015). Due to the complexities of clustering unfitted (raw data), most studies use MVA to fit the data and then apply cluster analysis to similarity/distance matrices derived from them (Ivanisevic et al. 2015; Naumann et al. 2007; Kruskal 1977). These cluster analyses are derived from parametric algorithms for normal distribution and compute an average around which to cluster samples. These averages can be computed in various ways based on different types of linkages, *e.g.*, centroid, complete, or single linkage (Nordhaug Myhre et al. 2018; Härdle and Simar 2015). An assumption of similarity between samples can lead to using similarity methods where a convergent algorithm is applied (agglomerative). A research question based on an assumption of dissimilarity/discrimination may, for instance, use divergent strategy such as centroid linkage HCA. These cases tend to be open-ended and result in hypothesis formation, making them popular for incorporation in non-parametric cluster analyses

(Radovanovic et al. 2016; Nordhaug Myhre et al. 2018; Edelmann et al. 2001).

Model optimization generally uses performance parameters as indicators for increasing the goodness-of-fit and performance. Improving the performance requires the use of latent variables. Variable contributions can be used to improve the efficiency (%EV) and variable weights can be used to improve sample clustering, both these and other parameters can be used for variable selection in the pre-processing stage (Wheeloc; Eriksson et al. 2006). Effective use of latent variables in pre-processing steps to improve model performance has been considered from an applications perspective (Iorgulescu et al. 2016) and a statistical method perspective (Jansen et al. 2013). Considering the previous data handling steps discussed in this review, ensuring improved model performance requires attention to detail from both perspectives (Gerretzen et al. 2015; Campos and Reis 2020).

Model optimization for unsupervised and supervised models can be done similarly from statistical and application perspectives (Iorgulescu et al. 2016). In the enology context, it is especially necessary to have both perspectives in mind when sensory evaluation is concerned. The number of samples that can be assessed by a specific method is often the limiting factor in Sensory (Valentin et al. 2012; Fleming, Ziegler, and Hayes 2015). Optimizing the data handling steps (collection, capture, and pre-processing) can sometimes be enough for optimization. A better way, though, is to start with a smart experimental design, since the experimental design can address the issues related to model optimization if the data handling options is considered from the beginning (Gerretzen et al. 2015; Yu, Low, and Zhou 2018). Based on principles of experimental design, model optimization requires looking at the number of samples, the variation in samples, and the variables measured.

Multivariate models (supervised and unsupervised) generally require the number of independent variables to be more than the number of samples, since the model is based on the correlations/covariance in the variables (McKillup 2005). Similar for supervised models, the calibration set (independent and/or dependent variables) and the validation set must have more variables than samples to optimize the calibration and validation (Jansen et al. 2013).

Model optimization from an application perspective is also important. Although more measurements (variables) can result in the optimization of the calibration by increasing variation, the nature of the relationship between variables is more important since it creates variability. Variation in the samples selected must be representative, when extrapolating results for the prediction of unknowns beyond a case study. A pre-modeling optimization which requires variable selection can be done in supervised modeling strategies based on the application or iterative statistical assessment of the model performance parameters. The mathematical and statistical aspects concerning supervised model optimization and pre-processing have been previously published (Lahat, Adali, and Jutten 2015; Rinnan, Berg, and Engelsen 2009; Jansen et al. 2013). Supervised models are more often optimized compared to unsupervised; this goes hand in hand with more applications using supervised than unsupervised modeling.

These principles for optimization applied in enology include variable selection, feature selection and engineering as pre-processing techniques coupled to supervised strategies such as PLS (Larsen, Van Den Berg, and Engelsen 2006; Guld et al. 2020; Seisonen, Vene, and Koppel 2016; Pereira et al. 2016; Petrovic, Kidd, et al. 2019). Variable selection has been used for choosing certain wavenumbers in IR modeling *a priori* (before the modeling based on the theoretical knowledge that the analytes of interest give a signal in a certain region) but also *a posteriori* (based on variable contributions to the classification of samples) (Genisheva et al. 2018). Feature selection has been done on similar data using IR, NMR, and UV-Vis for the selection of principal components (Pereira et al. 2016; Borràs et al. 2015) and/or the use of latent variables for optimizing untargeted spectral data (Godelmann et al. 2013; Brand, Panzeri, and Buica 2020; Cuadros-Inostroza et al. 2010).

The impact/success of these optimization strategies is assessed statistically by looking at the improvement of the performance parameters (*e.g.*, higher %EV, lower RMSEC/RMSD) and descriptively by looking at desirable sample clustering. The process is reiterative and may arrive at a point where the model can no longer be optimized, or the performance becomes compromised. It is at such a point that issues of overfitting can arise. It is then recommended to use at least two different types of parameters to track for this (*e.g.*, %EV for better fit and regression vector/RV coefficients for clustering).

Model output, visual aids, and interpretation

Multivariate data can be difficult to interpret; it is thus important to use both statistical and contextual interpretation: contextual interpretation in the form of background and domain knowledge of the application and experimentation, and statistical evaluation in the form of model performance and evaluation parameters. The statistical aspect is technical, and its significance must be interpreted not just using performance parameters, but also with the experimental context in mind. The use of visual aids provides a transition between the statistical and the contextual interpretation.

Accompanying every model are sets of tables containing performance parameters and latent variables that are specific for the type of model used (supervised or unsupervised, similarity or dissimilarity, correlation or covariance, *etc.*) (McKillup 2005). The latent variables are presented in tables of figures that show the relationship between variables, samples and/or both. These latent variables include ordinal model data, variable contributions, and variable weights among others (Eriksson et al. 2006; McKillup 2005; Wheeloc). From the fitted model, the coordinates are calculated for each dimension and then the contributions and weights are calculated (McKillup 2005).

Ordinal data is usually represented in two-dimensional Cartesian plot intersecting the first and second

dimensions with the highest explained variance; in certain cases (*e.g.,* when the %EV is not high) the third dimension is also explored. A Cartesian plot of either samples (*e.g.,* scores in PCA, individual factors in MFA) or the variables (*e.g.*, loadings in PCA, group factors in MFA) or a projection of the two (biplot) can be used for interpretation easier than the original tabulated data (McKillup 2005; Eriksson et al. 2006; Wheeloc). In enological studies, the first two dimensions are usually sufficient for visualizing the trends in chemistry data. Sensory data sets that contain lower %EV in the first two dimensions require greater probing beyond the first two dimensions. Studies have thus shown ingenuity by expressing the distribution of the %EV across all dimensions and using the first three dimensions in either multiple 2 D projections (Mafata et al. 2020) or as a 3 D graph (Ballester et al. 2005). This approach minimizes chances of misinterpretation of descriptive data models. An opportunity for misinterpretation of Cartesian plots can arise when using secondary identifiers, creating false visual impressions of associations/groupings among samples without running a cluster analysis. To overcome this, Cartesian plots are coupled with confidence ellipses, cluster analysis, and regression vector (RV) coefficients (Auf Der Heyde 1990; Radovanovic et al. 2016). Confidence ellipses can be imposed onto the projections to infer grouping of samples. This is based on analysis of variance (ANOVA) where the mean of certain repeats is common among samples, clustering them together (Pagés and Husson 2005). Confidence ellipses are applied on the Cartesian plot based on the distance to the model (*e.g.,* using Hotelling or bootstrapping), usually set at 95% standard deviation from the mean (Härdle and Simar 2015). Since repeats are not always possible, confidence ellipses often overfit the data depending on the variation between samples, this is especially the case for sensory data (Pagés and Husson 2005; Brand 2019).

Cluster analysis can be applied to the Cartesian plots, containing as many dimensions as needed for pattern recognition, visualized using a dendrogram. A table of co-occurrence latent values such as sample correlation matrix and RV coefficients can be calculated between samples, variables or data blocks in multiblock analyses (Abdi 2007; Kruskal 1977; Salkind and Kristin 2007). These matrices can be visualized as the Cartesian plots, a dendrogram for scores and blocks or using heatmaps for larger data such as loadings. Heatmaps have been mostly used in metabolomics (Ivanisevic et al. 2015). Unlike the Cartesian plots, heatmaps often include projections of dendrograms of scores and/or loadings. This means that, without bias, the clusters can be visualized for a sample set and simultaneously, the differences between variables across the samples. Heatmaps have been coupled with sensory methods for looking at the differences in sensory attributes across samples (Mafata et al. 2019; Brand, Panzeri, and Buica 2020). Other measurements of goodness-of-fit include distance to model (DModX), misclassification, and residuals which can be graphed to probe deeper into the model performance parameters (Eriksson et al. 2006; Wheeloc).

In sensory, when interpreting model output, it should be considered that experiments can result in the acquisition of primary and secondary data corresponding to primary and secondary tasks (Section 2.3.1). Primary data should be directly linked to the experimental/research question (hypothesis). Secondary data may be in the form of (tentative) annotations and often provides qualitative support to the main data. These data are often used as reasons for pattern recognition outcomes and, although they are important, it is necessary to understand their nature so as not to make inferences of correlation or causality. For example, sorting and Projective Mapping have the grouping and distances between samples respectively as the primary tasks and may incorporate annotations in the form of attributes using listing or ultra-flash profiling (Mafata, Buica, du Toit, et al. 2018; Cariou and Qannari 2018; Valentin et al. 2012; Hayward et al. 2020).

The design of experiments in these cases prioritizes and optimizes the primary task (*i.e.,* sorting and mapping) which directly addresses the research question. The statistical implications are that the sample variation for the primary task is based on the co-occurrence or ordinal matrix of samples, whereas for the secondary task it is based on the variability of attributes. These complexities of sensory data have significant implications on the statistical *vs* contextual interpretation of modeling results. Even though the secondary task may contribute contextual information to the research question, its results cannot be substituted with the primary task just because the results are more satisfactory. Secondary task may be forming a new hypothesis or be better suited to answer the research question, in such a case a new experimental design can be used to optimize and prioritize the task. For example, studies looking to profile sample sensory attributes may need to use a full-factorial DOE whereas those seeking to distinguish samples may not (Yu, Low, and Zhou 2018). Additionally, the manner (*i.e.*, the intuitiveness/level of difficulty) and order of execution of the tasks may influence the success of the modeling (Valentin et al. 2012; Brand 2019). It can happen that the judges are better at executing the secondary task, in which case the contextual interpretation of the results must take this into account.

Data fusion and advanced data modelling in enology

The most recent trends in data modeling for enology are toward the use of artificial intelligence (AI) (Garrido-Delgado et al. 2011; Valente et al. 2018) but there is an intermediary approach, which is data fusion. Data fusion is the combining of data sets from different sources into comprehensive and representative data models (Cocchi 2019a). Data fusion approaches can use algorithms from both classical multivariate modeling and AI at different levels of complexity using either supervised or unsupervised techniques (Cocchi 2019b).

Different data sets have different distributions and scale; they cannot always be simply combined. When data sets of

different distributions (variable scale and distribution) are modeled together in a simple concatenation, the results are skewed in such a way that it gives a false representation of the correlations between variables/samples. Hence, principles of data fusion must be used to properly integrate the data sets.

Data fusion frameworks

Data fusion is classified under low, medium, and high level according to increasing levels of complexity (Lahat, Adali, and Jutten 2015; Cocchi 2019a; Borràs et al. 2015), taking both statistical (Cocchi 2019a) and strategic approach (Lahat, Adali, and Jutten 2015). Enological data fusion strategies used for these levels have been reviewed by Borràs et al. (2015) in the context of food and beverages authentication.

The simplest form of data fusion, low-level, is heavily reliant on the prerequisite of matrix compatibility between different data sets (Cocchi 2019b). It is for this reason that it is often not called data fusion but rather data aggregation or concatenation (Borràs et al. 2015; Cocchi 2019b). The implications of data concatenation are that the data sets are dependent and vary similarly in scale and distribution (Härdle and Simar 2015). In enology, low-level data fusion is commonly done on targeted measurements but keeps the chemistry and sensory sets separate. For example, most low-level data fusion done on wine uses instrumental data and sensors as a proxy for sensory evaluation (Borràs et al. 2015; Seisonen, Vene, and Koppel 2016). Concatenation is more common for chemistry data sets since they are of the same type (correlation matrices) and can be scaled using simple methods such as unit conversion.

In sensory, overcoming matrix compatibility issues requires more sophisticated solutions than simple conversions; that is why fusion of sensory data is often done through mid-level or high-level data fusion strategies (Boccard and Rutledge 2014). Studies that have attempted to do simple concatenation of sensory and chemistry data used techniques such as PLS, which keep the chemistry set as an independent variables and sensory set as dependent variable set (Hopfer, Ebeler, and Heymann 2012; Seisonen, Vene, and Koppel 2016). One study has also attempted to use descriptive analysis profile of wine to predict typicality with good success (Coulon-Leroy et al. 2018). The low-level approaches that did not do simple concatenation were limited for reasons such as incompatible matrix types between data sets, and differences in variable distributions (discreet *vs* continuous) and matrix arrays (*e.g.,* 2D *vs* 3D); these are cases when the preceding steps in data handling (Section 2.3) must be re-assessed.

Mid-level data fusion involves the use of pre-processing and multiblock approaches to ensure matrix compatibility (de Juan et al. 2019; Cocchi 2019b; Borràs et al. 2015). Matrix compatibility, previously mentioned as a limitation to achieving low-level data fusion, is obtained through multiblock techniques such as factor analyses (MFA, GPA, PARAFAC, *etc.*) (Niimi, Boss, et al. 2018; Bro 1986; Silvestri et al. 2014). Pre-processing for matrix compatibility also includes mathematical transformations (rating converted to frequency data) and the use of exploratory modeling for scaling (Campos and Reis 2020; Jansen et al. 2013; Rinnan, Berg, and Engelsen 2009). Although supervised data fusion approaches are more common in enology, unsupervised approaches are gaining popularity. Since multiblock approaches (*e.g.,* MFA) have become commonplace for treatment of sensory data (*e.g.,* Projective Mapping), opportunities have risen where they are used for data fusion of multiple data sets. For example, MFA has been used for the fusion of chemical and sensory data related to volatile phenol compounds and smoke-related sensory descriptors (McKay et al. 2019) as well as furanmethanethiol (FMT) and coffee aroma in Pinotage wines (Garrido-Bañuelos and Buica 2020). Supervised mid-level data fusion approaches have been of relevance to enology due to increased use of untargeted analysis. Variations of partial least square (PLS) have been used on data such as UV-Vis, IR, GC-MS, NMR, and to predict sensory descriptors and/or sensory classes such style, cultivar or regionality (Cayuela, Puertas, and Cantos-Villar 2017; Cozzolino et al. 2005; Culbert et al. 2015; Fudge et al. 2013; Gambetta et al. 2019).

High-level data fusion involves extensive pre-processing, dynamic use of techniques from parametric (classical statistics) to advanced techniques (non-parametric), and mixed multiblock approaches that usually involve big data (Cocchi 2019a; Borràs et al. 2015). Also called decision-level data fusion, these approaches maximize informational value, precision, and accuracy (Borràs et al. 2015; Cocchi 2019b). The strategies generally require elements of both quantitative measures of variation (large sample size, biological, and/or instrumental repeats) and qualitative measures of variability (various equipment/types of measurements, sample variability in the form of representation within and outside the calibration ranges) (Petrovic, Luis Aleixandre-Tudo, et al. 2019). This means that model performance and optimization are very important aspects in these strategies. In enology, modeling mostly uses supervised methods of prediction and classification. Combinations of chemical data sets are used to create robust calibration models to predict wine-related concepts such as cultivar, designation of origin, and authenticity (Alañón, Pérez-Coello, and Marina 2015; Borràs et al. 2015). These high-level strategies involve process technology for acquisition, monitoring, and modeling process outcomes (Ríos-Reina et al. 2020; Cocchi 2019b; Borràs et al. 2015). Examples include the use of infrared spectroscopy for accurate predictions of enological parameters such as yeast assimilable nitrogen (YAN) (Petrovic, Luis Aleixandre-Tudo, et al. 2019) and total antioxidant capacity (TAC) (Versari et al. 2010). Although process analytical technology (PAT) strategies are not always considered data fusion, they integrate multiple measurements from different sources modes for prediction purposes (Cavaglia et al. 2020; Fourie et al. 2020; Borràs et al. 2015; Alañón, Pérez-Coello, and Marina 2015).

Even though the high-level data fusion strategies presented in the literature are generally hypothesis testing,

due to the large data variation and variability, prospects of data exploration could lead to hypothesis formation. This is an approach worth considering for future enological applications. This is especially true for cases that have used advanced modeling techniques for data mining and pattern recognition, which are presented in the next section.

In practice, the theoretical frameworks presented here are not always easy to distinguish. There are no hard borders between each level, and there may be some overlap. Since studies usually disclose the results of successful modeling strategies, the full process to the approach, which may contain elements of other levels of data fusion, is not always communicated. This can create misconceptions about the level of difficulty in fuzing multimodal data, which can be especially misleading when dealing with sensory data. Omitting intermediary steps of pre-processing creates gaps which are important for understanding the overall strategy and rationale behind choosing modeling types. There are so many modeling options that are available and interchangeable. Applications from a purely statistical approach can simply be based on the methodology but because the applied sciences need to address the contextual interpretation, communicating the rationale behind the approach is very beneficial for progression in the field.

Advanced data handling techniques

Advancements in data handling are motivated by the need to improve mathematical/statistical algorithms to better model performance and developing analytical algorithms for more user-friendly software. Advancements of algorithms can be based on classical statistics or AI systems. Using classical statistics, supervised modeling advancements have worked toward increasing the calibration and discriminative power of models (McKillup 2005; Eriksson et al. 2006; Härdle and Simar 2015). Both supervised and unsupervised modeling are advancing toward the use of nonparametric (non-classical) artificial intelligence techniques. These techniques have mostly been used to further pattern recognition in the form of clustering and classification, within the context of food analysis (Carvalho Rocha, Prado, and Blonder 2020).

Classical multivariate analyses derive linear relationships and linear regression algorithms based on normal parametric distribution (McKillup 2005; Härdle and Simar 2015). Although some advances in mathematical algorithms have been developed to improve on these methods, their limitations in solving complex applied science research questions cannot be overcome so simplistically, especially given the increase in data size and in variations. Since large data size and variability is a prerequisite for running AI analyses, AI as an approach is intuitively better suited for analyzing big data. Artificial intelligence is more nuanced in that it accommodates non-binary (*i.e.,* classifications) and non-linear (*i.e.,* calibrations) relationships (Carvalho Rocha, Prado, and Blonder 2020). This AI approach is especially motivating for work on complex natural products such as wine and is compatible with the nuances of sensory data, an avenue that has yet to be exploited. Additionally, AI can solve issues related to overfitting and model performance in classical MVA (de Andrade et al. 2020). In the wider field of food sciences, a recent review has also indicated a great advantage of coupling classical MVA with AI (Carvalho Rocha, Prado, and Blonder 2020). The review narrated some important behavioral barriers to the use of advanced techniques in food analysis and exemplified their use in food science, with only five of the 128 cases being wine related. With varying degrees of success, the review found that the AI approaches were better adapted for mapping the behavior of complex products and thus obtained models with better performance compared to classical MVA.

It is not just necessary to increase the discrimination power (classification, grouping, or prediction) of data analysis, it is also crucial ultimately to understand what drives/contributes to the observed patterns. Taking a non-classical approach to pattern recognition can result in extracting/obtaining greater information from the data (*e.g.,* compounds or sensory attributes). The strategy behind the use of advanced techniques is analogous to how mid and high-level data fusion uses low-level modeling as pre-processing steps. The strategy has been to use classical MVA followed by AI analysis for pattern recognition, *i.e.* the data is first normalized using classical MVA and then AI is applied (Nordhaug Myhre et al. 2018; Härdle and Simar 2015). In a proof of concept for the potential of non-parametric techniques, a few case studies have been documented for the successful use of artificial neural networks for mining unstructured/raw data (Nordhaug Myhre et al. 2018).

The most common documented uses of AI in enology include support vector machines (SVM), self-organizing maps (SOM), and k-nearest neighbors (k-NN) or k-means clustering (Carvalho Rocha, Prado, and Blonder 2020). Both generally and in enology, these techniques have been used in a supervised manner for prediction or classification, with either supervised or unsupervised classical MVA used for exploratory preceding steps (Carvalho Rocha, Prado, and Blonder 2020). In enology, SVM and k-NN have been coupled with other classical MVA supervised techniques such as PLS to increase model performance. With varying degrees of success, they had better performance compared to classical MVA (Gómez-Meire et al. 2014; Latorre et al. 1994; Borràs et al. 2015). SOM have previously been used for exploratory data mining of unstructured sensory data using Classification and Regression Trees (CART) coupled with CA to differentiate South African white wines styles; the study was successful in demonstrating mining of such data using advanced modeling techniques (Valente et al. 2018). Although these methods are theoretically and practically more complex compared to classical MVA, the examples and case studies presented have shown their potential in bettering data handling for enology. They could be capable of elucidating answers to big questions in enology such as sensory and chemistry markers of wine quality, as well as wine authenticity.

Conclusions

The aim of this review was to examine the different stages of the data handling process in Enology and elucidate the rules and rationale behind the decisions made. It specifically focused on the differences and similarities between the chemometric and sensometric treatments of the data. As well as addressing some misconceptions concerning data handling in Enology, this review identified the key decision-making aspects during the data input stage (capturing and pre-processing), the modeling, and the model output (visualization/interpretation). In terms of the success of a model in addressing the research question/hypothesis, what you put in is what you get out.[1] Hence, thorough data capturing chances of success increase since only that which was captured can be modeled. The pre-processing of the data was shown to impact on the performance of models as measured by the performance parameters. That is to say that the level of redundancies and "noise" in a model will be reflected in poor performance parameters such as the explained variance and calibration coefficients. Thus, as a reiterative process, model optimization techniques such as variable/feature selection and the choice of these were addressed. This review most importantly discussed the impactful nature of visual aids and offered rationale as to how to couple visual aids with each other and with performance parameters to enhance the interpretability of model outcomes. Furthermore, in this regard, the review rationalized the intertwining of statistical and applied reasoning for interpretation of modeling outcomes. The standing recommendation has thus been to have a design of experiments that is considerate of the stages of data handling and their impact on achieving the research question. The advantage of such a holistic approach is that it not only increases chances of successful hypothesis testing, but it can create opportunities for hypothesis forming scenarios. This would then encourage the advancement of data analysis in Enology toward techniques in Artificial Intelligence. Applying advanced data analyses is very much possible given that there are means (instrumental and software availability), motivation (optimizing model performance and applied interpretations), and opportunity (large data already available). It is important to communicate the strategies since this has critical contribution to the philosophy and progression of science and research.

Note

1. "Garbage in, garbage out: Used to express the idea that in computing and other spheres, incorrect or poor-quality input will always produce faulty output (often abbreviated as GIGO)." www.oxfordreference.com

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Research Foundation of South Africa.

References

Abdi, H. 2007. RV coefficient and congruence coefficient. In *Encyclopedia of measurement and statistics*, 849-853. Thousand Oaks, CA: Sage Publications, Inc.

Alañón, M. E., M. S. Pérez-Coello, and M. L. Marina. 2015. Wine science in the metabolomics era. *TrAC Trends in Analytical Chemistry* 74:1–20. doi: 10.1016/j.trac.2015.05.006.

Andrade, B. M. d., J. S. de Gois, V. L. Xavier, and A. S. Luna. 2020. Comparison of the performance of multiclass classifiers in chemical data: Addressing the problem of overfitting with the permutation test. *Chemometrics and Intelligent Laboratory Systems* 201:104013. doi: 10.1016/j.chemolab.2020.104013.

Ares, G., R. Deliza, C. Barreiro, A. Giménez, and A. Gámbaro. 2010. Comparison of two sensory profiling techniques based on consumer perception. *Food Quality and Preference* 21 (4):417–26. doi: 10.1016/j.foodqual.2009.10.006.

Auf Der Heyde, T. P. E. 1990. Analyzing chemical data in more than two dimensions: A tutorial on factor and cluster analysis. *Journal of Chemical Education* 67 (6):461. doi: 10.1021/ed067p461.

Ballester, J., B. Patris, R. Symoneaux, and D. Valentin. 2008. Conceptual vs. perceptual wine spaces: Does expertise matter? *Food Quality and Preference* 19 (3):267–76. doi: 10.1016/j.foodqual.2007.08.001.

Ballester, J., C. Dacremont, Y. L. Fur, and P. Etievant. 2005. The role of olfaction in the elaboration and use of the chardonnay wine concept. *Food Quality and Preference* 16 (4):351–9. doi: 10.1016/j.foodqual.2004.06.001.

Ballester, J., M. Mihnea, D. Peyron, and D. Valentin. 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):140–52. doi: 10.1111/ajgw.12024.

Boccard, J., and D. N. Rutledge. 2014. Iterative weighting of multiblock data in the orthogonal partial least squares framework. *Analytica Chimica Acta* 813:25–34. doi: 10.1016/j.aca.2014.01.025.

Borràs, E., J. Ferré, R. Boqué, M. Mestres, L. Aceña, A. Calvo, and O. Busto. 2016. Olive oil sensory defects classification with data fusion of instrumental techniques and multivariate analysis (PLS-DA). *Food Chemistry* 203:314–22. doi:10.1016/j.foodchem.2016.02.038.

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

Brand, J. 2019. "Rapid Sensory Profiling Methods for Wine: Workflow Optimisation for Research and Industry Applications." PhD Thesis, Stellenbosch University.

Brand, J., V. Panzeri, and A. Buica. 2020. Wine quality drivers: A case study on South African Chenin Blanc and Pinotage wines. *Foods* 9 (6):805. doi: 10.3390/foods9060805.

Bro, R. 1986. Chemometrics and intelligent laboratory systems, 1: 115–120. doi: 10.1016/0169-7439(86)80032-6.

Campo, E., J. Ballester, J. Langlois, C. Dacremont, and D. Valentin. 2010. Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy pinot noir wines. *Food Quality and Preference* 21 (1):44–55. doi: 10.1016/j.foodqual.2009.08.001.

Campos, M. P., and M. S. Reis. 2020. Data preprocessing for multiblock modelling – A systematization with new methods. *Chemometrics and Intelligent Laboratory Systems* 199 (January):103959. doi: 10.1016/j.chemolab.2020.103959.

Cardello, A. V., O. Maller, J. G. Kapsalis, R. A. Segars, F. M. Sawyer, C. Murphy, and H. R. Moskowitz. 1982. Perception of texture by trained and consumer panelists. *Journal of Food Science* 47 (4):1186–97. doi: 10.1111/j.1365-2621.1982.tb07646.x.

Cariou, V., and E. M. Qannari. 2018. Statistical treatment of free sorting data by means of correspondence and cluster analyses. *Food Quality and Preference* 68:1–11. doi: 10.1016/j.foodqual.2018.01.011.

Cariou, V., E. M. Qannari, D. N. Rutledge, and E. Vigneau. 2018. ComDim: From multiblock data analysis to path modeling. *Food Quality and Preference* 67:27–34. doi: 10.1016/j.foodqual.2017.02.012.

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

Cavaglia, J., D. Schorn-García, B. Giussani, J. Ferré, O. Busto, L. Aceña, M. Mestres, and R. Boqué. 2020. Monitoring wine fermentation deviations using an ATR-MIR spectrometer and MSPC charts. *Chemometrics and Intelligent Laboratory Systems* 201:104011. doi: 10.1016/j.chemolab.2020.104011.

Cayuela, J. A., B. Puertas, and E. Cantos-Villar. 2017. Assessing wine sensory attributes using Vis/NIR. *European Food Research and Technology* 243 (6):941–53. doi: 10.1007/s00217-016-2807-9.

Chollet, S., D. Valentin, and H. Abdi. 2005. Do trained assessors generalize their knowledge to new stimuli? *Food Quality and Preference* 16 (1):13–23. doi: 10.1016/j.foodqual.2003.12.003.

Chrea, C., D. Valentin, C. Sulmont-Rossé, D. Hoang Nguyen, and H. Abdi. 2005. Semantic, typicality and odor representation: A cross-cultural study. *Chemical Senses* 30 (1):37–49. doi: 10.1093/chemse/bjh255.

Cocchi, M. 2019b. "Introduction." In *Data handling in science and technology*, vol. 31, 1–26. Amsterdam, The Netherlands: Elsevier Ltd. doi: 10.1016/B978-0-444-63984-4.00001-6.

Cocchi, M., ed. 2019a. *Data fusion methodology and applications data handling in science and technology*. Vol. 31. https://www.sciencedirect.com/bookseries/data-handling-in-science-and-technology/vol/31/suppl/C.

Coetzee, C., E. Van Wyngaard, K. Šuklje, A. C. Silva Ferreira, and W. J. du Toit. 2016. Chemical and sensory study on the evolution of aromatic and nonaromatic compounds during the progressive oxidative storage of a Sauvignon Blanc wine. *Journal of Agricultural and Food Chemistry* 64 (42):7979–93. doi: 10.1021/acs.jafc.6b02174.

Coulon-Leroy, C., N. Pouzalgues, L. Cayla, R. Symoneaux, and G. Masson. 2018. Is the typicality of 'Provence Rosé Wines' only a matter of color? *OENO One* 52 (4):1–15. doi: 10.20870/oeno-one.2018.52.4.2125.

Cozzolino, D., H. E. Smyth, K. A. Lattey, W. Cynkar, L. Janik, R. G. Dambergs, I. L. Francis, and M. Gishen. 2005. Relationship between sensory analysis and near infrared spectroscopy in Australian Riesling and Chardonnay wines. *Analytica Chimica Acta* 539 (1-2):341–8. doi: 10.1016/j.aca.2005.03.019.

Cuadros-Inostroza, A., P. Giavalisco, J. Hummel, A. Eckardt, L. Willmitzer, and H. Peña-Cortés. 2010. Discrimination of wine attributes by metabolome analysis. *Analytical Chemistry* 82 (9):3573–80. doi: 10.1021/ac902678t.

Culbert, J., D. Cozzolino, R. Ristic, and K. Wilkinson. 2015. Classification of sparkling wine style and quality by MIR spectroscopy. *Molecules (Basel, Switzerland)* 20 (5):8341–56. doi: 10.3390/molecules20058341.

Deneulin, P., and F. Bavaud. 2016. Analyses of open-ended questions by renormalized associativities and textual networks: A study of perception of minerality in wine. *Food Quality and Preference* 47 (January):34–44. doi: 10.1016/j.foodqual.2015.06.013.

Dien, S. L., and J. Pagès. 2003. Hierarchical multiple factor analysis: Application to the comparison of sensory profiles. *Food Quality and Preference* 14 (5–6):397–403. doi: 10.1016/S0950-3293(03)00027-2.

Du Toit, W. J, and C. Piquet. 2016. Research Note: Effect of Simulated Shipping Temperatures on the Sensory Composition of South African Chenin Blanc and Sauvignon Blanc Wines. *South African Journal of Enology and Viticulture* 35 (2). doi:10.21548/35-2-1016.

Edelmann, A., J. Diewok, K. C. Schuster, and B. Lendl. 2001. Rapid method for the discrimination of red wine cultivars based on mid-infrared spectroscopy of phenolic wine extracts. *Journal of Agricultural and Food Chemistry* 49 (3):1139–45. doi: 10.1021/jf001196p.

Eriksson, L., E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikstrom, and S. Wold. 2006. Multivariate and megavariate data analysis basic principles and applications (Part I). Umea, Sweden: Umetrics.

Faye, P., P. Courcoux, A. Giboreau, and E. M. Qannari. 2013. Assessing and taking into account the subjects' experience and knowledge in consumer studies. Application to the free sorting of wine glasses. *Food Quality and Preference* 28 (1):317–27. doi: 10.1016/j.foodqual.2012.09.001.

Ferreira, S. L. C. 2019. Chemometrics and statistics | experimental design. In *Encyclopedia of analytical science*, 420. Elsevier. doi: 10.1016/B978-0-12-409547-2.14536-6.

Fleming, E. E., G. R. Ziegler, and J. E. Hayes. 2015. Check-All-That-Apply (CATA), sorting, and polarized sensory positioning (PSP) with Astringent stimuli. *Food Quality and Preference* 45:41–9. doi: 10.1016/j.foodqual.2015.05.004.

Fourie, E., J. L. Aleixandre-Tudo, M. Mihnea, and W. du Toit. 2020. Partial least squares calibrations and batch statistical process control to monitor phenolic extraction in red wine fermentations under different maceration conditions. *Food Control* 115 (March):107303. doi: 10.1016/j.foodcont.2020.107303.

Fudge, A. L., K. L. Wilkinson, R. Ristic, and D. Cozzolino. 2013. Synchronous two-dimensional MIR correlation spectroscopy (2D-COS) as a novel method for screening smoke tainted wine. *Food Chemistry* 139 (1–4):115–9. doi: 10.1016/j.foodchem.2013.01.090.

Gagolewski, M. 2012. *Data fusion*, In *Proceedings to 2012 5th international symposium on resilient control systems*, ed. O. Hryniewicz, J. Mielniczuk, W. Penczek, and J. Waniewski, 69–70. doi: 10.1109/isrcs.2012.6309295.

Gambetta, J. M., D. Cozzolino, S. E. Bastian, and D. W. Jeffery. 2019. Classification of chardonnay grapes according to geographical indication and quality grade using attenuated total reflectance mid-infrared spectroscopy. *Food Analytical Methods* 12 (1):239–45. doi: 10.1007/s12161-018-1355-2.

Garrido-Bañuelos, G., and A. Buica. 2020. Is there a link between coffee aroma and the level of furanmethanethiol (FMT) in pinotage wines. *South African Journal of Enology and Viticulture* 41 (2):245–50. doi: 10.21548/41-2-4224.

Garrido-Bañuelos, G., V. Panzeri, J. Brand, and A. Buica. 2020. Evaluation of sensory effects of thiols in red wines by projective mapping using multifactorial analysis and correspondence analysis. *Journal of Sensory Studies* 35 (4): e12576. doi: 10.1111/joss.12576.

Garrido-Delgado, R., L. Arce, A. V. Guamán, A. Pardo, S. Marco, and M. Valcárcel. 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): 471–79. doi: 10.1016/j.talanta.2011.01.044.

Gawel, R., A. Oberholster, and I. L. Francis. 2000. A 'mouth-feel wheel': Terminology for communicating the mouth-feel characteristics of red wine. *Australian Journal of Grape and Wine Research* 6 (3):203–7. doi: 10.1111/j.1755-0238.2000.tb00180.x.

Genisheva, Z., C. Quintelas, D. P. Mesquita, E. C. Ferreira, J. M. Oliveira, and A. L. Amaral. 2018. New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR). *Food Chemistry* 246 (April):172–8. doi: 10.1016/j.foodchem.2017.11.015.

Gerretzen, J., E. Szymańska, J. J. Jansen, J. Bart, H.-J. van Manen, E. R. van den Heuvel, and L. M. C. Buydens. 2015. Simple and effective way for data preprocessing selection based on design of experiments. *Analytical Chemistry* 87 (24):12096–103. doi: 10.1021/acs.analchem.5b02832.

Godelmann, R., F. Fang, E. Humpfer, B. Schütz, M. Bansbach, H. Schäfer, and M. Spraul. 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): 5610–19. doi: 10.1021/jf400800d.

Gómez-Meire, S., C. Campos, E. Falqué, F. Díaz, and F. Fdez-Riverola. 2014. Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International* 60:230–40. doi: 10.1016/j.foodres.2013.09.032.

Granato, D., and G. Ares. 2013. *Mathematical and statistical methods in food science and technology*, ed. D. Granato and G. Ares. Chichester, United Kingdom: John Wiley & Sons, Ltd. doi: 10.1002/9781118434635.

Granato, D., V. Ônica Maria de Araújo Calado, and B. Jarvis. 2014. Observations on the use of statistical methods in food science and

technology. *Food Research International* 55:137–49. doi: 10.1016/j.foodres.2013.10.024.
Guld, Z., D. Nyitrainé Sárdy, A. Gere, and A. Rácz. 2020. Comparison of sensory evaluation techniques for Hungarian wines. *Journal of Chemometrics* 34 (4): e3219. doi: 10.1002/cem.3219.
Härdle, W. K, and L. Simar. 2015. *Applied multivariate statistical analysis*. 4th ed. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-45171-7.
Hayward, L., H. Jantzi, A. Smith, and M. B. McSweeney. 2020. How do consumers describe cool climate wines using projective mapping and ultra-flash profile? *Food Quality and Preference* 86 (December):104026. doi: 10.1016/j.foodqual.2020.104026.
Hopfer, H., S. E. Ebeler, and H. Heymann. 2012. The combined effects of storage temperature and packaging type on the sensory and chemical properties of Chardonnay. *Journal of Agricultural and Food Chemistry* 60 (43):10743–54. doi: 10.1021/jf302910f.
Hunter, E., G. Anthony, B. Dijksterhuis, E. Mostafa Qannari, and H. J. MacFie. 1995. Second sensometrics meeting - Edinburgh, 16-18 September 1994: Introduction on behalf of the organising committee. *Food Quality and Preference* 6 (4):215–6. doi: 10.1016/0950-3293(96)80774-9.
Iorgulescu, E., V. A. Voicu, C. Sârbu, F. Tache, F. Albu, and A. Medvedovici. 2016. 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 extracts. *Talanta* 155:133–44. doi: 10.1016/j.talanta.2016.04.042.
Ivanisevic, J., H. P. Benton, D. Rinehart, A. Epstein, M. E. Kurczy, M. D. Boska, H. E. Gendelman, and G. Siuzdak. 2015. An interactive cluster heat map to visualize and explore multidimensional metabolomic data. *Metabolomics: Official Journal of the Metabolomic Society* 11 (4):1029–34. doi: 10.1007/s11306-014-0759-2.
Jansen, J. J., J. Engel, J. Gerretzen, L. Blanchet, E. Szymańska, G. Downey, and L. M. Buydens. 2013. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry* 50 (October):96–106. doi: 10.1016/j.trac.2013.04.015.
Johs, H., and H. Johs. 2018. Multiple correspondence analysis. In *Multiple correspondence analysis for the social sciences*, 31–55. Thousand Oaks (CA): Sage. doi: 10.4324/9781315516257-3.
Juan, A. d., A. Gowen, L. Duponchel, and C. Ruckebusch. 2019. Image fusion. In *Data handling in science and technology*, vol. 31, 311–44. Elsevier Ltd. doi: 10.1016/B978-0-444-63984-4.00011-9.
Kowalski, B. R. 1980. Chemometrics. *Analytical Chemistry* 52 (5):112–22. doi: 10.1021/ac50055a016.
Kreutz, C., and J. Timmer. 2009. Systems biology: Experimental design. *FEBS Journal* 276 (4):923–42. doi: 10.1111/j.1742-4658.2008.06843.x.
Kruskal, J. 1977. The relationship between multidimensional scaling and clustering. In *Classification and clustering*, 17–44. Elsevier. doi: 10.1016/b978-0-12-714250-0.50006-1.
Lahat, D., T. Adali, and C. Jutten. 2015. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE* 103 (9):1449–77. doi: 10.1109/JPROC.2015.2460697.
Lapalus, E. 2016. "Linking sensory attributes to selected aroma compounds in South African Cabernet Sauvignon wines," MSc Thesis, Stellenbosch University.
Larsen, F. H., F. Van Den Berg, and S. B. Engelsen. 2006. An exploratory chemometric study of 1H NMR spectra of table wines. *Journal of Chemometrics* 20 (5):198–208. doi: 10.1002/cem.991.
Latorre, M. J., C. Garcia-Jares, B. Médina, and C. Herrero. 1994. Pattern recognition analysis applied to classification of wines from Galicia (Northwestern Spain) with certified brand of origin. *Journal of Agricultural and Food Chemistry* 42 (7):1451–5. doi: 10.1021/jf00043a012.
Lawless, L. J., and G. V. Civille. 2013. Developing lexicons: A review. *Journal of Sensory Studies* 28 (4):270–81. doi: 10.1111/joss.12050.
Mafata, M., A. Buica, W. du Toit, V. Panzeri, and F. P. van Jaarsveld. 2018. The effect of grape temperature on the sensory perception of méthode cap classique wines. *South African Journal of Enology and Viticulture* 39 (1):132–40. doi: 10.21548/39-1-2620.
Mafata, M., A. Buica, W. J. Du Toit, and F. P. van Jaarsveld. 2018. The effect of grape temperature at pressing on phenolic extraction and evolution in méthode cap classique wines throughout winemaking. *South African Journal of Enology and Viticulture* 39 (1):141–8. doi: 10.21548/39-1-2621.
Mafata, M., J. Brand, V. Panzeri, and A. Buica. 2020. Investigating the concept of South African Old Vine Chenin Blanc. *South African Journal of Enology and Viticulture*. 41 (2):168–182. doi: 10.21548/41-2-4018.
Mafata, M., J. Brand, V. Panzeri, M. Kidd, and A. Buica. 2019. 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 Research International* 125 (November):108515. doi: 10.1016/j.foodres.2019.108515.
Makhotkina, O., B. Pineau, and P. A. Kilmartin. 2012. Effect of storage temperature on the chemical composition and sensory profile of Sauvignon Blanc wines. *Australian Journal of Grape and Wine Research* 18 (1):91–9. doi: 10.1111/j.1755-0238.2011.00175.x.
Makris, D. P., S. Kallithraka, and A. Mamalos. 2006. Differentiation of young red wines based on cultivar and geographical origin with application of chemometrics of principal polyphenolic constituents. *Talanta* 70 (5):1143–52. doi: 10.1016/j.talanta.2006.03.024.
McKay, M., F. F. Bauer, V. Panzeri, L. Mokwena, and A. Buica. 2019. Profiling potentially smoke tainted red wines: Volatile phenols and aroma attributes. *South African Journal of Enology and Viticulture* 40 (2):1–16. doi: 10.21548/42-2-3270.
McKillup, S. 2005. *Statistics explained. Statistics explained: An introductory guide for life scientists*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511815935.
Moyano, L., M. P. Serratosa, A. Marquez, and L. Zea. 2019. Optimization and validation of a DHS-TD-GC-MS method to wineomics studies. *Talanta* 192:301–7. doi: 10.1016/j.talanta.2018.09.032.
Murray, J. M., C. M. Delahunty, and I. A. Baxter. 2001. Descriptive sensory analysis: Past, present and future. *Food Research International* 34 (6):461–71. doi: 10.1016/S0963-9969(01)00070-9.
Naumann, D., P. Lasch, M. Diem, and W. Ha. 2007. Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. *Journal of Chemometrics* 20 (5):209–20. doi: 10.1002/cem.
Niimi, J., O. Tomic, T. Naes, D. W. Jeffery, S. E. Bastian, and P. K. Boss. 2018. Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition. *Food Chemistry* 256 (vember 2017):195–202. doi: 10.1016/j.foodchem.2018.02.120.
Niimi, J., P. K. Boss, and S. E. Bastian. 2018. Sensory profiling and quality assessment of research cabernet Sauvignon and Chardonnay wines; Quality discrimination depends on greater differences in multiple modalities. *Food Research International (Ottawa, ON)* 106 (December 2017):304–16. doi: 10.1016/j.foodres.2017.12.060.
Nordhaug Myhre, J., K. Øyvind Mikalsen, S. Løkse, and R. Jenssen. 2018. Robust clustering using a KNN mode seeking ensemble. *Pattern Recognition* 76:491–505. doi: 10.1016/j.patcog.2017.11.023.
OIV. 2006. Determination of chromatic characteristics according to CIELab. *Compendium of International Analysis of Methods*, no. Chromatic Characteristics: 1–16. http://www.oiv.int/oiv/files/6-Domaines scientifiques/6-4Methodesdanalyses/6-4-1/EN/OIV-MA-AS2-11.pdf.
Pagès, J. 2004. Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadistica* 27 (1):1–26.
Pagés, J., and F. Husson. 2005. Multiple factor analysis with confidence ellipses: A methodology to study the relationships between sensory and instrumental data. *Journal of Chemometrics* 19 (3):138–44. doi: 10.1002/cem.916.
Parr, W. V., J. Ballester, D. Peyron, C. Grose, and D. Valentin. 2015. Perceived minerality in Sauvignon wines: Influence of culture and perception mode. *Food Quality and Preference* 41 (April):121–32. doi: 10.1016/j.foodqual.2014.12.001.
Pereira, A. C., M. J. Carvalho, A. Miranda, J. M. Leça, V. Pereira, F. Albuquerque, J. C. Marques, and M. S. Reis. 2016. Modelling the ageing process: A novel strategy to analyze the wine evolution

towards the expected features. *Chemometrics and Intelligent Laboratory Systems* 154 (May):176–84. doi: 10.1016/j.chemolab.2016.03.030.

Pereira, A. C., M. S. Reis, P. M. Saraiva, and J. C. Marques. 2011. Madeira wine ageing prediction based on different analytical techniques UV-Vis, GC-MS, HPLC-DAD. *Chemometrics and Intelligent Laboratory Systems* 105 (1):43–55. doi: 10.1016/j.chemolab.2010.10.009.

Petrovic, G., J.-L. Luis Aleixandre-Tudo, and A. Buica. 2019. Unravelling the complexities of wine: A big data approach to yeast assimilable nitrogen. *OENO One* 53 (2):107–27. doi: 10.20870/oeno-one.2019.53.2.2371.

Petrovic, G., M. Kidd, and A. Buica. 2019. A statistical exploration of data to identify the role of cultivar and origin in the concentration and composition of yeast assimilable nitrogen. *Food Chemistry* 276 (March):528–37. doi: 10.1016/j.foodchem.2018.10.063.

Pickering, G. J., and P. Demiglio. 2008. The white wine mouthfeel wheel: A lexicon for describing the oral sensations elicited by white wine. *Journal of Wine Research* 19 (1):51–67. doi: 10.1080/09571260802164038.

Radovanovic, A., B. Jovancicevic, B. Arsic, B. Radovanovic, and L. G. Bukarica. 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* 49:42–8. doi: 10.1016/j.jfca.2016.04.001.

Ribéreau-Gayon, P., Y. Glories, A. Maujean, and D. Dubourdieu. 2006. *Handbook of enology, the chemistry of wine: Stabilization and treatments*. Vol. 2. 2nd ed, ed. P. Ribereau-Gayon. Chichester, England: Wiley Online Library. doi: 10.1002/0470010398.

Rinnan, Å., F. v d. Berg, and S. B. Engelsen. 2009. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 28 (10):1201–22. doi: 10.1016/j.trac.2009.07.007.

Ríos-Reina, R., R. M. Callejón, F. Savorani, J. M. Amigo, and M. Cocchi. 2019. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta* 198 (October 2018):560–72. doi: 10.1016/j.talanta.2019.01.100.

Ríos-Reina, R., S. M. Azcarate, J. M. Camiña, and H. C. Goicoechea. 2020. Multi-level data fusion strategies for modeling three-way electrophoresis capillary and fluorescence arrays enhancing geographical and grape variety classification of wines. *Analytica Chimica Acta* 1126:52–62. doi: 10.1016/j.aca.2020.06.014.

Robinson, A. L., P. K. Boss, P. S. Solomon, R. D. Trengove, H. Heymann, and S. E. Ebeler. 2014. Origins of grape and wine aroma. Part 2. Chemical and sensory analysis. *American Journal of Enology and Viticulture* 65 (1):25–42. doi: 10.5344/ajev.2013.13106.

Salkind, N. J., and R. Kristin. 2007. *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412952644.

Seisonen, S., K. Vene, and K. Koppel. 2016. The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. *Food Chemistry* 210:530–40. doi: 10.1016/j.foodchem.2016.04.134.

Serra-Cayuela, A., M. Jourdes, M. Riu-Aumatell, S. Buxaderas, P. L. Teissedre, and E. López-Tamames. 2014. Kinetics of browning, phenolics, and 5-hydroxymethylfurfural in commercial sparkling wines. *Journal of Agricultural and Food Chemistry* 62 (5):1159–66. doi: 10.1021/jf403281y.

Silvestri, M., A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. L. Vigni, A. Marchetti, and M. Cocchi. 2014. A mid level data fusion strategy for the varietal classification of lambrusco PDO wines. *Chemometrics and Intelligent Laboratory Systems* 137 (October):181–9. doi: 10.1016/j.chemolab.2014.06.012.

Sohail, A., and F. Arif. 2020. Supervised and unsupervised algorithms for bioinformatics and data science. *Progress in Biophysics and Molecular Biology* 151:14–22. doi: 10.1016/j.pbiomolbio.2019.11.012.

Terblanche, E. 2017. Development of novel methods for tannin quantification in grapes and wine. MSc Thesis., Stellenbosch University.

Du Toit, W. J., and C. Piquet. 2014. Research note: Effect of simulated shipping temperatures on the sensory composition of South African Chenin Blanc and Sauvignon Blanc wines. *South African Journal of Enology and Viticulture* 35 (2):278–82. doi: 10.21548/35-2-1016.

Torrens, J., M. Riu-Aumatell, S. Vichi, E. López-Tamames, and S. Buxaderas. 2010. Assessment of volatile and sensory profiles between base and sparkling wines. *Journal of Agricultural and Food Chemistry* 58 (4):2455–61. doi: 10.1021/jf9035518.

Trikas, E. D., M. Rigini, D. Papi, A. Kyriakidis, and G. A. Zachariadis. 2016. A sensitive LC-MS method for anthocyanins and comparison of byproducts and equivalent wine content. *Separations* 3 (2):18. doi: 10.3390/separations3020018.

Valente, C. C., F. F. Bauer, F. Venter, B. Watson, and H. H. Nieuwoudt. 2018. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. *Scientific Reports* 8 (1):4987. doi: 10.1038/s41598-018-23347-w.

Valentin, D., S. Chollet, M. Lelièvre, and H. Abdi. 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):1563–78. doi:10.1111/j.1365-2621.2012.03022.x.

Varela, P., and G. Ares. 2014. *Novel techniques in sensory characterization and consumer profiling*. Boca Raton, FL: CRC Press. doi: 10.1201/b16853.

Vera, L., L. Aceña, J. Guasch, R. Boqué, M. Mestres, and O. Busto. 2011. Discrimination and sensory description of beers through data fusion. *Talanta* 87 (1):136–42. doi: 10.1016/j.talanta.2011.09.052.

Versari, A., G. P. Parpinello, F. Scazzina, and D. D. Rio. 2010. Prediction of total antioxidant capacity of red wine by fourier transform infrared spectroscopy. *Food Control.* 21 (5):786–9. doi: 10.1016/j.foodcont.2009.11.001.

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

Waterhouse, A. L. 2002. Wine phenolics. *Annals of the New York Academy of Sciences* 957:21–36. doi: 10.1111/j.1749-6632.2002.tb02903.x.

White, F. E. 1991. Data fusion lexicon. *The data fusion subpanel of the joint directors of laboratories, technical panel for C3* 15 (0704):15. http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA529661%5Cnhttp://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA529661.

Yu, P., M. Y. Low, and W. Zhou. 2018. Design of experiments and regression modelling in food flavour and sensory analysis: A review. *Trends in Food Science & Technology* 71 (January 2017):202–15. doi: 10.1016/j.tifs.2017.11.013.